

**Neurodevelopmental outcomes of children born preterm:
analyses into the validity of data collection and
outcome reports**

**Sze Ying Hilary Wong
MBChB, MRCPCH, MSc**

**Section of Neonatal Medicine, Department of Medicine
Imperial College London**

**Thesis submitted towards the degree of
Doctor of Philosophy
Imperial College London**

DECLARATION OF ORIGINALITY

I, Sze Ying Hilary Wong, declare herewith that the work presented in this thesis is my own and that all data and findings in the work were generated as a result of my own original research. I confirm that where information has been derived from the published and unpublished work of others, this is clearly indicated in the text and a list of references is given in the bibliography. I have acknowledged all main sources of help.

Signed:

A handwritten signature in black ink, appearing to read 'Sze Ying Hilary Wong', written in a cursive style.

S.Y. Hilary Wong

Date: 7th September 2015

COPYRIGHT DECLARATION

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.

ABSTRACT

Background and aims: Information on the neurodevelopmental outcomes of children born very preterm is required for multiple purposes. Reliable and up-to-date data sources are lacking. The overall aim of this thesis is to evaluate the validity and usability of the neurodevelopmental outcome data of very preterm children available from current data sources. The specific objectives were:

- 1) to examine the validity of outcome data recorded during routine follow-up assessment
- 2) to explore early childhood social-communication difficulties exhibited by very preterm children
- 3) to assess the stability over time of neurodevelopmental diagnoses made in early childhood.

Methods: Three studies were conducted to meet the objectives. For studies 1 and 2, I recruited children born at <30 weeks' gestation at 2 years corrected age (age corrected for prematurity) from 13 participating study sites. In study 1, I compared the agreement between the neurodevelopmental outcomes of 190 children recorded at their routine NHS assessments and data obtained by a research assessment using the Bayley Scales of Infant Development, 3rd edition. In study 2, the social-communication skills of 141 children were determined using the parent-completed Quantitative Checklist of Autism in Toddlers (Q-CHAT) questionnaire and compared to published results from the general population. In study 3, I conducted a systematic review and using meta-analytic methods, I calculated the pooled sensitivity and specificity of early developmental assessment in identifying school-age cognitive deficit from 24 studies.

Conclusions:

- 1) Compared with research assessment, routine NHS follow-up assessment had a low sensitivity but high specificity for identifying children with neurodevelopmental impairment.
- 2) Very preterm children display greater early childhood social-communication difficulties and autistic behaviour than the general population as measured by their parents on the Q-CHAT.
- 3) Early neurodevelopmental assessment has high specificity but low sensitivity for identifying later school-age cognitive deficits.

ACKNOWLEDGEMENTS

I thank my supervisors, Professor Neena Modi and Professor Frances Cowan for their invaluable advice, guidance and support. They have been inspirational role models and I have been very fortunate in being able to draw on and learn from their areas of expertise and experience.

I credit Dr Angela Huertas-Ceballos for her vision in using the Quantitative Checklist for Autism in Toddlers to study children born preterm and thank her for her encouragement. I am grateful to Ms Betty Hutchon for her patience during my training on using the Bayley Scales of Infant and Toddler Development, 3rd edition. I also recognise Dr Margaret Morton's input during the early development of my projects.

My PhD is funded by the National Institute for Health Research under the programme grant awarded to the 'Medicines for Neonates Applied Research Programme' (RP-PG-0707-10010). I acknowledge the Medicines for Neonates Investigators, who in addition to Professor Neena Modi, are Professors Deborah Ashby (Imperial College London), Peter Brocklehurst (University College London), Kate Costeloe (Queen Mary University of London), Elizabeth Draper (University of Leicester), Azeem Majeed (Imperial College London), Stavros Petrou (University of Warwick), Alys Young (University of Manchester), Ms Jane Abbott (Bliss) and Ms Jacquie Kemp (London), as well as Professors Michael Goldarce and Andrew Wilkinson, for their feedback during the progress of my work.

My thanks also to my colleagues at Imperial College London for their comradeship and in particular, Shalini Shanthakumaran for her brilliant explanation of statistical methods.

I appreciate the hard work of the local collaborators (Drs Hashir Ariff, Vivien Chan, Michele Cruwys, Angela D'Amore, Ambilika Das, Swee Fang, Vimala Gopinathan, Ann Humphreys, Victoria Jones,

Anthony Kaiser, Richard Nicholl, Ann Opute, Sharon Richman, Caroline Sullivan), administrative staff and research nurses (Victoria Timms, Linda Evans) at the participating study sites for their assistance in recruiting participants, organising clinic rooms and in electronic data entry.

The meta-analysis (study 3) was enriched by the generosity of many investigators who shared their unpublished and supplemental data for the study. They are Dr Haim Bassan (Tel Aviv Sourasky Medical Center), Prof Arie Bos (University of Bonn), Dr Jenny Bowen (Royal North Shore Hospital, Australia), Dr Marie-Laure Charkaluk (Hospital Saint Vincent de Paul, Lille), Dr Sarale Cohen (retired), Prof Olaf Dammann (Tufts University), Prof Linda De Vries (University Medical Centre Utrecht), Prof Ermellina Fedrizzi (University of Padua), Prof Vineta Fellman (Lund University, Sweden), Prof Peter Gray (Mater Mothers' Hospital, Australia), Prof Maureen Hack (Rainbow Babies & Children's Hospital, USA), Dr Howard Kilbride (Children's Mercy Hospitals, USA), Prof Neil Marlow (University College London), Dr Jennifer Pinto-Martin (University of Pennsylvania), Dr Karin Rademaker (University Medical Centre Utrecht), Prof Jon Skranes (Norwegian University of Science & Technology), Dr Karen Smith (University of Texas), Prof Mary Sullivan (University of Rhode Island), Prof Paul Swank (University of Texas), Prof H. Gerry Taylor (Case Western Reserve University), Dr Viena Tommiska (Hospital for Children and Adolescent, Helsinki), Dr Norbert Veelken (Hannover Medical School), Prof Dieter Wolke (University of Warwick) and Prof Lianne Woodward (Washington University).

I am blessed with a family that cheers me on. Most of all, I thank my husband, Siong, the unsung hero in my life.

This thesis is dedicated to the memory of my mother,

Leong Mui Leng

能当您的女儿是我的福气

谢谢您的养育之恩

TABLE OF CONTENTS

Declaration of originality	1
Copyright declaration	2
Abstract	3
Acknowledgements	4
List of figures	13
List of tables	17
List of abbreviations	19
Chapter 1: Introduction	22
1.1 The original contribution of work to current knowledge	22
1.2 Scope of thesis	22
1.3 Role of the investigator.....	23
1.4 Structure of the thesis	24
Chapter 2: Background and literature review	25
2.1 Why conduct neurodevelopmental follow-up?	25
2.2 When should we assess neurodevelopment?	25
2.3 Types of neurodevelopmental outcome measures in early childhood	28
2.3.1 Motor function and cerebral palsy	28
2.3.2 Developmental/cognitive function	29
2.3.3 Language and communication	31
2.3.4 Neurosensory outcomes	32
2.3.5 Neuropsychiatric outcomes	33
2.4 How can we assess early neurodevelopmental outcomes?	37
2.4.1 Standardised developmental and neuropsychological tests	37
2.4.2 Developmental screening tools	40
2.4.3 Standard neurological examination	41
2.4.4 Assessment of autistic features	42
2.5 Classification of neurodevelopmental outcomes	45
2.5.1 The National Perinatal Epidemiology Unit/ Oxford classification of functional status at two years	45
2.5.2 Classifications of disability used by studies reporting neonatal outcomes	48

2.5.3	Functional classification of cerebral palsy	49
2.6	Sources of neonatal neurodevelopmental outcome data	50
2.6.1	Neonatal research studies	50
2.6.2	Neonatal follow-up programmes.....	51
2.6.3	Routine data from universal surveillance programmes.....	53
2.6.4	Parent-completed questionnaires	54
2.6.5	Electronic health records	56
2.6.6	The UK National Neonatal Research Database.....	57
2.7	Neonatal outcome reporting in the UK	58
2.7.1	Recommendations for neonatal outcome reporting.....	58
2.7.2	Current status of neurodevelopmental follow-up assessment and reporting in the UK	59
Chapter 3: Aim and objectives		60
3.1	Aim	60
3.2	Hypotheses	60
3.3	Specific objectives.....	60
Chapter 4: Methods and materials		62
4.1	Study designs	62
4.2	Research ethics committee approval/ research database inclusion	62
4.3	Study sites	62
4.4	Study participants	63
4.5	Recruitment	64
4.6	Training to conduct the research assessment	65
4.7	The research assessment.....	66
4.7.1	Timing of research assessment.....	67
4.7.2	Assessment of cognition, language and neuromotor development	67
4.7.3	Assessment for neurological deficits and cerebral palsy.....	71
4.7.4	Assessment of social-communication skills	72
4.7.5	Record of observed behaviour during the research assessment	74
4.8	Classification of impairment from the research assessment.....	74
4.8.1	Classification of impairment based on Bayley-III scores.....	75
4.8.2	Classification of impairment based on the NPEU/Oxford criteria	76
4.9	Collection of outcome data from the routine follow-up assessments	77
4.10	Retrieval of data from the routine assessments and classification of disability	79
4.10.1	Classification of disability based on routinely recorded clinical data.....	79

4.11	Statistical tests and measures used	81
4.11.1	Measures of test validity: sensitivity and specificity.....	81
4.11.2	Cohen’s kappa statistic	82
4.12	Sample size calculation	83
4.13	Study 1 analyses: The validity of routine NHS assessment.....	86
4.13.1	Examining the characteristics of the study population	86
4.13.2	Comparing the different methods of classification of impairments from the research assessment	86
4.13.3	Agreement in the classification of impairment between NHS and research assessments... ..	87
4.13.4	Variables associated with the validity of NHS neurodevelopmental data.....	89
4.14	Study 2 analyses: Social-communication in preterm children.....	89
4.14.1	Examining the characteristics of respondents.....	89
4.14.2	Comparison of Q-CHAT scores between the study population and the general population	90
4.14.3	Factors associated with Q-CHAT scores.....	90
4.14.4	Classifying children at risk for ASD using the Q-CHAT and Bayley-III Social-Emotional questionnaires.....	91
4.15	Study 3: Systematic literature review and meta-analysis.....	92
4.15.1	Eligibility criteria for study inclusion	92
4.15.2	Data sources and search strategy	93
4.15.3	Study selection	94
4.15.4	Study quality assessment.....	94
4.15.5	Data extraction and synthesis.....	97
4.15.6	Meta-analysis.....	98
4.15.7	Investigating heterogeneity by meta-regression and subgroup analysis	99
4.15.8	Post-hoc analysis to examine the change in mean developmental/ cognitive scores over time	100
4.15.9	Investigating publication bias	101
Chapter 5: Validity of standardised two-year neurodevelopmental data collected during NHS follow-up (Study 1 results)		103
5.1	Study population.....	103
5.1.1	Derivation of study population	103
5.1.2	Characteristics and representativeness of study population	105

5.2	Neurodevelopmental outcomes from research assessments	107
5.2.1	Categorisation of neurodevelopmental status using Bayley-III scores.....	108
5.2.2	Classification of impairment using NPEU/Oxford criteria.....	110
5.2.3	Concordance between Bayley-III and NPEU/Oxford criteria	111
5.2.4	Reliability of research assessments	112
5.3	Neurodevelopmental outcomes from routine NHS data	112
5.4	Agreement in the classification of outcomes between research and routine NHS assessments.....	115
5.4.1	Validity of NHS assessment against research assessment.....	115
5.4.2	Concordance in the assignment of the category of outcome between research and NHS assessments.....	122
5.4.3	Post-hoc analysis of the validity of NHS assessments using a different question set to identify ‘moderate-severe’ impairment.....	123
5.4.4	Variables affecting the validity of the NHS assessments	123
5.4.5	Behaviour during assessments and the effect on study findings	124
5.5	Results from the Hammersmith Infant Neurological Examination (HINE) and diagnosis of cerebral palsy	131
5.6	Discussion.....	132
5.6.1	Agreement between routine NHS data and research standard data	132
5.6.2	Explanatory factors that affected the concordance of NHS and research data	132
5.6.3	Strengths of study.....	135
5.6.4	Limitations affecting the internal validity of the study.....	136
5.6.5	Limitations affecting external validity (generalisability).....	140
5.7	Conclusions	140
Chapter 6: Early childhood social-communication difficulties in children born preterm (Study 2 results)		141
6.1	Characteristics of study population	141
6.2	Q-CHAT scores of the preterm population	144
6.3	Association of Q-CHAT scores with Bayley-III cognitive, language and motor scores.....	147
6.4	Neonatal and sociodemographic predictors of Q-CHAT scores.....	147
6.5	Comparison of Q-CHAT and Bayley-III Social-Emotional scores	149
6.6	Discussion.....	151
6.6.1	Social-communication skills of children born very preterm	151
6.6.2	Early screening for autism spectrum disorders	155

6.6.3	Strengths of the study.....	157
6.6.4	Limitations of the study	157
6.7	Conclusions	158
Chapter 7: Predictive validity of early developmental assessments in identifying school-age cognitive deficits in children born preterm or very low birth weight (Study 3 results)		159
7.1	Results of literature search	159
7.2	Description of included studies.....	159
7.2.1	Study populations	160
7.2.2	Developmental and cognitive assessments	164
7.2.3	Quality of included studies: results of QUADAS-2 appraisal.....	165
7.3	Predictive validity of early developmental assessment.....	169
7.3.1	Meta-analytic pooled estimates of sensitivity and specificity	171
7.3.2	Validity of early assessment assessed at different time points	173
7.3.3	Meta-regression: association of study-level variables with diagnostic validity	174
7.3.4	Post-hoc analysis: changes in the standardised mean difference between mean developmental and cognitive scores.....	178
7.3.5	Funnel plot for sample size-related effects and publication bias	179
7.4	Discussion.....	180
7.4.1	Internal validity	181
7.4.2	External validity (generalisability).....	183
7.4.3	Strengths and limitations of review and meta-analysis.....	184
7.4.4	Comparison with other studies.....	186
7.5	Conclusions	187
Chapter 8: General discussion and conclusions		188
8.1	Evaluation of evidence to original hypotheses	188
8.2	Clinical relevance of results	189
8.2.1	Reliability of neurodevelopmental assessments and outcome data recorded during routine clinical follow-up.....	189
8.2.2	Social-communication difficulties experienced by children born preterm	190
8.2.3	Predicting school-age cognitive impairment	190
8.3	Implications and recommendations for future practice.....	191
8.3.1	Individual versus population outcome data	191
8.3.2	Considerations and recommendations for neonatal follow-up programmes and data recording	193

8.3.3	Improvement in completeness of population outcome data.....	195
8.3.4	Intervention for social-communication difficulties.....	196
8.4	Areas for future research.....	198
8.4.1	Investigate barriers to uptake of follow-up programme	198
8.4.2	Improve validity of data collection form.....	199
8.4.3	Identification of risk factors for symptoms of ASD in the preterm population	200
8.4.4	Determination of factors that affect the predictive validity of early assessment.....	200
8.4.5	Linkage with school-age outcome data	201
8.5	Conclusions	201
References.....		203
Appendix 1: Research outputs		223
Appendix 2: The NPEU/Oxford classification of disability		234
Appendix 3: Definitions of disability used by studies reporting neonatal outcomes.....		236
Appendix 4: Study 1 data collection form.....		245
Appendix 5: Bayley-III Social-Emotional and Q-CHAT questionnaires		258
Appendix 6: Systematic electronic literature search strategy.....		266
Appendix 7: Study 1 sensitivity analyses		267
Appendix 8: Studies included in systematic review (study 3)		269
Appendix 9: Scatterplots showing the relationship of study-level variables with diagnostic validity (study 3)		272

LIST OF FIGURES

Figure 4.1	Types of Bayley-III scores	71
Figure 4.2	Algorithm for the classification of impairment using data from NHS assessment	80
Figure 5.1	Flowchart of children through research and NHS assessments to form the study population	104
Figure 5.2	Children recruited at each study site as a proportion of the total number of eligible infants born in 2008-2010 and discharged home from each site	107
Figure 5.3	Neurodevelopmental status by Bayley-III scores and the predicted BSID-II MDI.....	110
Figure 5.4	Classification of the severity of cognitive impairment of the children by research and NHS assessments.....	115
Figure 5.5	Classification of the severity of receptive communication, expressive communication and overall communication impairment of the participants based on the Bayley-III and NPEU/Oxford classification by research assessment, and by NHS assessments.....	116
Figure 5.6	Classification of the severity of fine motor, gross motor and overall motor impairment of the participants based on the Bayley-III and NPEU/Oxford classification by research assessment, and by NHS assessments	116
Figure 5.7	Classification of the neurodevelopmental outcome of participants by the severity of the worst impairment in the cognitive, communication and motor domains through research and NHS assessments.....	117
Figure 5.8	The prevalence of impairment and the estimated sensitivity and specificity of NHS assessment in identifying impairment in different neurodevelopmental domains, stratified by <i>gestation at birth</i>	125
Figure 5.9	The prevalence of impairment and the estimated sensitivity and specificity of NHS assessment in identifying impairment in different neurodevelopmental domains, stratified by <i>sex of participants</i>	126
Figure 5.10	The prevalence of impairment and the estimated sensitivity and specificity of NHS assessment in identifying impairment in different neurodevelopmental domains,	

stratified by <i>the requirement for supplemental oxygen at 36 weeks corrected gestational age</i>	126
Figure 5.11 The prevalence of impairment and the estimated sensitivity and specificity of NHS assessment in identifying impairment in different neurodevelopmental domains, stratified by <i>the IMD quintile of residence at the time of assessment</i>	127
Figure 5.12 The prevalence of impairment and the estimated sensitivity and specificity of NHS assessment in identifying impairment in different neurodevelopmental domains, stratified by <i>whether English was the only language spoken at home</i>	127
Figure 5.13 The prevalence of impairment and the estimated sensitivity and specificity of NHS assessment in identifying impairment in different neurodevelopmental domains, stratified by <i>the corrected age of participants at the time of the NHS assessment</i>	128
Figure 5.14 The prevalence of impairment and the estimated sensitivity and specificity of NHS assessment in identifying impairment in different neurodevelopmental domains, stratified by <i>whether a standardised neurodevelopmental test was used during the NHS assessment</i>	128
Figure 5.15 The prevalence of impairment and the estimated sensitivity and specificity of NHS assessment in identifying impairment in different neurodevelopmental domains, stratified by <i>the grade of the assessor who conducted the NHS assessment</i>	129
Figure 5.16 The prevalence of impairment and the estimated sensitivity and specificity of NHS assessment in identifying impairment in different neurodevelopmental domains, stratified by <i>the time interval between NHS and research assessments</i>	129
Figure 5.17 The prevalence of impairment and the estimated sensitivity and specificity of NHS assessment in identifying impairment in different neurodevelopmental domains, stratified by <i>the examiner rated behavioural score</i>	130
Figure 5.18 The prevalence of impairment and the estimated sensitivity and specificity of NHS assessment in identifying impairment in different neurodevelopmental domains, stratified by <i>whether the participant was considered 'difficult to test' during the NHS assessment</i>	130
Figure 6.1 Histogram of Q-CHAT scores of the preterm study population with superimposed distribution of published Q-CHAT scores of unselected toddlers (general population).....	145

Figure 6.2 Histogram of Bayley-III Social-Emotional composite scores of the preterm population with superimposed distribution of the standardised norm scores (mean 100, SD 15).....	149
Figure 6.3 Scatterplot showing the relationship in the distribution of Q-CHAT and Bayley-III Social-Emotional composite scores of the study population.....	150
Figure 7.1 PRISMA flow diagram depicting the literature search process.....	161
Figure 7.2 Proportions of studies with low, high or unclear risk of bias and concerns regarding applicability	165
Figures 7.3 Results of cross-tabulations and coupled forest plots of the estimated sensitivities and specificities of early developmental assessments in identifying the presence of (a) any cognitive impairment and (b) severe cognitive impairment.....	170
Figure 7.4 Scatterplot of the true-positive rate (sensitivity) against the false-positive rate (1-specificity).....	171
Figures 7.5a and 7.5b Hierarchical summary receiver operator characteristic (HSROC) curves for the pooled sensitivity and specificity of early developmental assessment in identifying (a) any impairment and (b) severe impairment.....	172
Figures 7.6a and 7.6b Line graphs demonstrating the change in (a) sensitivities and (b) specificities when early developmental assessments were repeated at different ages in three studies.....	173
Figures 7.7a and 7.7b Line graphs demonstrating the change in (a) sensitivities and (b) specificities when school-age cognitive assessments were repeated at different ages in four studies.....	174
Figures 7.8a and 7.8b Coupled forest plots of the estimated sensitivities and specificities of included studies, ordered by (a) the developmental assessment tool used and (b) the inclusion/ exclusion of participants with severe neurosensory impairment.....	177
Figure 7.9 Change in the standardised mean difference in mean developmental and cognitive scores relative to normative or control populations over time	178
Figure 7.10 Funnel plot of the log diagnostic odds ratio against the inverse of the square root of the effective sample size, with pseudo 95% confidence limits.....	180
Figures S1a and S1b Scatterplot of the (a) sensitivity and (b) specificity reported in each study against the mean gestational age of the study population.....	272

Figures S2a and S2b Scatterplot of the (a) sensitivity and (b) specificity reported in each study against the mean birth weight of the study population	272
Figures S3a and S3b Scatterplot of the (a) sensitivity and (b) specificity reported in each study against the mean age at early developmental assessment of the study population	273
Figures S4a and S4b Scatterplot of the (a) sensitivity and (b) specificity reported in each study against the mean age at school-age cognitive assessment of the study population.....	273
Figures S5a and S5b Scatterplot of the (a) sensitivity and (b) specificity reported in each study against the mean time difference between assessments of the study population.....	274
Figures S6a and S6b Scatterplot of the (a) sensitivity and (b) specificity reported in each study against the year of birth of the study population	274
Figures S7a and S7b Scatterplot of the (a) sensitivity and (b) specificity reported in each study against the prevalence of development impairment in the study population	275
Figures S8a and S7b Scatterplot of the (a) sensitivity and (b) specificity reported in each study against the prevalence of <i>severe</i> development impairment in the study population....	275

LIST OF TABLES

Table 4.1	Method for assigning neurodevelopmental outcome using either the Bayley-III composite scores or the scaled score as follows:	75
Table 4.2	Precision of estimated sensitivity for different sample sizes and sensitivity estimates	85
Table 4.3	Review-specific signalling questions and standards for appraisal of study quality	96
Table 5.1	Demographic and neonatal characteristics of study population versus non-participants born <30 weeks gestation in 2008-2010 and discharged from the participating study sites	106
Table 5.2	Mean Bayley-III scores (scaled scores and composite scores) of study population	108
Table 5.3	Cross-tabulation of classification of impairment by Bayley-III scores and the NPEU/Oxford criteria with agreement between the two methods measured by kappa coefficient (κ)*.....	112
Table 5.5	Results of cross-tabulations comparing the NHS and research categorisation of impairment and the sensitivities and specificities of the NHS assessment in identifying children with <i>any</i> impairment against the ‘gold-standard’ research assessment	120
Table 5.6	Results of cross-tabulations comparing the NHS and research categorisation of impairment and the sensitivities and specificities of the NHS assessment in identifying children with <i>severe</i> impairment against the ‘gold-standard’ research assessment	121
Table 5.7	Concordance in the assignment of the category of impairment between research and NHS assessments as measured by unweighted and weighted kappa coefficients (κ)	122
Table 5.8	Sensitivities and specificities of the NHS data using a broader ‘moderate-severe’ impairment category in identifying participants with Bayley-III scores lower than -2 SD below mean.....	123
Table 6.1	Comparing the characteristics of respondents, non-respondents and non-participants born <30 weeks gestation in 2008-2010 and discharged from the participating study sites	143

Table 6.2	Item-specific distribution of Q-CHAT scores	146
Table 6.3	Univariable association of neonatal and sociodemographic factors with Q-CHAT scores.....	148
Table 6.4	Final multivariable model of factors associated with Q-CHAT scores.....	148
Table 6.5	Cross-tabulations of the number of children classified to be ‘at risk’ for ASD by the Q-CHAT and the Bayley-III SE questionnaires.....	151
Table 7.1	Characteristics of studies included in review	162
Table 7.2	Quality assessment of included studies using the QUADAS-2 appraisal tool	166
Table 7.3	Association of study-level variables with estimated sensitivity and specificity	176
Table S1	The sensitivities and specificities of the NHS assessment in identifying children with <i>any</i> impairment against the research assessment, using all singleton births and only one randomly selected child from each multiple birth set	267
Table S2	The sensitivities and specificities of the NHS assessment in identifying children with <i>severe</i> impairment against the research assessment, using all singleton births and only one randomly selected child from each multiple birth set	268

LIST OF ABBREVIATIONS

ADHD	Attention deficit hyperactive disorder
AIMS	Alberta Infant Motor Scale
ASD	Autism spectrum disorder
ASQ	Ages and Stages Questionnaire
Bayley-III	Bayley Scales of Infant and Toddler Development, third edition
BAPM	British Association of Perinatal Medicine
BIA	Brief Intellectual Ability
BSID	Bayley Scales of Infant Development
BSID-II	Bayley Scales of Infant Development, second edition
CI	Confidence interval
CQUIN	Commissioning for Quality and Innovation
CSBS-DP-ITC	Communication and Symbolic Behavior Scales Developmental Profile Infant-Toddler Checklist
DOR	Diagnostic odds ratio
DQ	Developmental quotients
ELBW	Extremely low birth weight
ELGAN	Extremely Low Gestational Age Newborns
ESS	Effective sample size
FN	False-negatives
FP	False-positives
FPR	False-positive rate
FS-II	Functional Status II
GMDS	Griffiths Mental Development Scales
GMFCS	Gross Motor Function Classification Scale
HINE	Hammersmith Infant Neurological Examination

HSROC	Hierarchical summary receiver operator characteristic curve
ICF-CY	International Classification of Functioning, Disability and Health – Children and Youth Version
IMD	Index of Multiple Deprivation
IQ	Intelligence quotients
IQR	Inter-quartile range
ITSP	Infant/Toddler Sensory Profile
KABC	Kaufman Assessment Battery for Children
M-CHAT	Modified Checklist for Autism in Toddlers
MACS	Manual Ability Classification System (MACS)
MDI	Mental Development Index
NDAU	National Data Analysis Unit
NHS	National Health Service
NICHD	National Institute of Child Health and Human Development
NNAP	National Neonatal Audit Programme
NNRD	National Neonatal Research Database
NPEU	National Perinatal Epidemiology Unit
PARCA-R	Parent Report of Children’s Abilities – Revised
PEDS	Parent’s Evaluation of Developmental Status
OR	Odds ratio
PDDST-II	Pervasive Developmental Disorders Screening Test, second edition
PDI	Psychomotor Development Index
PPV	Positive predictive value
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analysis
Q-CHAT	Quantitative Checklist for Autism in Toddlers
QUADAS-2	Quality of Diagnostic Accuracy Studies version 2

RAKIT	Revision Amsterdam Children's Intelligence Test
RCPCH	Royal College of Paediatrics & Child Health
ROC	Receiver operator characteristic curve
ROP	Retinopathy of prematurity
SCPE	Surveillance of Cerebral Palsy in Europe
S-B-III/IV	Stanford-Binet Intelligence Scales, third or fourth edition
SD	Standard deviation
SE	Social-Emotional
SGS	Schedule of Growing Skills
SON-R	Snijders-Oomen Nonverbal Revised
TN	True-negatives
TP	True-positives
TPR	True-positive rate
TRPG	Thames Regional Perinatal Group
VLBW	Very low birth weight
WAIS-R	Wechsler Intelligence Scale for Adults – Revised
WASI	Wechsler Abbreviated Scale of Intelligence
WHO	World Health Organisation
WISC-III/R	Wechsler Intelligence Scale for Children, third or revised edition
WPPSI/-R	Wechsler Pre-school and Primary Scale of Intelligence or revised edition

CHAPTER 1

INTRODUCTION

1.1 THE ORIGINAL CONTRIBUTION OF WORK TO CURRENT KNOWLEDGE

Rapid advances in perinatal care over the past few decades have led to significant improvements in gestation-specific survival rates. In most developed countries, neonatal mortality and major disability rates are low and these outcome measures are no longer sufficiently sensitive measures of quality of care or assessment of new interventions. Current methods for gathering neonatal outcome information, for clinical and research purposes, stem from traditional practices focusing on early assessment of major disabilities and are limited by costs and resource availability. We should, therefore, reassess our current practices for neurodevelopmental follow-up and obtaining outcome data.

This thesis amalgamates my work in three separate studies, each focusing on a different key area:

1. The reliability of neurodevelopmental outcome data recorded during routine post-discharge clinical follow-up (Study 1).
2. The understanding and assessment of early social-communication skills among children born very preterm (Study 2).
3. The validity of developmental assessment conducted in early childhood for predicting cognitive deficits at school-age (Study 3).

1.2 SCOPE OF THESIS

Preterm birth is associated with multiple morbidities. In this thesis, I have focused my work on neurodevelopment, and specifically the assessment of cognitive, language, motor and early social-communication outcomes, rather than overall health. I recognise that neonatal follow-up, neurodevelopmental assessment and data acquisition require financial considerations. However,

this goes beyond the scope of my thesis and I have evaluated neurodevelopmental outcome data collection from only clinical and research perspectives.

1.3 ROLE OF THE INVESTIGATOR

Study 1 in this thesis comprises part of a National Institute for Health Research funded programme of work known as the “Medicines for Neonates Applied Research Programme” (RP-PG-0707-10010), which aims to develop the use of operational clinical electronic data captured at the point of care for multiple purposes. The idea for this project was conceived by the Medicines for Neonates Investigators (see Acknowledgements). I was responsible for executing all stages of the study including designing and producing the study protocol, obtaining approval from the research ethics committee, setting up the recruitment procedure at each study site, conducting the research assessments, compiling the datasets, analysing the results and completing the reports. I conceived and refined the idea for study 2 following discussions with Dr Angela Huertas-Ceballos (Consultant Neonatologist, University College London Hospital) and study 3 was developed with the guidance of my supervisors Professors Neena Modi and Frances Cowan. I planned and carried out all stages of studies 2 and 3, including the collection, quality assessment and analysis of the data, writing the papers and disseminating the findings. Throughout this process, I received guidance from Professor Deborah Ashby and Ms Shalini Santhakumaran (Imperial College London) on statistical methods although I undertook all statistical analyses myself. I also disseminated the findings from these studies through presentations at local and international meetings and publications of papers. The list of related outputs can be found in Appendix 1.

1.4 STRUCTURE OF THE THESIS

The thesis begins with a review on why, when and how neurodevelopmental assessments are conducted and the types of outcome measures that are commonly reported (chapter 2). This chapter also sets the scene for the three studies by detailing the available sources of neonatal neurodevelopmental outcome data and the current status of outcome reporting in the United Kingdom.

In chapter 3, I state the aim and objectives of the thesis.

In chapter 4, I describe in detail how participants were recruited from the participating study sites, the conduct of the research assessment and the process of data collection and synthesis. I also describe how I conducted the systematic literature search and extracted and synthesised relevant data for study 3. I explained the statistical methods I used to analyse the data for each study.

Chapters 5, 6 and 7 contain the results from studies 1, 2 and 3 respectively. In these chapters, I also discuss the strengths and limitations of the methods I adopted and the validity of my findings.

In chapter 8, I consider the implications of my findings for clinical practice, highlight important areas for future research and draw overall conclusions.

The appendices include a list of presentations and publications to date in relation to the research carried out during my PhD as well as details of the systematic literature search process for study 3 and results from sensitivity analyses.

CHAPTER 2

BACKGROUND AND LITERATURE REVIEW

2.1 WHY CONDUCT NEURODEVELOPMENTAL FOLLOW-UP?

Children born preterm, defined as birth at less than 37 weeks gestation, experience a major disturbance to their maturing body systems at a time of critical and rapid development. The preterm brain, in particular, is vulnerable to a constellation of prenatal, perinatal and postnatal insults that increase the risk for adverse neurodevelopmental outcomes. Reports on the outcome of surviving infants born at less than 32 weeks gestation had recorded severe disability rates between 5% to 56% (Milligan 2010). Longer term follow-up studies show that the adverse consequences of preterm birth are still apparent in adolescence and adulthood (Hille 2007, Doyle 2010a). Post-neonatal intensive care discharge follow-up and neurodevelopmental assessment of preterm children is necessary for multiple purposes. For an individual child, clinical outcomes are determined to ensure that disability, where present, is identified and timely intervention provided. Professionals require up-to-date outcome information to counsel, advise and support parents and to allow parents to make informed decisions about their child's care. Unit- and population-based outcome data are essential for service planning, benchmarking and evaluation. Neurodevelopment is also an outcome measure in epidemiological studies, in order to monitor trends and variations in population prevalence of disability, as well as in clinical research studies, to assess how outcome is altered by changes in clinical care. Furthermore, health care commissioners and health economists would be interested in the long-term impact of neonatal intensive care to assess the cost-effectiveness of running this service.

2.2 WHEN SHOULD WE ASSESS NEURODEVELOPMENT?

The optimal age for neurodevelopmental follow-up assessment is dependent on several factors, including the purpose of the assessment, the primary outcome measure and the resources available

for the follow-up programme. The majority of outcome studies report neurodevelopmental outcomes at 18 or 24 months corrected age (age corrected for prematurity). Similarly, most neonatal services provide follow-up assessment to around 2 years corrected age. 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 is generally felt that at 18 to 24 months of age, a meaningful assessment of neurodevelopment can be reliably conducted and it would still be relatively easy to achieve a good follow-up rate. Furthermore, if the data are then used to evaluate the quality of neonatal care, it would be pertinent that findings at this stage remain applicable to current clinical practice. There are unlikely to be major changes in practice that would significantly influence outcomes within a 2-year period and at this age, the effect of sociodemographic influences will be minimal (Lyon 2007).

Clearly if the follow-up assessment is for clinical purposes, then the timing would be driven by the child's clinical needs. An early assessment is of value as timely intervention can be facilitated if neurodevelopmental delay or impairment is identified. If the assessment is performed to study the impact of a perinatal event or intervention, it will need to be sufficiently early to be associated with the perinatal event. However, there are developmental and maturation changes that affect the diagnostic accuracy of early findings judged in the first couple of years. It has been recognised that transient neurological dystonia which mimic cerebral palsy can improve or resolve completely during the first year of life (Pedersen 2000). The literature suggests that a reliable early diagnosis of moderate to severe cerebral palsy can be made by 18 months corrected age and of mild cerebral palsy by 24 months corrected age (Paneth 2003). The signs of emerging cognitive, language or behavioural impairment may be subtle and only become evident at an older age. 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 recent cohort 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) (Munck 2012, Potharst 2012). 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 (Hack 2005). 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 (Roberts 2010).

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 (Marlow 2004, Doyle 2010a, Aylward 2002). These 'high-prevalence/ low-severity dysfunctions' include low average IQ scores, learning disabilities, attention and behavioural problems. The impact of these problems usually only becomes apparent at school-age when the child is faced with increased demand for more developmentally complex functions that were not required in early childhood. Longer-term follow-up studies are therefore necessary to uncover the prevalence and nature of these subtle disabilities, and are essential for health and educational resource planning. Increasing the duration of follow-up is, however, associated with higher attrition rate, which will affect the quality of the data collected. A systematic review of 20 papers that described neurodevelopmental outcomes of ELBW infants at 18 to 24 months found that cohorts with greater loss to follow-up reported higher rates of impairments (Guillen 2012). For randomised controlled trials, an attrition rate of greater than 20% was conventionally considered to pose serious threats to the validity and generalisability of the results (Fewtrell 2008). If neonatal interventional trials were to conduct long-term follow-up, the reduction in sample size due to loss of follow-up and the consequent loss of power of the study will need to be taken into consideration at the design

stage by increasing the required sample size. This will, in turn, have significant implications on the required resources for the study.

All follow-up programmes, whether for clinical or research purposes, incur significant costs related with the employment of trained staff, interim assessments, long-term tracking, data management and analysis, and need financial and logistic support to sustain long-term. The cost and resource limitations are often the main constrain on maintaining follow-up assessments.

2.3 TYPES OF NEURODEVELOPMENTAL OUTCOME MEASURES IN EARLY CHILDHOOD

2.3.1 Motor function and cerebral palsy

Cerebral palsy is the most commonly quoted outcome in neonatal follow-up studies. It is an umbrella term used to describe a group of non-progressive permanent disorders of movement and posture that occur following damage to the developing fetal or infant brain. There is a range of classification systems for cerebral palsy but it is most commonly described based on the nature of the neurological abnormality (e.g. spastic, dyskinetic or dystonic) and the topography of limb involvement. Spastic diplegia, resulting from injury to the internal capsule, is the most common form of cerebral palsy in preterm children. Among European population-based studies, the prevalence of cerebral palsy among infants born <32 weeks were reported to be between 6% and 9% (Milligan 2010). However, the inconsistencies in diagnostic definitions used by different studies have resulted in variability in the reported prevalence and restricted comparability between populations and over time. In 2000, investigators in 14 European centres maintaining cerebral palsy registers formed the Surveillance of Cerebral Palsy in Europe network and agreed on set of standardised definition and classification system to provide a framework for future collaborative research (SCPE 2000). More recently, investigators for the Extremely Low Gestational Age Newborns (ELGAN) study also developed an algorithm to diagnose and classify cerebral palsy to facilitate comparison between

studies (Kuban 2008). Using the algorithm, they reported an 11.4% prevalence of cerebral palsy among survivors of preterm birth before 28 weeks gestation in 14 institutions in the United States of America (USA).

Even in the absence of cerebral palsy, preterm infants experience abnormal patterns of motor development and neuromotor dysfunction (Bracewell 2002). Several authors have described the presence of transient dystonia, which may mimick cerebral palsy, in the first year of life in almost one-third of very low birth weight (VLBW; birth weight <1500g) cohorts (Pedersen 2000, Pallás Alonso 2000). Gross motor delay as well as poorer fine motor skills and coordination are also apparent among preterm children. A meta-analysis of studies that assessed the motor ability of children born very preterm (gestation ≤ 32 weeks) or with VLBW using standardised psychometric motor test reported that these children obtained motor scores between 0.57 and 0.88 standard deviations (SD) behind their term-born peers or typically developing children (de Kieviet 2009).

2.3.2 Developmental/cognitive function

Whilst many parents and professionals focused on the debilitating nature of severe motor impairment, the most common disability among preterm children is in fact developmental or cognitive delay (Wood 2000b, Stoelhorst 2003). Cognitive ability can be described using developmental quotients (DQ) or intelligence quotients (IQ) derived through standardised developmental or intelligence tests (see section 2.4.1). There is no criterion standard for determining developmental or cognitive delay. Conventionally, a standardised DQ or IQ more than 2 SD below the population mean is used to define impairment or disability, as it represents the lowest functioning 2.3% of the population. However, the accuracy of this cut-off in separating children with and without developmental delay depends on the validity of the assessment tool and the representativeness of the comparative population. Since the publication of the Bayley Scales of Infant and Toddler Development, third edition (Bayley-III) in 2006, concerns had been raised that it

underestimates developmental delay using the conventional cut-off. The mean Bayley-III cognitive scores of 'normal' term-born control groups were nearly 10 points above the normative mean (Anderson 2010, Lowe 2012) and when the preterm children were concurrently tested using both the Bayley-III and the BSID-II, cognitive and language scores were between 3 and 18 points higher than the mean BSID-II MDI scores (Vohr 2012, Moore 2012c, Lowe 2012). Possible explanations for this observed discrepancy suggested by the Bayley-III development team include differences in the demographic characteristics between the normative samples and the norming methodology adopted by the BSID-II and the Bayley-III (Bayley 2008). Crucially, one of the key differences in the standardisation procedure between the two editions was the inclusion of 'clinical cases' (children with cognitive, physical and behavioural issues) to constitute approximately 10% of the Bayley-III standardisation sample. This was made on the basis that excluding these conditions with higher risk for developmental impairment that are normally present in the general population would falsely inflate the average test scores. However, the effect of these clinical cases in the normative sample appeared to be an increased in discrepancy between BSID-II and Bayley-III scores particularly in the lower functioning range (Lowe 2012, Moore 2012c), leading to an overestimation of ability when the Bayley-III is used in children with suboptimal development. As the Bayley Scales are widely used in neonatal follow-up studies, the introduction of the new edition has made it difficult to interpret and compare results from longitudinal studies that had incorporated both the second and the third editions. Some of these studies had developed conversion algorithms or suggested different cut-off scores to determine developmental delay in order to allow comparison between cohorts (Lowe 2012, Moore 2012c, Johnson 2014).

The prevalence of developmental or cognitive impairment exists as a gradient that is inversely related to gestational age (Bhutta 2002). In the EPICure 2 study, which followed up surviving infants born before 27 weeks gestation in 2006 in England, 35% of survivors assessed at age 3 years had cognitive scores (predicted MDI) more than one SD below the normative mean (Moore 2012b). A

similar prevalence of developmental disability was also reported among survivors of extremely preterm birth in Sweden (Serenius 2013). In the Swedish cohort, survivors achieved 9.2 (95% confidence interval [CI] 6.9 to 11.5) points lower on the Bayley-III cognitive scale at age 2.5 years compared with matched controls, after adjusting for parental education and maternal country of birth. The magnitude of this difference in cognitive scores is equivalent to one SD.

Similar results can be observed in children evaluated at school-age. Bhutta et al conducted a review of 15 studies that compared the cognitive scores of 1556 preterm children at school-age with 1720 term-born controls (Bhutta 2002). Preterm children obtained cognitive test scores between 7.0 and 22.7 points lower than controls, with the meta-analytic weighted mean difference being 10.7 (95% CI 9.23 to 12.47). A more recent population-based comparison of school-age children born before 28 weeks gestation or with birth weight below 1000g to term-born controls revealed a 0.7 SD reduction in IQ points in the preterm children, after adjusting for sociodemographic factors and exclusion of children with neurosensory impairment (Hutchinson 2013).

2.3.3 Language and communication

Speech and language development is dependent on several processes including an intact auditory system, general cognitive function such as working memory and representational competence (the ability to extract commonalities from experiences and represent them abstractly or symbolically) (Rose 2009) and social interactions. In early language development, receptive (comprehension) and expressive (production) language are often considered and assessed separately. Language can be further divided into semantics (the meaning of words and sentences), morphology (word form), syntax (language structure), phonology (understanding of speech sounds) and pragmatics (the use of language in a social context). Preterm infants have documented delays in receptive language processing (Jansson-Verkasalo 2004), expressive language acquisition (Sansavini 2007, Wolke 1999), articulation and phonological short-term memory (Sansavini 2007, Pietz 2004, Vohr 2014). Current

evidence seems to suggest that although some preterm children have language abilities within the normal range, many are functioning behind their full-term peers (Foster-Cohen 2007, Anderson 2008). A meta-analysis of 12 studies published by Barre et al in 2011 reported that very preterm (gestational age less than 32 weeks gestation) and/or VLBW infants performed between 0.38 and 0.77 SD below their term-born counterparts in areas of expressive and receptive language (Barre 2011). A recent meta-regression of 6 studies for the difference in language scores between very preterm infants and term-born controls against the age at assessment between 3 and 12 years suggested that the deficit in language function deteriorated with increasing age (van Noort-van der Spek 2012).

2.3.4 Neurosensory outcomes

Hearing

Being born very preterm exposes an infant to multiple risk factors for hearing and visual deficits although, when compared with other neurodevelopmental outcomes, the incidence for neurosensory impairments remains relatively low. Some of the risk factors associated with hearing impairment are the use of mechanical ventilation, aminoglycoside antibiotics, loop diuretics, hypoxia and hyperbilirubinaemia (Cristobal 2008). Universal newborn hearing screen by otoacoustic emission testing is well established and offered to all infants prior to hospital discharge (Public Health England 2013). A study that compared the hearing screen results of 58 infants born <32 weeks gestation with infants from the well-baby nursery reported an 8-fold increase in failure rates among the very preterm infants (Korres 2005). However, this high failure rate may reflect the higher prevalence of middle ear effusions in preterm infants as the majority of infants who fail the screening test were found to have normal hearing or a mild conductive hearing loss by follow-up auditory brainstem evoked response assessment (Korres 2008). Published data in the past 20 years estimated that hearing impairment affected between 1.5% - 9% of infants born very preterm or VLBW although less

than 1% had severe bilateral sensori-neural hearing loss that was uncorrectable with hearing aids (Synnes 2012, Moore 2012b, Wood 2000b, D'Amore 2010, Ari-Even Roth 2006, Veen 1993).

Vision

Retinopathy of prematurity (ROP) resulting from disordered retinal vascular development is a major treat for vision loss in preterm infants and high risk groups receive regular screening ophthalmic examination (Royal College of Paediatrics and Child Health 2008). In the UK, ROP affects approximately 17% of infants born very preterm and/or VLBW (Dhaliwal 2008) and it accounts for around 3% of all childhood vision loss (Rahi 2003). A Canadian study reported that approximately 14% of infants with stage 3 or laser-treated ROP over a 10-year period developed visual impairment and 2.6% was blind at ages 4 to 6 years (Schiariti 2008). The preterm population also has an increased rate of strabismus, refractive errors (such as myopia and astigmatism) and deficiencies in visual fields, accommodation and visual perceptions (Holmstrom 2014, O'Connor 2002).

2.3.5 Neuropsychiatric outcomes

There has been an increasing awareness of the neuropsychiatric problems faced by survivors of preterm birth. Overall, the literature suggests an increased risk for attention deficit/hyperactivity (ADHD), emotional and social disorders, including autism spectrum disorders (ASD) among very preterm/VLBW children compared with the general population (Indredavik 2004, Farooqi 2007, Elgen 2002, Johnson 2011b). Most studies had examined the risk for psychiatric disorders in preterm cohorts using screening questionnaires that examine symptoms rather than through diagnostic evaluations. Hence, there is limited data on the estimated prevalence of psychiatric disorders in the preterm population.

Attention-deficit hyperactive disorder

Case-control studies have indicated a 2- to 3-fold increase in risk for ADHD in very preterm/VLBW infants compared to term-born controls (Johnson 2011b). A Norwegian population-based follow-up study reported that VLBW adolescents were more inattentive than controls but not more hyperactive (Indredavik 2004). Similar findings were reported in the EPICure 1 study (a longitudinal population-based study where all children born at less than 26 weeks gestation in the UK and Ireland between March and December 1995 were assessed at 30 months, 6 years and 11 years of age) (Johnson 2010b), suggesting that there is a specific risk for the ADHD inattentive subtype.

Autism spectrum disorders

Autism spectrum disorders (ASD) are a heterogeneous group of neurodevelopmental disorders characterised by impairments in communication, reciprocal socialisation and repetitive behaviour. Several studies have shown that preterm infants were more likely to screen positive on ASD early screening tools (Limperopoulos 2008). So far only 3 studies have validated results from the screening tests with clinical examinations and/or diagnostic instruments to estimate the prevalence of ASD among children born preterm (Johnson 2010a, Pinto-Martin 2011, Dudova 2014). From these studies, the estimated prevalence of ASD was reported to be 5% in children with birth weight <2000g (Pinto-Martin 2011), 9.7% among children born VLBW (Dudova 2014), and 8% in children born at <26 weeks gestation (Johnson 2010a). This represents an approximate 10-fold increase over the 2-9 per 1000 prevalence estimate in the general population (Williams 2006, Autism and Developmental Disabilities Monitoring Network Surveillance Year 2006 Principal Investigators; Centers for Disease Control and Prevention (CDC) 2009).

The Modified Checklist for Autism in Toddlers (M-CHAT) is the tool most frequently used in studies for early ASD screening in preterm populations (see section 2.4.4). Using the M-CHAT, high positive screening rates of 25% in VLBW infants (Limperopoulos 2008) and 21-41% in infants born before 28

weeks gestation (Kuban 2009, Moore 2012d) were found. Although an increased rate of positive screening for autism might be expected amongst children born preterm, it has also become apparent that the M-CHAT is poor at differentiating autistic symptoms from neurosensory, cognitive and motor impairments and that the specificity of screening for ASD in the preterm population is confounded by the high prevalence of these co-existing morbidities (Kuban 2009, Moore 2012d, Luyster 2011). Exclusion of children with these impairments reduced the positive screening rates at age two years from 21 to 10% in the ELGAN study (Kuban 2009) and from 41 to 16.5% in the EPICure study (Moore 2012d). Many 'critical' items on the M-CHAT such as "Does your child ever brings objects over to you to show you something?" and "Does your child respond to his/her name when you call?" depended on intact neurosensory and motor functions and were failed more frequently by children with overt motor and other disabilities (Luyster 2011).

Three studies have so far confirmed the low positive predictive value (PPV) of ASD screening tools in preterm populations. Using the Social Communication Questionnaire (Rutter 2003) to screen for ASD in children born extremely preterm at age 11 years, Johnson et al reported that only 9 out of 19 children with positive screens met the ASD diagnostic criteria (Johnson 2011a). The PPV of the M-CHAT in a cohort of children born ≤ 30 weeks gestation was 23% (3 children with ASD out of 13 positive screens) (Gray 2015) and in a cohort of children born VLBW, only 39.3% (13 out of 33) children who screened positive on one of three ASD screening tools and whose parents agree to a follow-up assessment had ASD (Dudova 2014). It is likely that, in addition to the presence of underlying neurosensory, cognitive and motor impairments affecting screening test interpretation, the high false-positive screening rate is also a reflection of an excess of ASD symptoms experienced by the preterm population. These symptoms may include difficulties with social awareness, communication and motivation (Movsas 2012). In the study by Johnson et al, after excluding children with serious functional disabilities, those with false positive screens had significantly lower cognitive scores and were more likely to have emotional, conduct, attention/hyperactivity and peer

problems than those with true negative screens (Johnson 2011a). In younger children, internalising behavioural problems (measured using the Child Behaviour Checklist (Achenbach 2000) were strongly associated with positive M-CHAT screen (Gray 2015, Limperopoulos 2008).

It is suggested that the ASD symptoms experienced by preterm children represent a 'preterm behavioural phenotype' characterised by behavioural difficulties (particularly in inattention/hyperactivity, social and emotional problems) that resemble, overlap or even meet the criteria for ASD and other psychiatric disorders (Johnson 2011b, Bowers 2015). 'Preterm ASD' may arise from a different causal pathway i.e. stemming from brain injuries and altered neurodevelopment associated with preterm birth. Some authors have found a direct relationship between shorter gestational age (Kuzniewicz 2014, Leavey 2013) or lower birth weight (Losh 2012, Stephens 2012) and increased risk of ASD. The role of fetal growth on the risk of ASD has also been implicated with several studies reporting higher risk among children born small for gestational age (Gray 2015, Moore 2012a, Moore 2012d). Perinatal and neonatal risk factors for ASD reported in the literature included chorioamnionitis, congenital or neonatal infections and hyperbilirubinaemia (Gardener 2009, Guinchat 2012). Recently, Bowers et al conducted a case-control study to explore the phenotypic differences between individuals with ASD born preterm and at term (Bowers 2015). Compared with those born at term, preterm subjects were significantly more likely to suffer from sleep apnoea (13.0% vs 3.0%), seizures (15.7% vs 8.3%) and attention-deficit/ hyperactivity disorders (13.0% vs 6.1%) and were more likely to be non-verbal (9.6% vs 4.6%). The implication of these findings is unclear and may reflect the presence of co-morbidities in the 'preterm phenotype'.

Emotional and behavioural problems

In terms of behavioural and emotional problems, 9 out of 12 case-control studies published in 1980-2001 and included in a meta-analysis reported an increase in internalizing behaviour among the very preterm/VLBW cases at age 5-12 years; 9 of 11 studies also reported an increase in externalizing

behaviour (Bhutta 2002). However, in a more recent meta-analysis based on parents and teachers ratings, the difference in internalizing behaviour scores reported between preterm/VLBW cases and term-born controls was small (preterm cases' scores were less than 0.28 SD below term controls' scores) and for externalizing behaviour, the difference was negligible (Aarnoudse-Moens 2009). Very preterm/VLBW children also experience more emotional problems, and in particular, with anxiety disorders. The EPICURE 1 study reported that extremely preterm children were 3.5 times more likely to have anxiety disorders than their term-born classmate controls (Johnson 2010b).

2.4 HOW CAN WE ASSESS EARLY NEURODEVELOPMENTAL OUTCOMES?

2.4.1 Standardised developmental and neuropsychological tests

Standardised developmental tests are considered the 'gold-standard' method for assessing a child's development and are the route by which most research studies obtain outcome data. The tests provide an inventory of key developmental milestones and are 'standardised' through administration to a large group of children (the normative sample) for whom the tests are designed (Johnson 2006). Standardised scores are age-adjusted scores with a normalised distribution and typically have a mean of 100 and SD of 15. Results from a standardised developmental test essentially compare an individual child's development to the normative sample. Standardised tests are designed to be administered by qualified examiners who would adhere to stringent administration and scoring protocol in order to optimise the objectivity and reliability of the results. From around 1930 to the present day, there has been a continuous and approximately linear increase in standardised test scores (Flynn 1999, Aylward 2011). This phenomenon had been described as the Flynn effect. Therefore, tests need to be updated and standardised with a contemporary normative sample to remain valid.

There is a range of standardised assessment tools available and the most suitable tool will depend on the developmental domain being assessed and the age of the child. A recent review on the existing standardised instruments that could be used to measure children's development at age 2 to 2½ years had been completed by the Policy Research Unit in the Health of Children, Young People and Families at the University College London Institute of Child Health (Bedford 2013). This review was commissioned by the Department of Health Policy Research Programme to consider tools that can be used as part of the 2 to 2½ years Healthy Child Programme (Shribman 2009) to monitor child development at population level. The report included a comprehensive analysis of the advantages and disadvantages of 13 different measures identified through a systematic literature search. In this thesis, I focus my review on the three standardised tools: (i) the Bayley Scales of Infant Development, which was used for the research assessment in study 1, (ii) the Griffiths Mental Development Scale and (iii) the Alberta Infant Motor Scale, which are tools that were used by clinicians during the routine NHS follow-up assessment of children born preterm in study 1.

Bayley Scales of Infant Development

The BSID was first published in 1969 and in a second edition (BSID-II) in 1993. A major revision and re-standardisation of the scales (Bayley-III) was published in 2005. The scales have been used extensively and translated into several languages to assess the development of infants and toddlers worldwide. In particular, the BSID-II, recognised to be highly reliable and valid, was the developmental test of choice among most major neonatal research studies, including the EPICure studies (Wood 2000b), the Victorian Infant Collaborative Study (The Victorian Infant Collaborative Study Group 1997) and the National Institute of Child Health and Human Development Neonatal Research Network (Vohr 2004). It yields Mental Development Index (MDI) and Psychomotor Index (PDI) that were standardised through a normative sample of 1700 children, stratified according to data from the 1988 US Census. Despite its popularity, the BSID-II has been criticised for the lack of separate assessments for language and non-verbal skills and for gross and fine motor performance.

The Bayley-III, standardised on a cohort of 1700 children in the USA in 2004 ameliorated these shortcomings by providing a more comprehensive assessment in separate cognitive, communication and motor domains, with sub-scale scores in receptive and expressive languages and fine and gross motor skills. The test had also been validated in 221 children from the UK and Ireland (Bayley 2010).

However, as described in section 2.3.2, studies have shown that concurrent scores achieved on the Bayley-III were on average higher than scores obtained on the BSID-II (Vohr 2012, Moore 2012c, Lowe 2012). These findings were inconsistent with the Flynn effect (Flynn 1999) and call into question the previous findings of morbidity using the older versions of the Bayley Scales as well as the validity of the current Bayley-III scales. The most common uses for the Bayley-III are in research setting or for the follow-up of children at high risk of developmental delay; however concerns that the scales are underestimating developmental delay will affect the interpretation and generalisability of the results. The high cost, levels of training and duration of time (approximately 90 minutes) required to administer the assessments further limits the use of the Bayley Scales as a population outcome measure.

Griffiths Mental Development Scales

The Griffiths Mental Development Scales (GMDS) were first published in 1954 and were the only developmental tests standardised in the UK. The original scales (GMDS 0-2) were designed to assess mental development in children below 2 years of age and underwent a major revision and re-standardisation on 665 children in 1996. Standardised scores (mean 100, SD 16) in 5 domains (Locomotor, Personal-Social, Language, Eye and Hand Coordination and Performance) were yielded that can be converted into a composite General Quotient (mean 10, SD 12). An extended scale (GMDS 2-8), comprising of an additional 'Practical Reasoning' domain was developed in 1970 for the assessment of children from 2 to 8 years with the third and most current edition published in 2006. As the GMDS 0-2 and the GMDS 2-8 were standardised using different normative population, they

are not directly comparable. This makes it tricky for neonatal follow-up assessment at 2 years when one could experience ceiling effect using the GMDS 0-2 with high achievers and children who were assessed later than intended. Concerns have also been raised about the representativeness of the normative population, the rigour of the standardisation process and the unknown psychometric properties (reliability and validity) of the scales (Johnson 2006). Despite these limitations, the GMDS remain a popular developmental test for neonatal follow-up.

Alberta Infant Motor Scale

The Alberta Infant Motor Scale (AIMS) was developed to assist physiotherapists and occupational therapists to measure motor development in high-risk infants (Piper 1992). It was standardised on a normative sample of 2202 infants between the ages of 1 week to 18 months living in Alberta, Canada between 1990 and 1992 and has established good reliability as well as concurrent and predictive validity. The AIMS assesses infant movement in prone, supine, sitting and standing positions and focuses on the attainment of motor milestones. An age-adjusted percentile rank is obtained. In my experience, the AIMS is usually used by allied health professionals who are involved in neonatal follow-up programmes rather than by medical staff. Liao et al reported that the scale has a ceiling effect with low precision after approximately 9 months of age, which significantly limit the utility of this tool in follow-up studies (Liao 2004).

2.4.2 Developmental screening tools

Developmental screening tools offer a cheap, quick and easy-to-use alternative to standardised assessments for determining neurodevelopmental outcomes. An example is the Schedule of Growing Skills (SGS) that is commonly used in community paediatrics and by health visitors in the UK. The SGS is based on Mary Sheridan's STYCAR developmental sequences (Bellman 1985) and underwent a revision in 1996 (Bellman 1996). Assessment scores in separate developmental skills areas are converted into developmental levels in months (developmental age equivalent based on

age-standardisation on a cohort of 348 children). Potential delay is raised if a child performs below the developmental level for their age. The validity of the original SGS as a screening tool was established by comparison with the GMDS (sensitivity 0.44 to 0.82; specificity 0.94 to 1.0 depending on the developmental domain) (Bellman 1985). Amess et al suggested that it was feasible for health visitors to follow-up preterm infants, for example as part of the Healthy Child Programme, using this tool as the proportions of children identified to have normal development and 'mild', 'moderate' or 'severe' delay were similar to other studies (Amess 2010). However, a major flaw of this study was that the accuracy of the classification was not examined. Screening tools are designed to be used for universal surveillance to identify children who require more thorough assessments. Such tools, when applied to a high-risk population, may have a low negative predictive value, particularly for mild developmental delay. Therefore, the utility of developmental screening tool for neonatal follow-up is still questionable.

2.4.3 Standard neurological examination

A neurological examination is an integral component of the neurodevelopmental assessment. The use of a standard neurological examination in conjunction with a gross motor functional assessment has been shown to increase the diagnostic accuracy for cerebral palsy (Vohr 2005, Paneth 2003). In study 1, I use the Hammersmith Infant Neurological Examination (HINE), which is a simple, quantitative method for assessing children between the ages of 2 and 24 months. The standard neurological examination assesses the cranial function, posture, movement, tone and reflexes of children and can be scored to yield an optimality score. The HINE was standardised in 135 low-risk term-born children between 12 and 18 months of age (Haataja 1999). The optimality score had been found to be valid for use in children born preterm, and was unaffected by the degree of prematurity or the age at assessment (Frisone 2002). An optimal score on the HINE between 9 and 18 months of age had been found to be predictive of the ability to walk at 2 years (Frisone 2002).

2.4.4 Assessment of autistic features

There is emerging evidence that developmental benefits may be achieved from early intervention programmes for children with ASD (Howlin 2009, Eldevik 2009, Eldevik 2010). Therefore, screening of very preterm infants could be important as they are considered to be at high risk for ASD. However, the pattern of early development of social-communication and autistic traits among children born preterm is unknown and there is no screening tool specifically designed for or validated in this population. The M-CHAT had shown promising test characteristics (sensitivity 87%, specificity 99%, PPV 80%, NPV 99%) when validated in a mixed population of unselected and high risk children (Robins 2001). It consists of 23 yes/no items (6 critical questions and 17 non-critical questions). A failed screening is defined if the parent reports that the child failed on any two critical questions or any three questions overall. As stated in section 2.3.5, when applied to the preterm population, high false-positive screening rates were found. High positive screening rates were also found with other screening tools including the Communication and Symbolic Behavior Scales Developmental Profile Infant-Toddler Checklist (CSBS-DP-ITC) (Wetherby 2008), the Infant/Toddler Sensory Profile (ITSP) (Dunn 2002) and the Pervasive Developmental Disorders Screening Test, 2nd edition (PDDST-II) (Seigel 2004) (Dudova 2014, Stephens 2012).

Dudova et al compared the utility of the M-CHAT, the CSBS-DP-ITC and the ITSP in preterm VLBW children at the corrected age of 2 years (Dudova 2014). The CSBS-DP-ITC is a 24-item parent-report questionnaire to be used between 6 and 24 months of age that screens for a wide range of disorders including global developmental delay, general language delay and autism. It was reported to have strong psychometric properties (sensitivity 87%-93%; specificity 75%; PPV 75%). The ITSP is a 48-item caregiver questionnaire that measures sensory modulation abilities in children aged 7 to 36 months. Its use as an early ASD screening tool was first evaluated in this study. The authors defined the criterion for positive screening to be test scores outside two SD of population norms. They found that the CSBS-DP-ITC yielded the highest number of positive screens (27.1%) and correspondingly

had the highest sensitivity (84.6%) but lowest specificity (84.9%) among the three tools. This was not unsurprisingly as the CSBS-DP-ITC is a general broadband screen rather than being specific for ASD. The utility of the M-CHAT and the ITSP was reported to be not significantly different (M-CHAT: positive screen 17.8%, sensitivity 69.2%, specificity 92.6%; ITSP: positive screen 14.3%, sensitivity 46.2%, specificity 94.2%). As expected, the PPV for all three tools were low (M-CHAT 50.0%; CSBS-DP-ITC 37.9%; ITSP 46.2%). Whilst the tools agreed substantially in negative cases, the positive cases varied between tools. 9% of the population screened positive on all three tools. Although the use of the tools in combination increased the PPV to 75%, this reduced the sensitivity to an unacceptable level of 23.1%.

Stephens et al also conducted a comparison between the PDDST-II, 'Response to Joint Attention' and 'Response to Name' (latter two are items adapted from the Autism Diagnostic Observation Scales, and had previously been used as individual screens for ASD (Nadig 2007, Sullivan 2007)) in 554 infants born <27 weeks gestation at 18-22 months of age (Stephens 2012). Once again, a high proportion of children (20%) screened positive on one or more tests. However, there was little overlap in the positive cases among the three screens (3% were positive for two tests and 1% had three positive screens). All the children who screened positive on all three tests had language impairment, which is of course a core symptom of ASD. A major limitation of this study is the lack of ASD diagnostic confirmation to fully assess the utility of these screening tools in the preterm population.

It is apparent from these studies that each ASD screening tool focuses on different behavioural and autistic traits and the choice of the most appropriate tool to assess early autism symptoms in preterm children is unclear. A major revision of the M-CHAT, the Quantitative Checklist for Autism in Toddlers (Q-CHAT) was published prior to the start of my PhD (Allison 2008). In addition, the Bayley-

III included a new social-emotional questionnaire in the current edition. My PhD provided an opportunity to examine the first use of these 2 assessment tools in the preterm population.

Quantitative Checklist for Autism in Toddlers

The Q-CHAT is a parent-completed questionnaire that aims to identify children at risk for autism with an improved sensitivity through updated items on the checklist and by having a 5-point rating scale (0 to 4) instead of a binary scoring system for each item. The Q-CHAT produces a score with a possible range from 0 to 100, with higher score reflecting greater social communication difficulties. Assessment of the test properties and clinical validity of the Q-CHAT is ongoing. In a preliminary report, the authors showed that the Q-CHAT scores from an unselected group of 754 toddlers aged between 17 and 26 months (mean age 21.2 months), living in Cambridgeshire in the UK, followed a near-normal distribution and were significantly lower (more normal) than the scores of children with ASD (Allison 2008). The quantitative nature of the Q-CHAT and the near-normal distribution of its scores make it a useful tool as an autistic trait measure.

Bayley Scales of Infant and Toddler Development Social-Emotional Scale

The Bayley-III social-emotional (Bayley-III SE) questionnaire was derived from the Greenspan Social-Emotional Growth Chart, which was reported to have a sensitivity of 67.2% and a specificity of 97.8% in identifying children with ASD (Casenhiser 2007). The questionnaire is designed to be completed by parents and is structured according to the anticipated acquisition of functional emotional milestones between birth and 42 months of age. It was standardised on the same normative cohort for the Bayley-III and therefore produces a standardised composite score with mean of 100 and SD for 15, with lower composite scores indicating greater social emotional difficulty.

2.5 CLASSIFICATION OF NEURODEVELOPMENTAL OUTCOMES

In 1980, the World Health Organization (WHO) proposed an international classification for describing the consequences of disease, including impairment, disability and handicap as an extension to the international classification of disease coding system. The WHO defined impairment as “any loss or abnormality of physiologic or anatomic structure”; disability was “any restriction or loss of ability (attributable to an impairment) in performing an activity in a manner and range considered normal for a human being” and a handicap was “a disadvantage for a given individual, resulting from a disability or impairment, that limits or prevents the fulfilment of a role that is normal for that individual”. In 2007, the WHO published the International Classification of Functioning, Disability and Health – Children and Youth Version (ICF-CY) (World Health Organization 2007), which is a modification of the original classification system to provide a more comprehensive framework to document child functional characteristics at the body, person and societal levels. Feasibility studies on the ICF-CY have found the content of the tool to contain sufficient breadth and depth for useful documentation of children’s body functions and activities (Bjorck-Akesson 2010). However, the ICF-CY is a complex classification and implementing it is a long-term project, hence it has not been used for describing outcomes from preterm birth.

2.5.1 The National Perinatal Epidemiology Unit/ Oxford classification of functional status at two years

In 1993, a working group of experts formed by the National Perinatal Epidemiology Unit (NPEU) and the former Oxford Regional Health Authority developed a standard minimum dataset that contained information relevant to the measurement of health status in early childhood (National Perinatal Epidemiology Unit and the former Oxford Regional Health Authority 1994). The minimum dataset consisted of patient identifiers (NHS numbers of mother and child, and child’s date of birth), socio-demographic measures (postcode, mother’s age at delivery, age last in full-time education and

support status at birth), perinatal variables (birth weight, gestation, gender, plurality, hospital of birth, and presence of congenital anomaly) and information on the child's health and functional status in eight clinical domains at age two years, based on responses to eleven key questions. This set of eleven key questions was designed through professional consensus to be used during assessment of children at 2 years of age without the need for specific training. They allowed description of the level of function of the child within each clinical domain and provided clear definitions of severe disability for outcome studies (Appendix 2). A 'severe disability' was considered as one that was likely to put the child in need of physical assistance to perform daily activities (A report of two working groups convened by the National Perinatal Epidemiology Unit and the former Oxford Regional Health Authority 1994). The intention was identify children with major functional loss who would go on to experience severe lifelong disabilities and information on the prevalence of children in this category may be useful to parents in the process of perinatal decision making. The set of key questions became known as the 'Health Status Questionnaire' or the 'NPEU/Oxford criteria for disability'. Moderate agreement between the NPEU/Oxford criteria and other methods of assessing disability had been reported (Jones 2002). The NPEU/Oxford classification had been used by several studies in the UK to report two-year outcomes of preterm children (Bohin 1999, D'Amore 2010), most notably the EPICure 1 study (Wood 2000b). Comparing the disability profile of the EPICure 1 cohort at 30 months and 6 years, the use of the NPEU/Oxford classification at 30 months corrected age had 50% sensitivity and 93% specificity for moderate or severe disability at 6 years of age (Marlow 2005).

In 2007, a working group was convened by the British Association of Perinatal Medicine (BAPM) and the National Neonatal Audit Project (NNAP) based in the Royal College of Paediatrics and Child Health (RCPCH) to evaluate and provide recommendations for the reporting of perinatal outcomes for audit, service evaluation and research purposes, to allow for benchmarking and comparison with international studies (British Association of Perinatal Medicine Working Party 2008). The working

group specified a dataset based on the model of the NPEU/Oxford criteria to provide a two-year health outcome measure to meet these multiple purposes. Clear definitions modified from the NPEU/Oxford criteria were described to allow standardised classification of preterm children at two years corrected age into one of three outcome categories as below:

Outcome	Definitions
(1) Severe neurodevelopmental disability	Cerebral palsy with GMFCS level 3-5; or Cognitive score <-3 SD below norm; or No useful hearing even with aids (>90dB hearing level); or Blindness or can only perceive light or light reflecting objects; or No meaningful words/signs; or Unable to comprehend cued command
(2) Neurodevelopmental impairment (moderate neurodevelopmental disability)	Cerebral palsy with GMFCS level 2; or Cognitive score -2 SD to -3 SD below norm; or Hearing loss corrected with aids (40 - 90dB hearing level); or Moderately reduced vision but better than severe visual disability, or unilateral blindness with good vision in contralateral eye; or Some but few than 5 words/signs; or Able to comprehend cued command but not un-cued command
(3) 'Normal'	Absence of any of the above

As with the NPEU/Oxford criteria, the 'severe neurodevelopmental disability' category aimed to describe children who would experience impaired independence throughout childhood. The 'moderate neurodevelopmental disability' category was developed to provide a "more inclusive classification with children with less severely impaired outcome which might serve to use as a comparator across populations and with international studies". Normality was judged based on the absence of an impairment. The working group recommended the inclusion of cerebral palsy in outcome reporting, classified using the Gross Motor Function Classification System (GMFCS; see section 2.5.3) to facilitate comparison with other international database.

2.5.2 Classifications of disability used by studies reporting neonatal outcomes

In Appendix 3, I display a table that lists the definitions of disability used by major population-based and multi-centre studies that have reported neurodevelopmental outcomes from preterm birth in the past two decades. All studies had defined impairments in the visual, hearing and neuromotor domains for the diagnosis of disability with most also included definitions in the developmental (cognitive) and/or communication domains. Many studies had adopted a scheme similar to that of the NPEU/ Oxford classification, such as defining severe visual impairment as bilateral blindness or the perception of light only and severe hearing impairment as deafness that could not be improved by aids. However, there were some differences that hindered direct comparison of disability rates between studies. Whilst some studies had defined the severity of disability into three levels (severe, moderate and mild), others had created a combined 'moderate-severe' category or did not include a 'mild' category. In the National Institute of Child Health and Human Development (NICHD) Neonatal Research Network follow-up study (Hintz 2011), a different terminology of 'profound' disability was used. Studies that had adopted the original NPEU/Oxford classification such as the EPICure (Wood 2000b), the Former Trent region cohort study (Rattihalli 2011) and EPIBEL (De Groot 2007) focused on functionality and did not include the diagnosis of cerebral palsy in the definition of disability. In these studies, participants who experience non-febrile seizures and/or other significant health problems affecting daily living were also considered to have at least mild disability. There were also differences in the 'cut-offs' for levels of severity in different studies which further limited comparison. For example, the presence of cerebral palsy of GMFCS level 3 would be considered 'severe disability' in the EPICure 2 (Moore 2012b) and the Swiss national cohort study (Schlapbach 2012) but was 'moderate disability' in the Norwegian national cohort (Leversen 2011) and the Victorian Infant Collaborative studies (Doyle 2010b). In addition, whereas most studies considered developmental scores of 1 – 2 SD below mean for age to represent mild developmental delay, in the EPIBEL study, participants achieving BSID-II MDI of 70 or greater (i.e. higher than -2 SD below mean) were considered to be normal (De Groot 2007). The Finnish national follow-up study (Tommiska

2003) did not include developmental scores in the classification of disability. Some studies such as the Swiss (Schlapbach 2012), the Netherlands (de Waal 2012) and the Norwegian (Leveresen 2011) national cohorts that had used the BSID-II MDI or the WPPSI-R full-scale IQ (which had incorporated language assessment in the mental scales) to classify disability did not include separate definitions for communication impairment.

2.5.3 Functional classification of cerebral palsy

The Gross Motor Function Classification System (GMFCS) is a method for categorizing the gross motor functional abilities of children with cerebral palsy (Palisano 1997). It is an age-related system and describes five levels of gross motor function, in which level I represents the least limitation and level V the most. It is widely used internationally and has proven to be reliable for classifying children with cerebral palsy to allow comparisons between different studies (Jahnsen 2006, Palisano 2006, Wood 2000a). The motor function of a child with cerebral palsy can change with time. Classification using GMFCS at age two years had been found to be stable over time and could be used to predict later mobility - the positive predictive value of GMFCS level I - III at age two years to predict walking by age 12 years was 0.74 and the negative predictive value was 0.77 (Wood 2000a).

The Manual Ability Classification System (MACS), developed by Eliasson and colleagues (Eliasson 2006), provides a description of how children with cerebral palsy use their hands to manipulate objects in daily activities. Similar to the GMFCS, the MACS is a 5-level classification system that focuses on the child's functional performance, with children at level I being able to handle objects easily and successfully and children at level V having limited ability in performing even simple actions. The MACS is designed for children between the ages of 4 and 18 years and the levels are determined based on age-appropriate activities. Interrater reliability (intraclass correlation coefficients between 0.83 and 0.98) (Jeevanantham 2015) and the stability of the MACS over time (intraclass correlation coefficients 0.97 for 3- to 5-year intervals) (Ohrvall 2014) were reported to be high.

2.6 SOURCES OF NEONATAL NEURODEVELOPMENTAL OUTCOME DATA

2.6.1 Neonatal research studies

Neonatal follow-up research usually focuses on studying the outcomes of specific groups of infants, determining the association between early risk factors and later findings and/or the clinical or cost-benefits of interventions. Most studies conduct primary data collection, using standardised assessment tools, to ensure objective and high-quality outcome data. However, several methodological issues are common among follow-up studies and affect the robustness of the results. Many neonatal studies use birth weight-based selection criteria, resulting in a heterogeneous study population of infants who were less mature but of appropriate size and infants who were more mature but small for gestational age. In addition, many studies collect data on survivors from tertiary centres alone and were not, therefore, representative of a complete geographical population (Kutz 2009). Although standardised assessment tools are used, children who display challenging behaviour or were too difficult to test are often excluded, leading to underestimation of the prevalence of disability. Wolke and colleagues also highlighted that the lack of a comparable control group flawed the evaluation of cognitive development of very preterm infants since the existing test norms might be outdated or demographically invalid (Wolke 1994). In the past few decades, there had been several large, population-based long-term follow-up studies such as the EPICure studies (<http://www.epicure.ac.uk>), the EPIPAGE studies (<http://www.perinat-france.org>) and the Victorian Infant Collaborative Studies (<http://www.vics-infantstudy.org.au>) that had contributed significantly to the understanding of the neurodevelopmental outcomes of extremely or very preterm infants. The eligibility criteria for these cohort studies are described in Appendix 3. However, considerable funding and resources were invested to set up and maintain these studies and the availability of future similar studies are not guaranteed. Furthermore, differences in the study populations, service infrastructure, healthcare policies, assessment tools and definitions of

impairments between studies limit direct comparisons of outcomes between populations and do not allow assessment of trends over time.

2.6.2 Neonatal follow-up programmes

Although it has been widely acknowledged that neonatal follow-up activities are a key part of delivering high-quality care for children who survived neonatal intensive care, the provision of follow-up appointments and assessments is still not universal in the UK (see current status of neurodevelopmental follow-up in the UK, section 2.7.2). The National Institute for Health and Care Excellence (NICE), in 2010, published a list of statements that define what high-quality specialist neonatal care should be (National Institute for Health and Care Excellence Topic Expert Group and project team 2010). The purpose of these statements were to set out markers of best practice that could be used by commissioners, managers and providers to improve the structure, process and outcomes of health care. One of the quality measures listed was to have ‘evidence of processes to enable collection of health outcome data within the network for babies who receive specialist neonatal care’. In the past year, the Department of Health in England has asked NICE to develop a clinical guideline on the developmental follow-up of preterm infants (NICE). It is anticipated that the guideline will be published in 2017.

The establishment of a coordinated neonatal follow-up programme at a regional, network or national level, in which eligible children are assessed at uniform ages with a common standardised set of assessments, would be an ideal means to obtain valid outcome data. The data recorded would be of high quality and could serve multiple purposes, including audit, benchmarking and research. In the UK, neonatal services are organised into ‘clinical networks’ with hospitals within each network providing different categories of neonatal care. As a result, many infants born very preterm would receive neonatal care in more than one hospital. The coordination of post-discharge follow-up varies widely among different neonatal networks. The most common model of practice is for follow-up

appointments to be organised at the time of hospital discharge and for many infants, the discharging hospital would be the local neonatal unit rather than the tertiary centre. It is unclear how often outcome data is shared between different units within a network. The NNAP, delivered by the RCPCH, has been using electronic data from neonatal units to audit two-year health status of children born at less than 30 weeks gestation and has reported its findings in annual publications since 2009 (National Neonatal Audit programme). For this project, the hospital of birth was assigned responsibility for collecting the outcome data for eligible all infants. However, this method could result in significant level of 'missing data' as infants often receive follow-up assessment coordinated through the hospital he/she was discharged from. In the 2013 NNAP report, two-year outcome data was available from only 44% of all infants born at less than 30 weeks gestation in England and Wales between July 2010 and June 2011. It is paramount that an appropriate denominator population is selected to meet the purpose of the data collection. For example, if the prevalence of disability is to be compared between networks, then data needs to be collected from the whole population that each network serves in order to avoid referral/ inclusion bias. Some neonatal networks have set up regional projects involving collaboration to achieve population-based clinical and outcome assessments that conform to standardised definitions (Salt 2006). The cost of setting up and running a follow-up programme, which would include training staff and funding continuing employment, is considerable (Dorling 2006) and the benefit of a follow-up programme over other data collection methods would need to be clarified in order to justify the expenditure.

In Switzerland, the Swiss Neonatal Network and Follow-up Group have set up a national follow-up programme and have recently published the two-year outcomes for a Swiss national cohort of extremely preterm infants born over an 8 year period between 2000 and 2008 (Schlapbach 2012). Even with a robust process for follow-up data collection based on a national prospective database, the attrition rate at 2 years was still 19%. As the infants lost to follow-up were less preterm and had a lower rate of broncho-pulmonary dysplasia than the study group, it is possible that the overall rate

of impairment from this national cohort was underestimated. The Canadian Neonatal Follow-up Network (<http://cnfun.ca>) is also establishing a collaborative group to deliver standardised neonatal follow-up assessments and develop a national neonatal outcome database. There had already been successful linkage of the Canadian Neonatal Network's national NICU admissions database with the national perinatal and perinatal surgical databases (Skarsgard 2006) and the creation of a perinatal outcome database will no doubt facilitate outcome evaluation research.

2.6.3 Routine data from universal surveillance programmes

In the UK, as in most developed countries, there is an established surveillance programme to monitor the health and development for all children (Shribman 2009). Routinely recorded data from the surveillance programme may provide an on-going source of neurodevelopmental outcome information that is easily accessible and can be analysed at low cost.

In the 1990s, several studies investigated the extent to which data recorded during routine service delivery can be used to report the outcomes of survivors of neonatal intensive care (Dawson 1997, Bohin 1999, Johnson 1999). The Trent Neonatal Follow-up Project reported that most of the data required to meet the NPEU/Oxford minimum dataset could be extracted from routinely available information systems (Dawson 1997). However, the quality of the data was variable and there was no standardisation in the interpretation or documentation of clinical assessments. An exercise on data linkage between the neonatal register and community child health surveillance database produced "error-free" linkage (using the identifiers date of birth, birthweight and gestation) in only 53.9% of children who had received neonatal intensive care. Modi and Carpenter reported similar problems when they reviewed the use of district and regional child health database in the North Thames Region to ascertain the two-year health status of children born at less than 29 weeks' gestation (Modi 1997). They were able to retrieve child health surveillance records for only two of 80 children surviving to two years. When Johnson and King used the routine child health information system to

compile a list of children with motor or sensory disability, they failed to identify 162 (36.3%) of 446 children listed on the co-existing population register of CP, sensorineural deafness and severe vision loss (Johnson 1999). The variable timing of routine screening had led to misclassifications in the outcomes of children, particularly if outcomes were ascertained based on developmental achievements reported at a younger age (Field 2001). Therefore, the overall opinion was that the routine information systems in place in the 1990s were inadequate for the provision of complete and accurate outcome data on children born preterm.

Even now, the coverage of the screening programme and the quality of the collected data remain unclear. In the USA, despite a policy statement from the American Academy of Pediatrics on universal developmental surveillance by paediatricians, it was estimated that nearly 50% of children did not receive a developmental assessment (Halfon 2004) and almost three-quarter of paediatricians do not use a standardised screening instrument (Sand 2005). Informal clinical assessment, when conducted in isolation had been found to be inadequate to detect children with developmental disabilities (Smith 1978).

Population registers of children with a particular impairment or disability such as a CP register are also established sources of data which may be of value in neonatal outcome reporting. They use clear case definitions for inclusion and exclusion that should improve accuracy of diagnoses. Nevertheless, registers require significant resources to ensure completeness. Moreover, single-condition registers would not reflect the global functional status of children born preterm and therefore insufficient to provide a comprehensive assessment of their outcomes.

2.6.4 Parent-completed questionnaires

Parent-completed questionnaires have been developed as a low-cost alternative to developmental tests to identify children with disabilities. They could be used in isolation, or as part of a research

study or follow-up programme, as a time-efficient and cost-effective source of outcome information. By providing a greater sense of parental involvement, they may also aid to improve parents' satisfaction.

The reported level of agreement of children's developmental status between parental perceptions and paediatricians' assessments had been inconsistent (Fooks 1999, Kim 1996, Bortolus 2002) and may be influenced by parental socio-demographic factors. The validity of the revised Parent Report of Children's Abilities (PARCA-R) (Johnson 2004, Johnson 2008), the Parent's Evaluation of Developmental Status (PEDS) (Pritchard 2005), the Functional Status II (FS-II) questionnaire (Da Costa 2009), the Ages and Stages Questionnaire (ASQ) (Skellern 2001) and a questionnaire adapted from the Griffiths Developmental Scales (Fooks 1997) had been evaluated in the preterm population. In particular, the PARCA-R were found to have good diagnostic utility for moderate to severe cognitive and language impairment when validated against the BSID-II (reported sensitivity 85%; specificity 87% (Johnson 2008)) and the Bayley-III (sensitivity 75-94%; specificity 79-89% (Martin 2013)) and had been used in a couple of neonatal clinical trials (Marlow 2006, Brocklehurst 2011). The ASQ has also been validated against the BSID-II and had a reported sensitivity of 100% and specificity of 87% at 24 months for severely delayed status (Gollenberg 2010). Nevertheless, the questionnaires were evaluated in relatively small and selected groups of participants (the PARCA-R was given to parents of infants enrolled in trials of enhanced parental support and neonatal sepsis; the modified Griffiths' questionnaire was completed by parents whose children had post-haemorrhagic ventricular dilatation) and their utility in a large-scale preterm population remains unclear. Although the typical response rates to postal questionnaires were reported to be between 52 - 61% (Cummings 2001), Field et al, when testing parent-completed questionnaires as a source of outcome data at 2 years following neonatal discharge, recorded a 90% response rate by maintaining contact with the families in the form of Christmas and birthday cards (Field 2001). However, only 51% of the questionnaires were returned within the requested 6 week period of the child reaching a

corrected age of 2 years and in one-quarter of cases, parents judged their child's development based on observations recorded before 18 months' corrected age. To this end, the extent of information bias in the reporting of neurodevelopmental outcomes by parental questionnaires has not been investigated.

2.6.5 Electronic health records

The use of electronic health record systems have been associated with improved delivery of care (Adams 2003). In the UK, most community child health services hold clinical information from child health surveillance programmes on electronic information systems, although these systems vary from one NHS trust to another and the data are not routinely passed back to the neonatal units for children born preterm. In the past decade, all neonatal units in the UK have moved towards routinely recording clinical data in an electronic record to facilitate shared care within neonatal networks. The BadgerNet, previously known as the Standardised Electronic Neonatal Database, is the most widely used platform (<http://www.clevermed.com/BadgerNet-Platform>). The data in these electronic records are based upon standard definitions developed by the BAPM (British Association of Perinatal Medicine Working Party 1997). In 2007, a standardised format for the recording of two-year neurodevelopmental and health status, adapted from the NPEU/Oxford classification of disability, was developed by the Thames Regional Perinatal Group (TRPG) Outcomes Group. This format was incorporated into the database in 2008, providing an electronic platform to record the results of routine neurodevelopmental assessment at two years that was in direct linkage with neonatal data. Since 2009, the NNAP, delivered by the RCPCH, has been using electronic data from neonatal units to audit two-year health status of children born at less than 30 weeks gestation and has reported its findings in annual publications (National Neonatal Audit programme). The programme has promoted outcome data collection activity with an increase in the number of participating neonatal units documenting any two-year outcome data on the eligible infants from 51 out of 170 units (30%) in 2009 (Watkinson 2009) to 158 out of 179 units (88%) in 2013 (Oddie 2014).

2.6.6 The UK National Neonatal Research Database

In 2007, the Neonatal Data Analysis Unit (NDAU) was established at Chelsea and Westminster NHS Foundation Trust and Imperial College London. It is an academic unit led by a Steering Board of key stakeholders, including clinicians, nurses, managers, academics, and representatives from the National Perinatal Epidemiology Unit, the RCPCH, the BAPM and Bliss (national newborn charity supporting parents). Following public consultation, the NDAU defined the 'Neonatal Dataset' of approximately 400 items, comprising demographics, daily care processes, medications, clinical parameters and outcome measures that are extracted from real-time electronic patient records containing point-of-care clinician entries held on BadgerNet. In 2013, the Neonatal Dataset received Health and Social Care Information Centre approval as an NHS Information Standard (ISB1595).

The National Neonatal Research Database (NNRD) contains the variables from the Neonatal Dataset after the data had been checked for duplicates, internal inconsistencies and outliers. It is managed by the NDAU and regulatory approvals had been received from the National Research Ethics Service (10/H0803/151), the Health Research Authority Confidentiality Advisory Group (8-05(f)/0210) as well as from the Caldicott Guardians and lead clinicians of participating NHS Trusts for its use in health service evaluations and approved research. The NNRD is unique in that up-to-date information is recorded in an approach that imposes no additional burden upon NHS staff. A UK Neonatal Collaborative comprising of NHS Trusts that contribute to the NNRD was formed and as of January 2014, this consisted of all 176 neonatal units in England and Wales and 13 of 15 neonatal units in Scotland. The NNRD is updated quarterly and is linked to the Office of National Statistics. This powerful resource that has complete population coverage of all neonatal unit admissions could potentially be an easily accessible source of national neonatal outcome data.

2.7 NEONATAL OUTCOME REPORTING IN THE UK

2.7.1 Recommendations for neonatal outcome reporting

Since 1992, several official reports have highlighted the need for data collection on the later morbidity of survivors of neonatal intensive care (House of Commons Health Committee Session 1991-2 1992, Audit Commission 1993, Clinical Standards Advisory Group 1993, Cumberledge 1993). In particular, the Audit Commission (Audit Commission 1993) proposed a national data collection exercise with all neonatal units collecting data in a nationally agreed format. The Cumberledge Report (Cumberledge 1993) recommended the development of “a system of data collection which can allow meaningful comparison of statistics relating to perinatal care”.

The BAPM published a set of standards for hospitals providing neonatal services in 2001, stating that “the later health status of survivors at particular risk of disability should be ascertained up to at least a corrected age of two years and the use of standardised guidelines for the definition of disability (based on the NPEU/Oxford classification) is recommended (British Association of Perinatal Medicine 2001). The BAPM/RCPCH 2007 working group further refined the recommendations for neonatal services to carry out follow-up evaluations at two years of age (corrected for prematurity) for all children born before 32 weeks gestation or with birth weight less than 1500g. This criteria is consistent with the RCPCH/ Royal College of Ophthalmology National Guidelines for Screening for Retinopathy of Prematurity. The NICE guideline for the developmental follow-up of preterm infants is currently being developed and is anticipated to be published in full in 2017 (NICE).

Despite these official recommendations, a robust system for routine outcome reporting of health outcomes following preterm births remains largely unavailable. In 2007, the National Audit Office reported that evidence of neonatal outcomes, other than the traditional indicator of mortality rates, was still sparse (The National Audit Office study team 2007).

2.7.2 Current status of neurodevelopmental follow-up assessment and reporting in the UK

Most neonatal units in the UK do offer routine clinical follow-up for high risk infants, including those born very preterm, to a corrected age of two years (information from correspondence with members of the TRGP group which had conducted a national survey) but the proportions of eligible infants that actually receive the assessment is unknown. In the past few years, there had been efforts to encourage neonatal outcome assessment and data collection. In addition to the publication of the NICE quality statement (section 2.6.2) and the impetus provided by the NNAP for data collection at a national level (section 2.6.5), some neonatal networks have, at regional and local levels, included a target for two-year assessment of very preterm infants in the Commissioning for Quality and Innovation (CQUIN) payment framework to encourage post-discharge follow-up and data collection.

Despite the recommendations from the BAPM, the approach to the assessment of neurodevelopment during routine clinical follow-up varies widely throughout the country. Some neonatal services base their follow-up criteria on gestational age at birth, some on a birth weight cut-off and some use both criteria. Children may be assessed by a neonatal or community paediatrics consultant or a doctor in a staff, associate specialist or training grade. It is also possible that the assessment is performed by an Advanced Neonatal Nurse Practitioner or an occupational or developmental therapist. There is no standardisation of the assessment methods used and many children are assessed based on informal clinical judgment. Where, how and what findings from the assessment are documented are also inconsistent although the NNAP do provide a framework for standardised data recording on the electronic platform. Overall, therefore, whilst the NNAP and NNRD had demonstrated feasibility of the electronic records as a national follow-up database, the utility of the data collected is unclear.

CHAPTER 3

AIM, HYPOTHESES AND OBJECTIVES

3.1 AIM

The purpose of this thesis is to evaluate the validity and usability of the reported neurodevelopmental outcome information of children born very preterm based on the current practices of follow-up and assessment in clinical and research settings.

3.2 HYPOTHESES

Main hypothesis for study 1:

The neurodevelopmental outcome data of children born preterm as determined through routine National Health Service (NHS) follow-up assessments at age two years is valid in identifying children with neurodevelopmental impairment and classifying them into levels of severity.

Main hypothesis for study 2:

Children born very preterm exhibit greater social-communication difficulties (as measured using the parent-completed Quantitative Checklist for Autism in Toddlers) than the general population at age two years.

Main hypothesis for study 3:

Early developmental assessments are poor at predicting the presence of cognitive impairment at school-age in children born very preterm.

3.3 SPECIFIC OBJECTIVES

Study 1:

1. To compare the agreement of outcome data obtained from routine National Health Service (NHS) follow-up assessments and data collected through a formal neurodevelopmental assessment that I conducted to research standards using the Bayley Scales of Infant and Toddler

Development, third edition (Bayley-III), in identifying children born preterm with neurodevelopmental impairments at age two years and classifying them into categories of neurodevelopmental status.

2. To examine factors that influence the agreement between routine NHS and research neurodevelopmental follow-up data.

Study 2:

3. To characterise the early social-communication skills and autistic-like traits in children born very preterm at age two years using the parent-completed Quantitative Checklist for Autism in Toddlers (Q-CHAT) questionnaire.
4. To compare the Q-CHAT scores of the preterm cohort with the reported scores of the general population.
5. To examine the associations of Q-CHAT scores with neonatal and sociodemographic factors as well as the Bayley-III cognitive, language and motor scores.
6. To determine the proportions of children born very preterm that would be classified as being at 'high risk' for autism spectrum disorder by the Q-CHAT and the Bayley-III Social-Emotional questionnaires and to compare the agreement between the two questionnaires.

Study 3:

7. 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 one and three years, for school-age cognitive deficit in children born very preterm or very low birth weight.
8. To assess the quality of the studies included in the systematic review.
9. To determine the meta-analytic (pooled) sensitivity and specificity of early developmental assessment in predicting school-age cognitive deficit in children born very preterm or very low birth weight from all available data.

CHAPTER 4

METHODS AND MATERIALS

4.1 STUDY DESIGNS

I conducted three studies to meet the objectives. Study 1 was a cross-sectional study that compared the validity of neurodevelopmental data collected during routine NHS assessments to data collected through a research assessment employing a standardised assessment tool (Bayley-III scales). Study 2 was a comparative analysis of the early social-communication skills displayed by preterm children against that of the general population as measured on the parent-completed Q-CHAT questionnaire. Data collected from participants recruited for study 1 were used in the analyses for study 2. Study 3 was a systematic review and meta-analysis on the predictive validity of early developmental assessment on identifying cognitive deficit at school-age.

4.2 RESEARCH ETHICS COMMITTEE APPROVAL/ RESEARCH DATABASE INCLUSION

Studies 1 and 2 received approval from the Royal Free Hospital Research Ethics Committee (REC 10/H0720/35) on 29th April 2010. Research ethics committee approval was not required for systematic review and meta-analysis (study 3). Study 1 was adopted into the UK Clinical Research Network Portfolio (ID 8626) and study 3 was registered on PROSPERO, an international prospective register of systematic reviews (CRD42012002168).

4.3 STUDY SITES

There were 13 hospitals where study participants were recruited from and where I conducted the research assessment for Study 1. These study sites are listed as follow (letter assigned to denote hospital in study):

- Addenbrooke's Hospital, Cambridge (A)
- Chelsea & Westminster Hospital (B)

- Ealing Hospital (C)
- Hillingdon Hospital (D)
- Homerton University Hospital (E)
- Newham Hospital (F)
- North Middlesex University Hospital (G)
- Northwick Park Hospital (H)
- Queen’s Hospital, Romford (I)
- Royal London Hospital (J)
- St Thomas’ Hospital (K)
- West Middlesex Hospital (L)
- Whipps Cross Hospital (M)

The study sites were selected to provide a representation of infants from a wide range of ethnic and socio-economic backgrounds as well as to include units where clinicians of different grades and specialties follow-up preterm infants. Study sites were restricted to within Greater London and Cambridge to maintain practical travel distances to the hospitals for the conduct of the research assessments. According to the London Perinatal Networks Annual Report for 2008, 945 infants were born at less than 30 weeks gestation and discharged alive from all London neonatal units that year (London Perinatal Group 2008). As the participating units included approximately 50% of neonatal unit admissions in London, I estimated that a study population of the desired sample size (section 4.11) could be recruited from these study sites during the two-year recruitment period. At the start of the study, I approached the lead consultant responsible for post-discharge follow-up at each hospital to be the local collaborator for the study. No hospital declined participation. Approval from the local NHS Research & Development department was sought prior to the commencement of research activities at each study site.

4.4 STUDY PARTICIPANTS

Eligible participants were children born at less than 30 weeks gestation who had attended or were going to attend routine NHS follow-up assessments at the participating hospital sites between the corrected ages of 20 and 28 months (age adjusted for prematurity, calculated as months from

expected due date of birth) during the recruitment period (June 2010 - July 2012). Exclusion criteria included children who had received Bayley-III assessment, either as part of their routine NHS assessment or due to enrolment in other research studies, in order to prevent 'practice effect' bias from repeated testing. A test-retest interval of 6 months is recommended for the Bayley-III (Bayley 2006b). Moreover, the inclusion of children who were tested with the same tool for both the routine and research assessments would lead to greater agreement between reported outcomes. Children from non-English speaking families whose parents require interpretation of the English language were also excluded for the following reasons: (i) no provision for English translation was available to ensure that informed consent was obtained; (ii) it was not possible to conduct the Bayley-III neurodevelopmental assessment, which was developed to be administered in English, reliably and (iii) the parents would not be able to complete the Q-CHAT and Bayley SE questionnaires independently.

4.5 RECRUITMENT

Recruitment to study 1 occurred between June 2010 and July 2012. In order to respect patient confidentiality, eligible participants were identified and approached in the first instance by the local collaborators (i.e. the clinical consultants). I requested that the local collaborators regularly identify all eligible participants with routine neurodevelopmental follow-up appointments scheduled within the following 3 months. The parents of these children were sent the study information sheet and a letter of invitation to participate in the study by post. They were asked to provide their contact details on a pre-printed response form and send it to me in a pre-paid envelope if they were interested in participating or would like to discuss the study in detail. Alternatively, the parents were given the study information and the response form at the time of the NHS follow-up appointment. I followed-up any response by phone and if the parents agreed to participate, I would organise an appointment to conduct the research assessment.

I took several steps to promote recruitment during study. At the start of the study, I gave a presentation of the aims and the recruitment process to all staff involved in neonatal post-discharge follow-up at each study site. I maintained regular contact with the local collaborators to encourage continual recruitment activity. Mid-way through the recruitment phase, I requested and received administrative support from two newly appointed research nurses at Northwick Park and St Thomas' Hospitals in sending out the letters of invitations to the parents of eligible participants. Finally, as much as possible, I attended the scheduled neurodevelopmental clinics at each study sites to have the opportunity to speak in person to the parents of any eligible child who had not responded by the time of their routine NHS appointment. This was very time-consuming as the average return journey time to each study site was 3 hours. However, it was the most crucial step in achieving consistent recruitment and more than 60% of participants were recruited by me meeting and speaking to the parents face-to-face.

I obtained informed written consent from the parents of all participating children. The final study population was a sample of the original cohort of infants discharged from the study sites. Non-participants would include children who had died or were lost to follow-up, children who were missed in the recruitment process or those whose parents did not respond or declined participation and children who were deemed ineligible for the study. I did not have access to data to determine the numbers of non-participants.

4.6 TRAINING TO CONDUCT THE RESEARCH ASSESSMENT

I received training on Bayley-III assessment techniques through attendance at the two-day training workshop (Cambridge, 21st-22nd January 2010), followed by practice sessions supervised by Bayley-III expert trainers (Ms Betty Hutchon, Consultant Neurodevelopmental Therapist, North Central London Perinatal Network and Dr Angela Huertas-Ceballos, Consultant Neonatologist, University College

London Hospital). My assessment techniques were accredited through a pilot assessment that was independently scored by Dr Huertas-Ceballos. I achieved 100% agreement on all items on the assessment scales with her. To ensure reliability and consistency during the course of the study, I attended validation sessions with Ms Hutchon who observed and independently scored assessments that I administered on non-study participants who were born at less than 30 weeks gestation and were 20-28 months old (corrected age) at the time of the assessment. The inter-observer agreement between our scores was evaluated and I received feedback on my assessment skills and on any difference in scores.

I also received training to perform a standardised neurological examination based on the Hammersmith Infant Neurological Examination (Haataja 1999) through practice sessions supervised by expert trainer (Professor Frances Cowan).

4.7 THE RESEARCH ASSESSMENT

At the time of assessing the participant, I was blinded to the results from the NHS assessment. I designed a data collection form (Appendix 4) and recorded the participant's demographic information including date of birth, gestation at birth, sex, ethnicity, singleton or multiple pregnancy, maternal age at birth, current post code and languages spoken at home as well as the presence of any hearing or vision problem as reported by the parents.

The research assessments took place in an outpatient clinic room at the same site as the routine NHS assessments. Each participant was accompanied by one or both parent(s) or caregiver throughout the assessment. For the Bayley-III assessment, the participant was either seated on an appropriately sized chair at a children's table or on his/her caregiver's lap at the office desk and I administered the test items by sitting across the table facing the participant. For twins or triplets, I assessed one

participant at a time and only the child being tested was invited into the clinic room. Each research assessment took 1.5 to 2 hours to complete.

4.7.1 Timing of research assessment

I aimed to conduct the research assessment within one month before or after the participant's NHS follow-up assessment. To minimise potential information bias due to changes in development during the interval between the NHS and the research assessments, I had aimed for approximately half the cohort to receive the research assessment before their routine assessment and the other half to receive it after their routine assessment. In practice, most of the participants recruited through a postal response had the research assessment prior to or on the same day as their routine NHS assessment whereas participants who were approached and recruited during their routine assessment had the research assessment on a later date arranged at the convenience of the parents. I requested the use of outpatient clinic rooms on an ad hoc basis whenever a participant was recruited. Hence, the timing of the research assessment was also largely dependent on the availability of the clinic rooms.

4.7.2 Assessment of cognition, language and neuromotor development

Bayley-III scales and subtests

Participants' cognitive, language and motor development were assessed using the Bayley-III (Bayley 2006b). The Bayley-III cognitive scale includes items that assess exploration and manipulation, sensorimotor development, object relatedness, conception formation and other aspects of cognitive processing. The language scale consists of separate receptive communication and expressive communication subtests. The receptive communication subtest assesses preverbal behaviour and verbal comprehension, such as the ability to identify specific objects and pictures when asked. The expressive communication subtest assesses preverbal communication, such as babbling, gesturing

and joint referencing, vocabulary development (e.g. naming objects) and morpho-syntactic development (e.g. using two-word utterances). The motor scale is separated into fine motor and gross motor subtests. The fine motor subtest measures functional hand skills, reaching, grasping, visual tracking, and object manipulation skills. The gross motor subtest measures static positioning (e.g. sitting and standing) as well as dynamic movement, balance, coordination and motor planning. For all Bayley-III scales, test items are ordered by difficulty, following the typical developmental pattern in children.

Administration of the Bayley-III assessment

I adhered strictly to the administration rules stated in the Bayley-III Administration Manual (Bayley 2006a). When appropriate, I used the suggested anglicised version of test items (e.g. use of the term 'biscuit' rather than 'cookie') listed in the Bayley-III UK and Ireland Supplement Manual (Bayley 2010). There are designated test items (start points) to commence Bayley-III assessment based on the participants' corrected age. The basal level for each scale was determined by 'passing' three consecutive test items from the start point. If the participant 'failed' any of the first three test items from the designated start point, testing was re-commenced from one start point below the age-designated start point. Test items were then administered sequentially until the ceiling level was reached when the participant could not complete five consecutive items. Certain test items were scored through incidental observation of the skills demonstrated by the participant during the appointment. I allowed accommodations during the assessment of participants with physical disability to maximize the opportunity for the demonstration of the skill being measured and reduce any non-relevant impact of the disability on the performance of the child. In children with hearing impairment, additional hand gestures were sometimes used to support verbal instructions. Participants with motor impairment were positioned using adaptive equipment. Certain items were scored as 'passed' if an indication of understanding and intent was demonstrated (e.g. clearly

looking at the correct response). Even with these accommodations, I could not complete the assessment for one participant with severe ataxic cerebral palsy.

Overall, I found the Bayley-III assessment to be child-friendly and it was relatively easy, using the test items, to develop a rapport with the participant at the start of the assessment. However, there were a few test items which I found hard to elicit a response from the participant or the response could be equivocal and difficult to interpret. For example, to assess for the development of representational play in the cognitive scale, I would use a block as a bar of soap and demonstrate 'giving the doll a bath' in accordance with the Bayley-III administration rules. The understanding of my action is clearly dependent on prior experience of the participant. I found that most participants were unfamiliar with the use of bar soaps since bath or shower gels are more commonly used. In the receptive communication subtest, I found it tricky to judge if a participant understood inhibitory words as the expected positive response of pausing during a play routine when I said 'stop' could be very subtle. Furthermore, significant attention and motivation from the participant was necessary for the completion of the assessment. In my experience, the participants generally found the test items for the cognitive and motor scales interesting but it was challenging to engage the participants in looking at a picture book for the purpose of testing their language ability.

Although children whose parents required interpretation of English were deemed ineligible for this study, there were participants who were raised in a bilingual or predominantly non-English speaking environment and selectively spoke a different first language. The Bayley-III Technical Manual and Administration Manual do not declare any potential limitations in assessing children with limited English proficiency. I would expect a lack of English proficiency to have little impact on the scores for the cognitive and motor scales as many of the test items do not fully depend on understanding of verbal instructions, for example placing pegs on the pegboard, matching shape puzzles, scribbling on a piece of paper and kicking a ball. Nevertheless, I am aware that participants learning English as a

second language may respond incorrectly to test items, particularly on the language scale, due to a lack of exposure to English rather than having delayed language development. In these cases, I made no modification to the administration and scoring of the Bayley-III assessment but made a note of their language ability based on my observation of the communication between the participants and their caregivers and by conferring with the caregivers. Whilst it was possible to involve the caregiver in translating instructions, I would not be able to control potential bias and variations in the translated instructions and the administration of the test could not then be considered standardised.

Bayley-III assessment scores

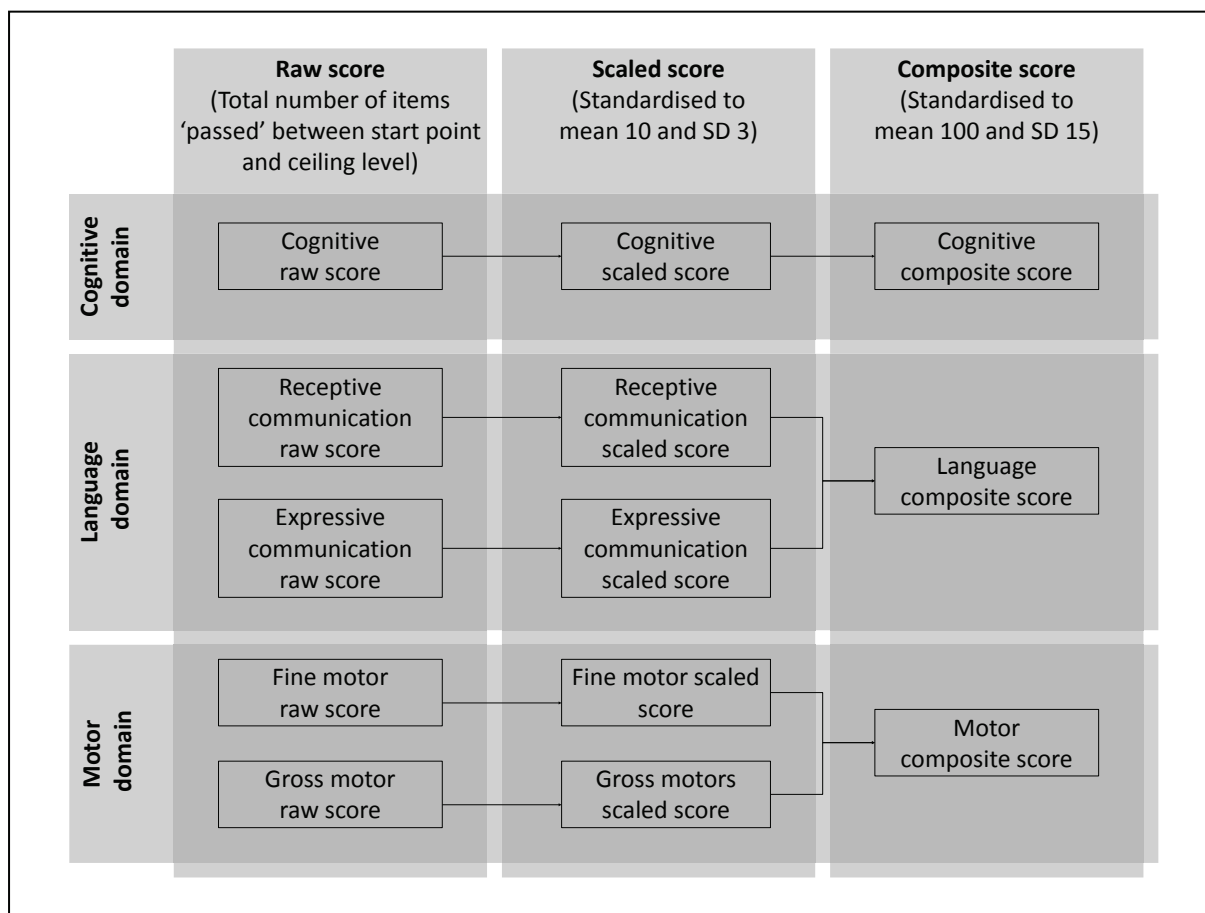
Each test item was scored as 1 (pass) or 0 (fail). If the participant refused to respond to any test item, it would be scored as 'failed' as it was difficult to determine if such refusal represented a lack of ability or if the participant was unwilling to cooperate for whatever reason. For each scale, the sum of the scores for all items tested between the basal and ceiling levels constitute the participant's raw score. Two types of norm-referenced scores were obtained: scaled scores, which are standardised to a mean of 10 and SD of 3; and composite scores, which have a mean of 100 and SD of 15. For the cognitive scale, both the scaled score and the composite score were derived from the raw score (figure 4.1). The language composite score was derived from the sum of the receptive communication and expressive communication scaled scores. Similarly, the motor composite score was obtained from the sum of the fine motor and gross motor scaled scores. The 95% confidence intervals (CI) for the all scaled and composite scores as well as the participants' percentile ranks for each scale were also obtained.

In view of the concerns that the Bayley-III scores over-estimated neurodevelopmental outcomes compared with the Bayley Scales of Infant Development, second edition (BSID-II) (Bayley 1993), I used the algorithm developed by Moore et al (Moore 2012c) to convert the Bayley-III cognitive and

language scores into a predicted BSID-II MDI, for the purpose of comparing the classification of neurodevelopmental outcomes into categories of severity based on the two scores. The algorithm is:

$$\text{Predicted BSID-II MDI} = 88.8 - (61.6 \times (\text{Bayley-III language composite score}/100)^{-1}) + (0.67 \times \text{Bayley-III cognitive composite score})$$

Figure 4.1 Types of Bayley-III scores



4.7.3 Assessment for neurological deficits and cerebral palsy

Hammersmith Infant Neurological Examination

I performed a standardised neurological examination, based on the HINE (Haataja 1999). The examination includes 26 items that assess cranial nerve function (5 items), posture (6 items), movements (2 items), tone (8 items) and reflexes (5 items). Each item can be individually scored on a

scale from 0 to 3 according to pre-defined criteria, 3 being the optimum score. Scores in each subsection were added up to give a global optimality score ranging between 0 and 78. Suboptimal scores were defined as <15 for cranial nerve function, <16 for posture, <6 for movements, <22 for tone, <13 for reflexes and <74 for global score. These cut-off scores were the 10th percentile of scores achieved by a validation population of normal children examined at 18 months of age.

ELGAN algorithm for cerebral palsy

I used the ELGAN algorithm as a structured guide to diagnose and classify cerebral palsy into topography-based categories of quadriplegia (at least 3-limbs involvement), diplegia (involvement of one or both lower limbs) and hemiplegia (involvement of one side of the body). I had opted not to use the SCPE framework for diagnosing cerebral palsy as it was designed to be used in children aged 4 years and older.

Gross Motor Function Classification System

The functional severity of cerebral palsy was also classified into 5 levels based on the Gross Motor Function Classification System (Palisano 1997, Palisano 2006, Palisano).

4.7.4 Assessment of social-communication skills

The social-communication abilities of the participants were judged using the parent-completed Q-CHAT (Allison 2008) and Bayley-III SE (Bayley 2006b) questionnaires. The parents of all participants were sent the questionnaires (combined into a single booklet as shown in Appendix 5) prior to the appointment for the research assessment. If the questionnaires were not completed by the time of the appointment, I gave the parents a further copy of the questionnaires at the appointment and asked them to return it by post. I requested that the parents complete the questionnaires

independently at home rather than at the appointment to avoid the tendency by the parents to seek my opinion, which could potentially introduce information bias.

Q-CHAT

The Q-CHAT consisted of 25 items that assessed children on targeted autistic-like behaviours expressed during toddlerhood. Each of the 25 questionnaire items was scored using a 5-point Likert scale (0 to 4 points) with higher scores indicating a higher frequency of autistic-like behaviour. All returned Q-CHAT questionnaires were scored according to the methods described by the research team who developed it (Allison 2008). Responses that were ambiguous or incomplete were scored 0, in accordance with the conservative approach adopted by the developers of the Q-CHAT. Questionnaires with more than six incomplete responses were excluded. The scores from all items were summed to obtain a total Q-CHAT score within a possible range of 0 to 100. To examine different aspect of autistic-like behaviour, I classified the Q-CHAT items into categories that explored social-relatedness (9 items), restricted, repetitive, stereotyped behaviour (9 items), communication (4 items) and sensory abnormalities (3 items), based on the nature of the questions.

Bayley Social-Emotional questionnaire

On the Bayley-III SE questionnaire, parents were asked to rate how often their child demonstrated certain behaviours described. The questionnaire items were ordered according to the anticipated trajectory of social-emotional development and the number of questions the parents were asked to complete were based on the corrected age of the participant. Scores for each item were allocated according to behaviour frequency as follow: all of the time (5 points), most of the time (4 points), half of the time (3 points), some of the time (2 points), none of the time (1 point) and can't tell (0 point). Once again, a conservative approach to allocation of scores was adopted. A score of 0 (equivalent to 'can't tell') was given to questions with incomplete responses; if more than one response was given, the response with the highest score was used. The Bayley-III SE composite score,

standardised to mean of 100 and SD of 15, takes into account the participant's corrected age at the time of assessment, and was obtained by referring the total raw score on the questionnaire to the conversion table in the Bayley-III Administration Manual (Bayley 2006a). Higher Bayley-III SE composite scores represent more advanced social-emotional development and hence, a lower likelihood for ASD.

4.7.5 Record of observed behaviour during the research assessment

To examine how participants' behaviour during the research assessment affected the study findings, I used the Behavioural Observation Inventory included in the standard Bayley-III record to document behaviour observed during the assessment. Thirteen types of behaviour were noted: positive affect, enthusiasm, exploration, ease of engagement, cooperativeness, appropriate activity level, adaptability to change, alertness, distractibility, appropriate motor tone, tactile defensiveness, fear or anxiety and negative affect. Numerical scores were assigned for each behaviour: 2 if the behaviour was 'observed most of the time', 1 if 'observed some of the time' and 0 if 'never or rarely observed'. The presence of 'distractibility', 'tactile defensiveness', 'fear/anxiety' and 'negative affect' were reverse-scored. Using the same form, the parent(s) or caregiver accompanying the participant was asked to rate how much the child's behaviour during the assessment was representative of his/her usual conduct. A score of 2 points was given for 'very typical (child is like this most of the time)', 1 for 'somewhat typical' and 0 for 'not at all typical'. Hence, 2 behavioural rating scores, each with maximum score of 26, were obtained – an examiner rated behavioural score for the frequency of positive behaviour and a parent rated score for the typicality of behaviour.

4.8 CLASSIFICATION OF IMPAIRMENT FROM THE RESEARCH ASSESSMENT

Using data from the research assessment, participants were classified into categories of neurodevelopmental status using two methods:

- (i) into SD score groups 'higher than -1 SD', '-1 to -2 SD' and 'lower than -2 SD' based on their Bayley-III scores (table 4.1);
- (ii) into 'no', 'mild-moderate' and 'severe' impairment groups according to the NPEU/Oxford criteria that was used for the NHS data and described in detail in sections 4.9 and 4.10.

4.8.1 Classification of impairment based on Bayley-III scores

The Bayley-III composite score was used to assign the SD score group in the cognitive domain (table 4.1). In the language and motor domains, composite scores were derived from combining scaled scores from the receptive and expressive communication subtests and the fine motor and gross motor subtests, respectively. Therefore, if a child had a specific impairment in only one subtest, it is possible for compensation from the other subtest to occur, resulting in a composite score within the normal range. Hence, the scaled score was used to identify specific impairment in the sub-domains of receptive communication, expressive communication, fine motor and gross motor. In the combined language and motor domains, impairment was taken as the worst category of outcome assigned in the respective sub-domains and based on the Bayley-III composite scores. The overall outcome of each participant was based on the worst category of impairment from the cognitive, language and motor domains.

Table 4.1 Method for assigning neurodevelopmental outcome using either the Bayley-III composite scores or the scaled score as follows:

SD score groups	Scaled score	Composite score	
Higher than -1 SD	≥7	≥85	(scores within 1 SD below standardised mean)
-1 to -2 SD	4-6	70 - 84	(scores between 1 and 2 SD below standardised mean)
Lower than -2 SD	<4	<70	(greater than 2 SD below standardised mean)

I considered participants who received Bayley-III scores lower than -1 SD to have at least a mild form of impairment and scores lower than -2 SD to represent at least moderate to severe impairment.

Using the same cut-off for composite scores as listed in table 4.1, I also examined how the classification of SD score groups changed when, instead of the Bayley-III cognitive and language scores, the predicted BSID-II MDI score was used instead.

4.8.2 Classification of impairment based on the NPEU/Oxford criteria

The NPEU/Oxford criteria for disability (National Perinatal Epidemiology Unit and the former Oxford Regional Health Authority 1994) adapted for use on the electronic ‘two-year outcome’ form on BadgerNet was applied to the NHS data to determine to determine the functional outcomes of high-risk infants (see sections 4.9 and 4.10). There are important methodological differences between the classification of neurodevelopmental status using Bayley-III scores and the NPEU/Oxford criteria that could account for observed differences in the outcomes. Firstly, the Bayley-III scores were standardised to take into account the participants’ corrected age at the time of assessment. The NPEU/Oxford criteria were designed to be used at 24 months corrected age and do not make allowance for participants younger or older than the designated age at assessment. Secondly, for the language domain, the Bayley-III assessment was based on verbal communication whereas the NPEU/Oxford criteria allowed the use of signs as a form of communication. Therefore, in order to have a direct ‘like-for-like’ comparison of agreement between the research and the NHS assessments, I also classified the participants, based on my judgment of their performance at the research assessment, into levels of functional outcomes using the NPEU/Oxford criteria. However, I only focused on the communication and motor domains as the outcome criteria in these areas were clearly defined (e.g. the number of meaningful words or signs and ability to sit or walk) and could be judged objectively from the research assessment. For the cognitive domain, the NPEU/Oxford criteria assigned outcomes according to ‘how far behind’ each child was deemed to be functioning.

For example a child considered to be functioning at an age-equivalent level of more than 12 months behind his/her corrected age would be classified as being severely impaired. This judgment is subjective. The Bayley-III does provide a chart to convert a participant's raw score into a 'developmental age equivalent' score. However, the developmental age equivalent score is an indication of the average age at which a given raw score is typical and does not specify the functional level of the child. Hence, I did not assign the cognitive outcome of the participants using the NPEU/Oxford criteria.

4.9 COLLECTION OF OUTCOME DATA FROM THE ROUTINE FOLLOW-UP ASSESSMENTS

The participants were assessed by their local clinicians as part of their routine NHS post-discharge follow-up. Their assessors were blinded to the results of the research assessment. Results of the NHS assessment for each participant were entered into the electronic 'two-year outcome' form on BadgerNet by as required for the National Neonatal Audit Programme (NNAP). In the participating hospitals, completion of the electronic 'two-year outcome' form was also one of the targets for the Commissioning for Quality and Innovation (CQUIN) payment framework. The 'two-year outcome' form was in a questionnaire format modified from the NPEU/Oxford criteria for disability, with each question yielding a 'yes/no' answer. The specific questions for the development (cognitive), communication and motor domains were:

Question reference

Development (Cognitive)

- D1 Is the child's development between 3-6 months behind corrected age?
- D2 Is the child's development between 6-12 months behind corrected age?
- D3 Is the child's development more than 12 months behind corrected age?

Receptive communication

- RC1 Does this child have difficulty with understanding outside of familiar context?
RC2 Is this child unable to understand words or signs?

Expressive communication

- EC1 Does this child have any difficulty with communication?
EC2 Does this child have difficulty with speech (<10 words/signs)?
EC3 Does the child have <5 meaningful words, vocalisation or signs?

Fine motor

- FM1 Does this child have any difficulty with the use of one hand?
FM2 Does this child have difficulty with the use of both hands?
FM3 Is this child unable to use hands (i.e. to feed)?

Gross motor

- GM1 Does this child have any difficulty walking?
GM2 Is this child's gait non-fluent or abnormal reducing mobility?
GM3 Is this child unable to walk without assistance?
GM4 Is this child unstable or needs to be supported when sitting?
GM5 Is this child unable to sit?

A positive response to any of the questions implied the presence of impairment. Questions D3, RC2, EC3, FM3, GM3 and GM5 denote the criteria for severe impairment. Additional information on whether the child was diagnosed with cerebral palsy, if a standardised neurodevelopment test was used during the NHS assessment and if the child was difficult to assess were also entered. The electronic form could be completed by the examining health professional or by administrators such as secretaries or data entry clerks based on the information given to them by the examiner.

4.10 RETRIEVAL OF DATA FROM THE ROUTINE ASSESSMENTS AND CLASSIFICATION OF DISABILITY

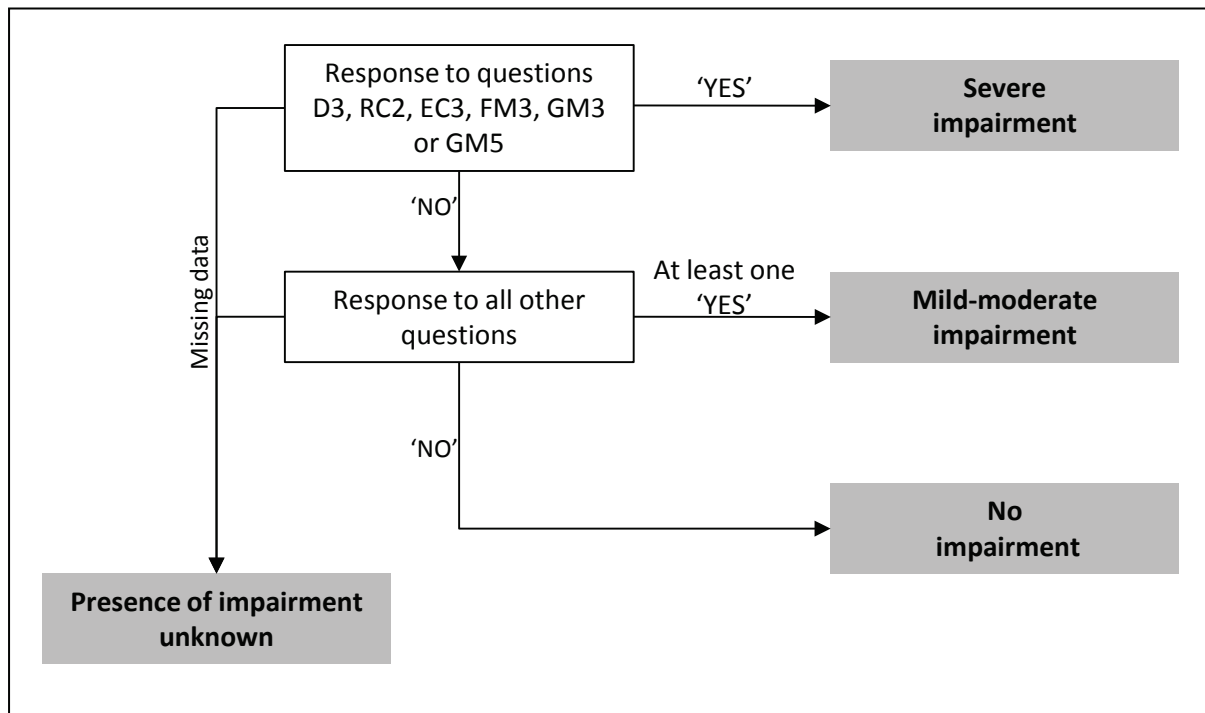
All the hospitals participating in this study are members of the UK Neonatal Collaborative which aims to support neonatal services and research through the release of anonymised patient-level electronic data, with permission from the Caldicott Guardians of NHS trusts, to the National Neonatal Data Unit (NDAU) where they are cleaned and merged to create the National Neonatal Research Database (NNRD) (see section 2.6.6). The NNRD is approved by the UK National Research Ethics Service (10/80803/151). Therefore, clinical information from the participants which is prospectively collected during their neonatal period as well as data captured from the 'two-year outcome' form was available on the NNRD. With consent from the parents, I obtained the participants' unique identifier on the NNRD from the local collaborator at each study site and extracted the neonatal and two-year outcome data for each participant with assistance from the NDAU data managers.

4.10.1 Classification of disability based on routinely recorded clinical data

Participants were classified into categories of 'no', 'mild-moderate' and 'severe' impairment within each outcome domain (cognitive, receptive communication, expressive communication, fine motor and gross motor) using the electronic two-year outcome data and according to the algorithm outlined in figure 4.2. The criteria used for classification was based on guidance from the Thames Regional Perinatal Group (TRPG) that had adapted the NPEU/Oxford classification for the electronic database.

A missing response does not count as a 'no'; therefore complete data entry is required to assign participants as having no impairment. An overall level of impairment was defined based on the worst outcome from the 5 domains.

Figure 4.2 Algorithm for the classification of impairment using data from NHS assessment



In addition, for the purpose of assessing selection bias, the following data were extracted from the NNRD for all infants born between 1 January 2008 and 31 December 2010, at gestational ages below 30 weeks, and discharged from the participating study sites (the 'baseline population'): gestation at birth, birth weight, sex, ethnicity, singleton or multiple pregnancy, mode of delivery, days of mechanical ventilation, oxygen therapy at 36 weeks' corrected gestational age, maternal age and the index of multiple deprivation (IMD) based on maternal residence at birth. Multiple deprivation relates to the concurrent occurrence of several forms of social and economic disadvantage. The IMD is a summary measure of relative area deprivation, calculated through a weighted combination of scores from 38 different indicators covering factors such as income, employment, education, health, living environment and crime for each area in England, using national census data. The IMD was obtained based on the post code of the mother at the time of birth of her child and according to the English Indices of Deprivation 2010 (Department for Communities and Local Government 2011).

4.11 STATISTICAL TESTS AND MEASURES USED

Data recorded on standardised forms and parent-completed questionnaires were encoded for analysis using Microsoft Office Excel 2007 (Microsoft Corp, Washington, USA). Data were double-entered, examined and outliers were verified to ensure accuracy. All analyses were performed using Stata statistical package version 11.0 (StataCorp, Texas, USA).

In general, quantitative variables are presented as means and standard deviations (SD) for normally distributed data or median and inter-quartile range (IQR) when the distribution of the data was skewed. Qualitative variables are presented as numbers of subjects (n) and percentages (%). Differences between categorical variables were analysed using the Pearson's chi-squared test. For continuous variables, the Student's t-test was used for parametric comparison and the Mann-Whitney U test was used for non-parametric comparison. *P*-values derived from statistical tests are presented and the conventional 5% level is used to define statistical significance. Several key statistical measures used in the analyses are described below.

4.11.1 Measures of test validity: sensitivity and specificity

The validity of an assessment, in the context of this thesis, refers to the ability of the assessment to accurately differentiate between children with and without neurodevelopmental impairment as defined. It is described using sensitivity and specificity, which are derived through a 2 x 2 table:

Assessment under evaluation	Reference 'gold-standard' assessment	
	Children with impairment	Children without impairment
Tested positive for impairment	True-positives (TP)	False-positives (FP)
Tested negative for impairment	False-negatives (FN)	True-negatives (TN)

Sensitivity is the proportion of children with impairment who were accurately identified as having impairment from the assessment under evaluation. It is calculated as

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Specificity is the proportion of children without impairment who were accurately identified by the assessment under evaluation and is calculated with the formula

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

Sensitivity and specificity calculations are expressed as either proportions or percentages with corresponding 95% CI. Values <0.7 (or 70%) were interpreted as low, 0.7 to 0.85 (70% to 85%) as moderate and >0.85 (>85%) as high (American Educational Research Association 1999).

4.11.2 Cohen's kappa statistic

Sensitivity and specificity can only measure the accuracy of an assessment from a binary classification. Hence, I used the Cohen's kappa statistic (κ) to compare the agreement in classifying neurodevelopmental outcomes into the three categories of 'no', 'mild-moderate' and 'severe' impairment (ordinal data). The κ coefficient is a measure of the proportion of agreement above that is due to chance alone and is calculated by

$$\kappa = \frac{\text{observed agreement} - \text{chance agreement}}{1 - \text{chance agreement}}$$

A κ value of 1 indicates perfect agreement and a value of 0 reflects agreement that is no better than by chance. Unweighted and weighted forms of κ coefficient were obtained. The purpose of the weighting was to derive a coefficient that provided a closer reflection of the clinical implications of disagreement between the ordinal categories. It is clinically more important to distinguish patients with impairments from those without impairment than to differentiate between the severity of 'mild-moderate' and 'severe' impairments. Hence, in the calculations for the weighted κ coefficient, discrepancy between 'mild-moderate' and 'severe' impairments was considered partial agreement.

The weighting matrix used was as follow:

Level of impairment based on Assessment I*	Level of impairment based on Assessment II*		
	None	Mild-moderate	Severe
None	1	0	0
Mild-moderate	0	1	0.5
Severe	0	0.5	1

Note: 1=full agreement, 0.5=partial agreement, 0=no agreement

*The Cohen's kappa statistic was used for comparisons where

- (i) Assessment I = classification using Bayley-III scores and Assessment II = classification based on the NPEU/Oxford criteria
- (ii) Assessment I = NHS assessment and Assessment II = research assessment

I interpreted the κ values according to the standards proposed by Landis and Koch: 0 - 0.4 = slight to fair agreement; 0.4 - 0.6, moderate agreement and 0.6 - 1, substantial to perfect agreement (Landis 1977).

4.12 SAMPLE SIZE CALCULATION

A precision analysis for the estimated sensitivity of the NHS assessment in identifying children with Bayley-III scores lower than -2 SD in study 1 was used to calculate the target sample size for recruitment. The desired sensitivity of a developmental test is conventionally between 70% and 80%. The precisions (widths of CI) of the observed sensitivity and specificity of a test vary depending on sample size and the observed estimates. I aimed for a sample size to achieve a precision of 95% CI half-width within 10% for the estimated sensitivity of identifying children with Bayley-III scores lower than -2 SD by the NHS assessment.

Based on the London Perinatal Networks 2008 Annual Report (London Perinatal Group 2008), I estimated that approximately 500 children were born <30 weeks gestation and survive to discharge

from the participating hospitals per year. Assuming that 10% of these children have Bayley-III scores lower than -2 SD, with an unstratified random sample, 650 participants would be required to achieve a CI half-width within 10% for an estimated sensitivity of 80%. Therefore, I attempted to recruit a stratified sample to include higher proportions of children with medium and high risk for impairment to improve the precision of the study while maintaining a practical sample size ((Obuchowski 2002); table 4.2).

Gestation at birth is the strongest predictor of neurodevelopmental impairment. Among infants born at less than 30 weeks gestation who survived to discharge in London in 2008, 20% were born at or less than 25 weeks (higher risk group), 30% were born at 26 to 27 weeks (medium risk group) and 50% were born at 28 to 29 weeks gestation (lower risk group) (London Perinatal Group 2008). Assuming 25% of higher risk, 15% of medium risk and 0% of lower risk children achieve Bayley-III scores lower than -2 SD, and the sensitivity of identifying different severity of impairment is the same for all risk groups, the table 4.2 shows various sample size options and the resulting CI half-width for different sensitivity estimates. I set out to recruit 500 children (200 from the higher risk group, 200 from the medium risk group and 100 from the lower risk group) over the two-year recruitment period.

Table 4.2 Precision of estimated sensitivity for different sample sizes and sensitivity estimates

Size of strata		Sample size	Estimated proportion with Bayley-III scores lower than -2SD	Estimated sensitivity	95% CI half-width
Unstratified sample		650	10%	80%	9.7%
Unstratified sample		650	10%	50%	12.2%
Higher risk	200		25%		
Medium risk	200	500	15%	80%	8.9%
Lower risk	100		0%		
Higher risk	200		25%		
Medium risk	200	500	15%	50%	11.2%
Lower risk	100		0%		
Higher risk	100		25%		
Medium risk	150	300	15%	80%	11.4%
Lower risk	50		0%		
Higher risk	100		25%		
Medium risk	150	300	15%	50%	14.2%
Lower risk	50		0%		

4.13 STUDY 1 ANALYSES: THE VALIDITY OF ROUTINE NHS ASSESSMENT

4.13.1 Examining the characteristics of the study population

In order to assess the effect of selection bias, I would ideally have compared the characteristics of the study population with the target population from which I was recruiting (i.e. eligible infants attending their routine NHS assessment at the corrected ages of 20 - 28 months). However, as recruitment depended on referral from the local collaborators and there was no information available on children lost to post-discharge follow-up or whose parents refused to participate, I was unable to determine the characteristics of the target population. Hence, to evaluate the representativeness of the study population, I compared neonatal and sociodemographic characteristics, with data extracted from the NNRD, of the study participants with non-participants from the 'baseline population' of infants born between 1 January 2008 and 31 December 2010, at gestational ages below 30 weeks, and discharged from the participating study sites. As raw IMD values are non-linear, for the purpose of analysis, IMD was categorised into quintiles based on ranking with IMD Quintile One presenting the least deprived 20% of areas in England.

4.13.2 Comparing the different methods of classification of impairments from the research assessment

Using graphical displays, I compared the differences in proportions of participants in each SD score group as per Bayley-III scores and the predicted BSID-II MDI in the cognitive and language domains.

I wanted to judge how comparable the 3 levels of impairment (none, mild-moderate and severe) based on the NPEU/Oxford criteria are to the 3 Bayley-III SD score groups of 'higher than -1 SD', '-1 to -2 SD' and 'lower than -2 SD'. To do this, using only the data obtained from the research assessment, I cross-tabulated the NPEU/Oxford levels of impairment against the Bayley-III SD score

groups and calculated the unweighted and weighted Cohen's κ coefficients to measure the concordance between using the two classification methods

4.13.3 Agreement in the classification of impairment between NHS and research assessments

For each neurodevelopmental domain, the classification of outcomes from the NHS assessment was cross-tabulated against the results from the research assessment (classified using both Bayley-III scores and the NPEU/Oxford criteria).

I considered any participant with Bayley-III scores lower than -1 SD to have at least mild impairment. Taking the research assessment to be the reference 'gold standard', the sensitivity and specificity of the NHS data in identifying children with *any* impairment were calculated. I then calculated the sensitivity and specificity of the *severe* impairment category in the NHS data for identifying children with Bayley-III scores lower than -2 SD.

The data were likely to be clustered by study sites because participants recruited from the same sites were more likely to share similar characteristics and to have received the same type of assessments by the same NHS assessor/ team of assessors. To account for this correlation, robust standard errors were used to calculate the 95% confidence interval for the estimated sensitivities and specificities. In addition, analyses were repeated on all singleton births and only one randomly selected child from each multiple birth set, to examine the effect of correlated outcomes within multiple birth sets.

Weighted and unweighted κ coefficients were used to measure the concordance between the research and the NHS assessments, again matching the 'no impairment' category to Bayley-III scores

higher than -1 SD, 'mild-moderate' to Bayley-III scores between -1 to -2 SD and 'severe' to Bayley-III scores lower than -2 SD.

Secondary post-hoc analysis to identify question sets with improved validity for identifying severe impairments

I am aware that the NPEU/Oxford expert group had suggested using a criterion of -3 SD scores to represent 'severe cognitive (developmental) disability' at age 2 years as it would be more predictive of later severe disability and impaired IQ (A report of two working groups convened by the National Perinatal Epidemiology Unit and the former Oxford Regional Health Authority 1994). However, this cut-off score was not feasible due to the floor effect of the Bayley-III cognitive composite scores, which ranged between 55 and 145. Besides, only one participant received the minimum composite score of 55. Therefore, I conducted a post-hoc analysis to evaluate if applying a broader criteria at the severe end of the impairment spectrum would improve the validity of the NHS data in identifying children with Bayley-III scores lower than -2 SD. Instead of classifying impairments into 'none', 'mild-moderate' and 'severe' impairment, I redefined the impairment categories into 'none', 'mild' and 'moderate-severe'.

Referring back to the list of questions on the electronic '2-year outcome form' in section 4.9, in addition to those that denote severe impairment, I re-categorised participants who received a positive response to the following questions into the 'moderate-severe' category:

- D2 Is the child's development between 6-12 months behind corrected age?
- RC1 Does this child have difficulty with understanding outside of familiar context?
- EC2 Does this child have difficulty with speech (<10 words/signs)?
- FM2 Does this child have difficulty with the use of both hands?
- GM2 Is this child's gait non-fluent or abnormal reducing mobility?

GM4 Is this child unstable or needs to be supported when sitting?

Participants who received a positive response to any of the other questions are classified as having mild impairment. I then calculated the sensitivity and specificity of the 'moderate-severe' category in predicting children with Bayley-III scores lower than -2 SD. I also calculated the concordance of the NHS data classified into these new categories with the Bayley-III SD score groups, matching 'no impairment' to Bayley-III scores higher than -1 SD, 'mild impairment' to Bayley-III scores between -1 to -2 SD and 'moderate-severe' impairment to Bayley-III scores lower than -2 SD.

4.13.4 Variables associated with the validity of NHS neurodevelopmental data

I examined the effect of the following factors on the validity of the NHS data: gestation at birth, sex, supplemental oxygen requirement at 36 weeks corrected gestational age, IMD quintile at the time of assessment, English as the only language spoken at home, corrected age at NHS assessment, use of a standardised neurodevelopmental test or screening test during NHS follow-up, grade of NHS assessor, time interval between NHS and research appointments, behaviour during the research assessments as measured by the examiner rated behavioural score and if the NHS assessor thought that the child was difficult to test during the NHS assessment. Cross-tabulations and the calculation of the sensitivities and specificities of NHS assessment, stratified by the factor under study, were performed for each domain of neurodevelopment.

4.14 STUDY 2 ANALYSES: SOCIAL-COMMUNICATION IN PRETERM CHILDREN

4.14.1 Examining the characteristics of respondents

For the purpose of assessing the applicability of the Q-CHAT for the majority of children born preterm, I excluded children with cerebral palsy and severe neurosensory impairments (defined as a

hearing deficit not correctable with hearing aids or a visual deficit not correctable with glasses) from this analysis to reduce the potential confounding effect of co-existing neurosensory and physical impairments on the association between preterm birth and Q-CHAT scores. Differences in characteristics between respondents whose parents returned the Q-CHAT questionnaire and non-respondents, and between respondents and the 'baseline population' were compared to evaluate selection bias.

4.14.2 Comparison of Q-CHAT scores between the study population and the general population

The overall and sex-specific Q-CHAT scores from the study population were compared with the published scores from the general population (general population overall mean 26.7, SD 7.8; mean for boys 27.5, SD 7.8; mean for girls 25.8, SD 7.7) (Allison 2008), using the Student's t-test. Differences in the distributions of item-specific scores between the study cohort and the general population in each category of autistic-like behaviour were examined by Chi-square tests. To overcome the Chi-square test restriction for low expected numbers, I combined the proportions in adjacent score categories to ensure that all expected values were larger than five (Cochran 1954).

4.14.3 Factors associated with Q-CHAT scores

I explored the correlation between the Q-CHAT scores and the Bayley-III cognitive, language and motor composite scores using linear regression to determine if any observed difference in Q-CHAT scores between the study population and the general populations were explained by delayed neurodevelopment in the preterm population. Post-hoc analysis of the correlation between sub-categorical Q-CHAT scores (total score from items within each category of autistic-like behaviour) and Bayley-III cognitive, language and motor composite scores was carried out with Bonferroni correction for multiple testing.

The following neonatal and sociodemographic factors were analysed for possible association with Q-CHAT scores: gestation at birth, birth weight z-score, sex, single versus multiple pregnancy, white versus non-white ethnicity, maternal age at birth, mode of delivery, length of mechanical ventilation, supplemental oxygen requirement at 36 weeks corrected gestational age and IMD quintile at the time of completion of the Q-CHAT. I had chosen to use the current IMD quintile the participant was living in rather than the one at birth because the data for latter were retrieved from the NNRD and could be erroneous. Comparing the IMD quintiles at birth and at the time of assessment, 177 (83.9%) participants continued to live within the same IMD quintile, 13 (9.2%) moved to a more deprived IMD quintile and 15 (10.6%) moved to a less deprived quintile. Linear regression models were created to determine the association between predictive variables with Q-CHAT scores. To account for correlated outcomes within multiple birth sets, I used cluster bootstrap to estimate standard errors and the resultant 95% CIs. Variables identified to be significant at a 5% level in univariable models were included in forward stepwise multivariable regression analyses to determine the independent effect of each factor on Q-CHAT scores. Post-hoc analysis was conducted to explore possible interactions between ethnicity, Bayley-III language scores and IMD.

4.14.4 Classifying children at risk for ASD using the Q-CHAT and Bayley-III Social-Emotional questionnaires

Both the Q-CHAT and the Bayley-III SE questionnaires were proposed as potential screening tools for early detection of ASD in toddlers although no formal 'cut-off' scores had been suggested. As described in section 2.4.4, the positive predictive values of available ASD screening tools are low and there is significant disagreement in the cases that screened positive among different tools. I therefore aim to determine the proportions of 'positive screens' using the Q-CHAT and the Bayley-III SE to allow for comparison with other tools described in the literature. Using an arbitrary cut-off score of 2 SD above the general population mean for Q-CHAT scores and 2 SD below the

standardised mean for Bayley-III SE scores, participants were classified as 'at risk for ASD'. I also wanted to explore the agreement between the Q-CHAT and the Bayley-III in describing children at risk for ASD. A scatterplot was used to examine the relationship in score distribution between the two questionnaires and the agreement between the questionnaires in identifying children 'at risk' was measured using Cohen's κ statistic.

4.15 STUDY 3: SYSTEMATIC LITERATURE REVIEW AND META-ANALYSIS

I conducted a systematic electronic literature search on MEDLINE to seek comprehensive data on the early developmental outcomes and corresponding school-age cognitive outcomes of preterm children to meet objectives 7, 8 and 9. The methods adopted in this review were based on recommendations outlined in the Cochrane Handbook for Diagnostic Test Accuracy Reviews (Deeks 2013). Results were reported in accordance with the 'Preferred Reporting Items for Systematic Reviews and Meta-Analysis' (PRISMA) Statement (Moher 2009).

4.15.1 Eligibility criteria for study inclusion

Any cohort or matched-control studies published since 1 January 1990 on study populations of infants born ≤ 32 weeks gestation and/or had birth weight < 1500 g (very low birth weight, VLBW), in which at least two serial assessments, consisting of a neurodevelopmental assessment conducted 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, GMDS, Wechsler Preschool and Primary Scale of Intelligence) were considered for inclusion in the review. I did not review assessments conducted before 1 year of age as impairment, particularly if mild, may not be evident at this stage. Studies with populations that did not meet the gestation or birth weight criteria or reported outcomes using non-standardised assessments including measures of academic attainment were excluded. 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. In addition, I excluded case reports, narrative reviews, editorials, letters and comments on published articles.

4.15.2 Data sources and search strategy

The electronic search was conducted on MEDLINE through the PubMed interface on 13th April 2012, covering English-language literature published between 1st January 1990 and 31st March 2012. The time period spanning more than 20 years would allow a reasonable number of studies to be included in the review. Studies published prior to 1990 were not included in order to focus the review on more contemporaneous preterm populations.

Search terms were selected *a priori* through a preliminary review of the literature. 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 ‘explode’ feature was used with subject headings to include articles categorised under more specific subheadings. The detailed search strategy was:

```
((("preterm children"[tiab] OR "premature children"[tiab]) OR ("premature infant"[tiab] OR ("preterm infant"[tiab]) OR ("preterm neonate"[tiab] OR "premature neonate"[tiab]) OR ("Infant, Premature"[MeSH]) OR ("Infant, Very Low Birth Weight"[MeSH]) OR ("very low birth weight"[tiab] OR "very low birthweight"[tiab]) OR ("extremely low birth weight"[tiab]) OR ("extremely low birthweight"[tiab])) AND ((cogniti*[tiab]) OR (neurodevelopment*[tiab]) OR (mental retardation) OR ("Developmental Disabilities"[Mesh] OR disability[tiab]) OR (intelligence[tiab] OR IQ[tiab]))
```

A summary of the full search process with the number of articles retrieved during each step is outlined in Appendix 6. The electronic search was then supplemented by a manual search of the reference lists of studies that met the inclusion criteria.

4.15.3 Study selection

The titles and abstracts of studies retrieved from the literature search were screened to identify studies that reported developmental and/or cognitive outcomes among preterm children born at less than 32 weeks gestation and/or VLBW. These articles were grouped into 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. The author lists for articles in groups (ii) and (iii) were matched in order to identify assessments and publications on the same population at different time points. Studies that satisfied the initial screening process were retrieved for full text evaluation for final inclusion in the review.

4.15.4 Study quality assessment

I assessed the quality of included studies using a checklist adapted from the Quality of Diagnostic Accuracy Studies version 2 (QUADAS-2) appraisal tool (Whiting 2011). The aim was to provide a qualitative judgment for the risk of bias and the applicability of each study to the review question. The QUADAS-2 tool use 'signalling questions' to assess 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. An essential feature of QUADAS-2 was the tailoring of the signalling questions to enable review-specific appraisal. Table 4.3 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. No summary 'quality score' was generated as such scores lack statistical justification and are not comparable across different scoring systems (Whiting 2005). I decided not to exclude any study on the basis of its quality to achieve a review on the topic that was as comprehensive as possible.

Table 4.3 Review-specific signalling questions and standards for appraisal of study quality

Domain	Patient selection	Index test (Early developmental assessment)	Reference standard (School-age cognitive assessment)	Flow and timing
Signalling questions	(1) Was a consecutive or random sample of patients enrolled? (2) Did the study avoid inappropriate exclusion?	(1) Was an age-appropriate validated standardised assessment tool used?	(1) Was an age-appropriate validated assessment standardised assessment tool used? (2) Were the assessors blinded to the results of the early developmental test?	(1) Was all eligible infants participants receive the same assessments? (2) Were all participants included in the analysis?
High risk of bias	Non-consecutive or random sampling methods; additional inclusion criterion based not on birth weight or gestational age	Inappropriate test used for population under study	Inappropriate test used for population under study or if assessors were not blinded to results of early developmental test	If participants receive different assessments or if drop-out rates >30%
High concerns regarding applicability	Subcohort of infants (e.g. only IUGR infants were included) recruited. Infants born before 1990 as they would differ to target population in terms of neonatal care received and severity/pattern of diseases experienced	Non-universal tests (e.g. only standardised in a specific population). Older versions of assessments (validated in normative populations that were no longer representative of contemporaneous populations)	Non-universal tests (e.g. only standardised in a specific population). Assessment tools that may not be representative of current populations (e.g. published before 1990)	

4.15.5 Data extraction and synthesis

From each included study, the following data were extracted, synthesised and systematised into a table. Unpublished data were sought from study authors through email requests.

(i) Data on study characteristics:

Study location (hospital(s), city, country)

Population sampling method (single-centre, multi-centres, population-based)

Inclusion and exclusion criteria

Participation and/or follow-up rates (as percentage of eligible survivors)

Final sample size to be included in meta-analysis (number of participants who completed both early and school-age assessments)

Early developmental and school-age cognitive assessment tool used

(ii) Data on study population characteristics:

Year(s) of birth of participants

Mean or median gestational age

Mean or median birth weight

Ages at early and school-age assessments

Mean test scores at early assessment and school-age assessments

(iii) Data on the predictive validity of early developmental assessments:

For this review, mild-moderate deficit was defined as developmental or cognitive test scores between 1 and 2 SD below the standardised or control group means. Severe deficit was defined as test scores lower 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 was used as the references for defining the presence of deficits. Data on the number of 'true-positive', 'false-positive', 'false-negative' and 'true-negative' cognitive deficits identified by early assessments were collated from each study. If serial assessments were performed at different time points, all available data were extracted although only data obtained from participants at the oldest age were included in the meta-analysis. The estimated sensitivity and specificity with corresponding 95% CI for mild-moderate and severe deficits were calculated.

4.15.6 Meta-analysis

The goals of the meta-analysis were to use statistical tools to (i) evaluate the variation in the estimates of the diagnostic accuracy (sensitivity and specificity) of early developmental assessments between studies and (ii) combine results from all studies to yield a more precise estimate than is possible from individual studies. I generated coupled forest plots to depict the ranges of sensitivity and specificity derived from the studies. Homogeneity of the sensitivities and specificities from the studies were tested using Chi-squared tests. It has been noted that, in meta-analyses of diagnostic tests, significant between-study heterogeneity often exists (Deeks, 2010). One source of heterogeneity is due to variations in diagnostic threshold and the related 'trade-off' between sensitivity and specificity (Moses 1993, Deeks 2013). This may occur even when the same diagnostic criterion was applied across the studies, as was in this review, due to for example, inherent differences in the spectrum of impairments in the patient populations or inter-observer interpretation of test performances. To examine for this, a scatterplot of the true-positive rate (TPR; or sensitivity) against the false-positive rate (FPR; or 1-specificity) for each study was created and the Spearman correlation coefficient was computed. 'Threshold effect' was demonstrated when the points assume the shape of a receiver operator characteristic (ROC) curve and the sensitivity and specificity were significantly correlated. In this circumstance, separate pooling of sensitivities and specificities which ignore the correlation between the two measures would lead to an

underestimation of the diagnostic accuracy (Deeks 2001). It is possible to combine estimates using the Moses-Littenberg method to generate a summary ROC curve (Moses 1993, Littenberg 1993). However, this does not allow for between-study variation. Instead, I used the Rutter and Gatsonis approach to fit a hierarchical summary ROC (HSROC) curve of the data (Rutter 2001). The HSROC model accounts for both sampling variation within study at a lower level and between-study heterogeneity at a higher level using random effects. It models the log odds of a positive test result in each study and each impairment group as a function of the positivity threshold in each study and the true impairment status, with model parameters describing the accuracy and asymmetry of the ROC curves. The output includes a summary operating point (pooled values for sensitivity and specificity) with 95% confidence region and a 95% prediction region for a forecast of the true sensitivity and specificity in a future study. As this is a hierarchical model, the summary operating point represents an average of study effects rather than a common effect. Individual study effects may differ considerably due to heterogeneity; this variation is represented by the 95% prediction region.

4.15.7 Investigating heterogeneity by meta-regression and subgroup analysis

I investigated the possible association of the diagnostic validity with study-level variables that could account for the observed heterogeneity among studies, using meta-regression methods for continuous variables and subgroup analysis for categorical variables. The variables were gestational age, birth weight, age at early assessment, age at late assessment, time interval between assessments, year of birth of participants, prevalence of total and severe impairment, the developmental assessment tool used and the inclusion/exclusion of neurosensory impaired participants. For categorical variables, couple forest plots stratified by the subgroups were generated to allow for visual assessment of the differences in diagnostic validity between subgroups.

For continuous variables, I generated scatterplots of sensitivity and specificity against each study-level covariates, taking the mean value for continuous variables within each study except for year of birth, where I used the earliest date as the mean/median value was not available. Bivariate models (Reitsma 2005) were used to test formally whether sensitivity and specificity were associated with study-level covariates. Bivariate models are equivalent to HSROC models when no covariate is included (Harbord 2007). When including covariates, the bivariate model measures the association with sensitivity and specificity (on the logit scale) while the HSROC model measure association with the accuracy and threshold parameters, therefore the former was chosen for ease of interpretation. For each study-level covariate, associations with sensitivity and specificity were tested separately; likelihood ratio test was then used to test both associations jointly. Results are reported as estimated odds ratios (OR) with associated 95% CI and p-values.

For the studies that had reported data from multiple assessments at different time points, I created scatterplots of sensitivity and specificity against mean age at assessment to explore the stability of sensitivity and specificity estimates over time.

4.15.8 Post-hoc analysis to examine the change in mean developmental/ cognitive scores over time

The diagnosis of developmental or cognitive ‘impairment’ is based on arbitrary cut-off points and the comparison of the association between categories of ‘developmental impairment’ and ‘cognitive impairment’ assumes that developmental and cognitive abilities follow the same scale. The feasibility of exploring the relationship between early developmental scores and school-age cognitive scores within this systematic review is limited by the lack of patient-level data (only mean developmental and cognitive scores of study populations were available). However, using the data available, I was able to examine the change in the standardised mean difference (SMD) in

developmental and cognitive scores over time. In the context of this review, the SMD is a summary statistic that expressed the difference in the mean developmental or cognitive scores between the study populations and the normative or control populations, relative to the variability observed. The SMD for each study population is calculated as:

SMD

$$= \frac{(\text{Mean developmental or cognitive score of study population}) - (\text{Mean developmental or cognitive score of normative or control populations})}{\text{Standard deviation of developmental or cognitive score of normative or control populations}}$$

The change in SMD in developmental and cognitive scores over time is calculated as:

$$(\text{Cognitive score SMD}) - (\text{Developmental score SMD})$$

The direction and magnitude of the difference between SMD of developmental scores in early childhood and SMD of school-age cognitive scores may provide an indication of the dimension of change in the two measurements over time. Due to the significant heterogeneity in the data, I did not perform a meta-analysis to pool together the differences in developmental and cognitive score SMD from individual studies.

4.15.9 Investigating publication bias

There is currently little understanding of the determinants and impact of publication bias on the reviews of screening and diagnostic test accuracy. For reviews of interventional trials, the funnel plot, a graphical display of the estimates of study effects plotted against their sample size or precision. Statistical tests such as Egger's regression test and Begg's rank correlation are used to test for funnel plot asymmetry which would indicate the presence of publication bias and other sample size-related effects. The appropriate method for investigating publication bias for studies of diagnostic test accuracy is unclear. Funnel plot of the estimates of log diagnostic odds ratio (DOR) against

corresponding precision was proposed (Song 2002). The DOR is a single statistic measure of diagnostic performance that is defined as:

$$\text{DOR} = \frac{\text{TP} \times \text{TN}}{\text{FP} \times \text{FN}}$$

Therefore, the larger the DOR, the more accurate is the test. In the Cochrane Handbook (Deeks 2013), the application of tests for funnel plot asymmetry designed for use in randomised trials, including the Egger's and Begg's tests, was specifically discouraged as these were associated with inflated type I error rates. Instead, a regressions test for the association between the log DOR and the 'effective sample size (ESS)' developed by Deeks et al (Deeks 2005) was suggested. The ESS is a function of the number of non-diseased (n_1) and diseased (n_2) participants, where

$$\text{ESS} = (4n_1n_2) / (n_1 + n_2)$$

Following the proposed methods outlined in the paper by Deeks et al (Deeks 2005), I investigated the possibility of publication and other sample size related effects by developing funnel plots of log DOR against $1/\text{ESS}^{1/2}$ and tested for plot asymmetry using linear regression of the two variables, weighted by ESS.

For this review, forest plots were generated using Review Manager (RevMan) version 5.2 (The Cochrane Collaboration, Oxford, UK). All other analyses were performed using Stata statistical package version 11.0 (StataCorp, Texas, USA) and SAS 9.3 (SAS Institute, North Carolina, USA).

CHAPTER 5

STUDY 1 RESULTS: VALIDITY OF STANDARDISED TWO-YEAR NEURODEVELOPMENTAL DATA COLLECTED DURING NHS FOLLOW-UP

5.1 STUDY POPULATION

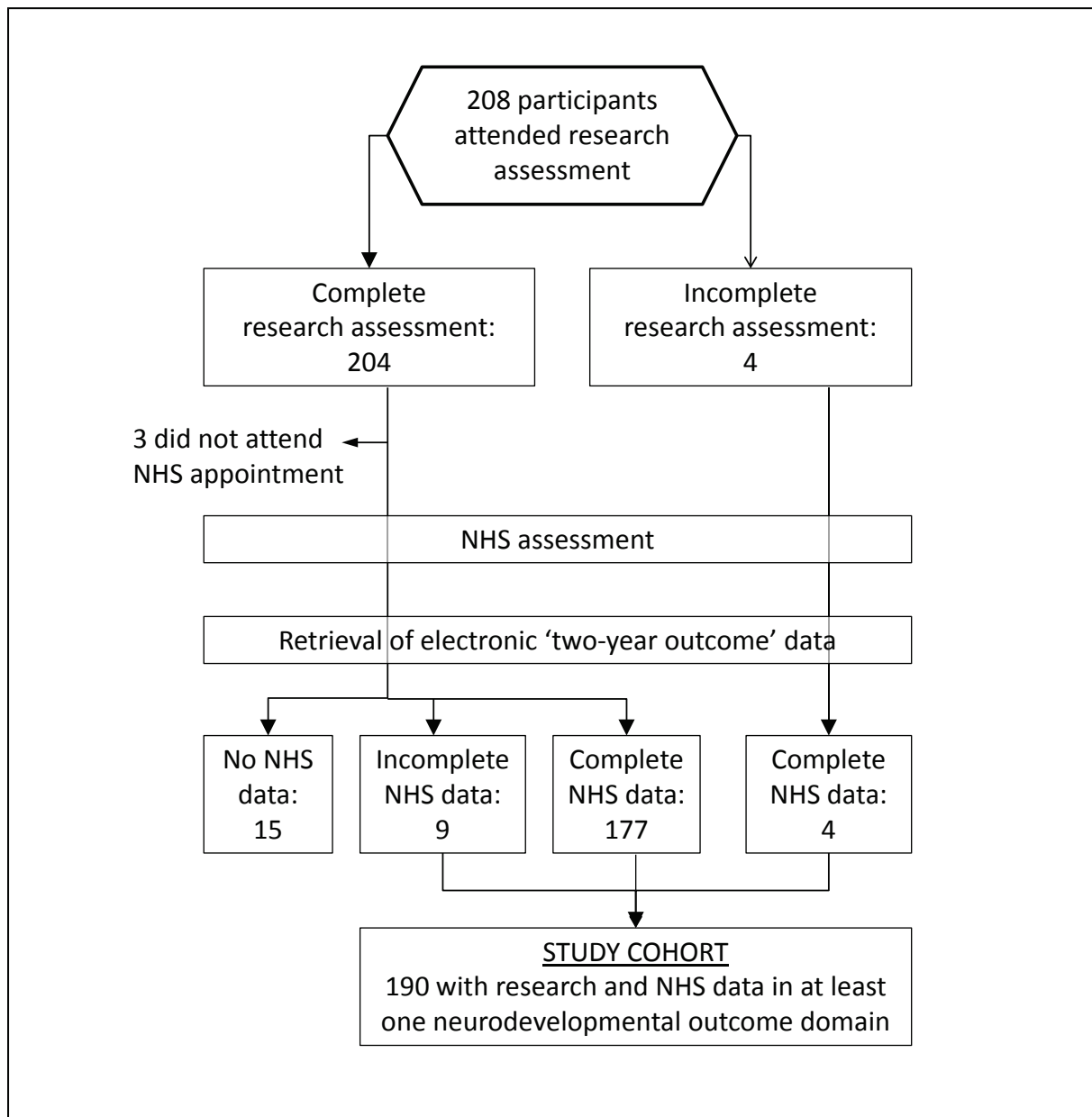
5.1.1 Derivation of study population

208 children were recruited to this study. Figure 5.1 shows the flow of children through recruitment to the completion of research and NHS assessments. I do not have data on the number of children considered ineligible (already received Bayley-III assessment and non-English speaking families) and not invited to participate by the local collaborators, nor the total number of parents who were approached and those who declined to participate. The representativeness of the study population is considered in the next section (5.1.2).

Two hundred and four children completed all the subtests of the Bayley-III assessment. One child with ataxic cerebral palsy could not be assessed for the cognitive and language scales; two children did not cooperate for the receptive communication assessment and one for the gross motor assessment. Of the children who completed the research assessments, three did not attend their routine NHS follow-up visit. Data from the NHS assessment were not entered on the electronic 'two-year outcome' form by the examining clinician in 15 cases and the overall category of impairment could not be assigned due to missing data on the electronic database for nine children. The 190 children on whom both research and NHS data were available in at least one outcome domain form the study cohort. A complete set of data in all outcome domains was available for 177 children. Although the original plan was to stratify recruitment based on the gestational ages of the eligible children, as it became clear during the recruitment phase that the final study cohort would be smaller than the initial projection, I recruited and assessed all children whose parents agreed to

participate. I do not think that the final sample size being smaller than the original target invalidated my findings and I discussed the reasons and implications of this in section 5.6.4.

Figure 5.1 Flowchart of children through research and NHS assessments to form the study population



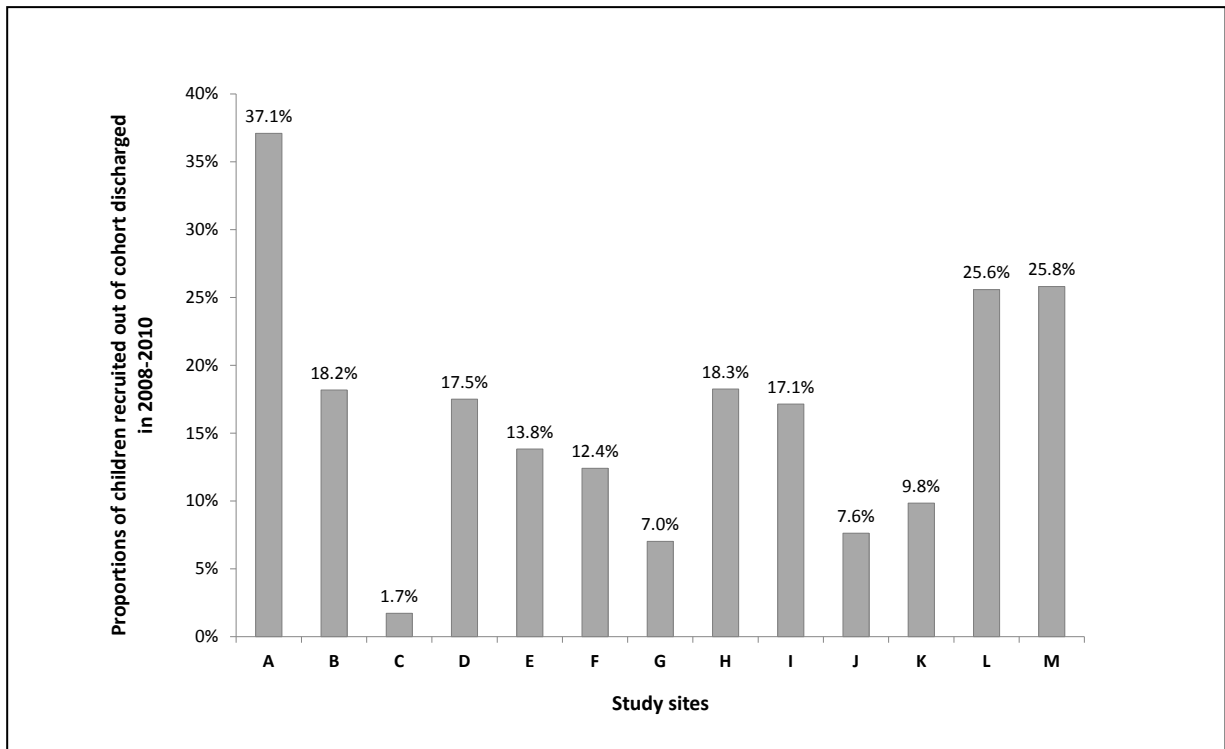
5.1.2 Characteristics and representativeness of study population

There was likely to be selection bias as the participation rates and completeness of electronic data entry were inconsistent between study sites (figure 5.2). I compared the characteristics of the study population with the 'baseline population' (all infants born between 1 January 2008 and 31 December 2010, at gestational ages below 30 weeks, and discharged from the participating hospitals); the results are shown in table 5.1. The most significant differences were that study participants received shorter duration of mechanical ventilation and were less likely to be receiving oxygen therapy at 36 weeks' postmenstrual age than the baseline population ($p < 0.001$). The study population was comparable to the baseline population in terms of gestational age, birth weight, sex, proportions of singleton, mode of delivery and maternal age. The study population consisted of larger proportions of children of white ethnicity and born to mothers living in the least deprived IMD quintile compared with the 'baseline population'. Nevertheless, a wide range of ethnic groups was represented, reflecting the diversity of the population living in London. Consequentially, although children from non-English speaking families were not included in the study, 92 (48.4%) children were raised in a bilingual or multilingual environment and in 38 (20.1%), English made up less than 50% of their total language exposure (as estimated by the parents).

Table 5.1 Demographic and neonatal characteristics of study population versus non-participants born <30 weeks gestation in 2008-2010 and discharged from the participating study sites

Characteristics	Study population (n=190)	'Baseline population' (n=1,037)	p- value
Gestation (completed weeks)	27 (26 - 29),	27 (26 - 29),	0.25
Median (IQR) range	23 - 29	22 - 29	
Birth weight (g)	965 (790 - 1140),	1000 (812 - 1200),	0.08
Median (IQR), range	490 - 1720	455 - 1990	
Sex			
Girls, n (%)	99 (52.1)	444 (42.8)	0.19
Boys, n (%)	91 (47.9)	503 (48.5)	
Missing, n (%)	0 (0.0)	90 (8.7)	
Ethnicity			
White, n (%)	88 (46.3)	364 (35.1)	0.03
Black, n (%)	50 (26.3)	287 (27.7)	
Asian, n (%)	41 (21.6)	239 (23.1)	
Mixed, n (%)	0 (0.0)	33 (3.2)	
Other, n (%)	11 (5.8)	52 (5.0)	
Missing, n (%)	0 (0.0)	62 (6.0)	
Pregnancy			
Singleton, n (%)	147 (77.4)	690 (66.5)	0.26
Multiples, n (%)	43 (22.6)	250 (24.1)	
Missing, n (%)	0 (0.0)	97 (9.4)	
Mode of delivery			
Vaginal, n (%)	74 (39.0)	475 (45.8)	0.22
Caesarean, n (%)	103 (54.2)	540 (52.1)	
Missing, n (%)	13 (6.8)	22 (2.1)	
Maternal age (years)			
Mean (SD)	31.9 (6.7)	31.0 (6.4)	0.08
IMD quintile at birth			
One (least deprived), n (%)	19 (10.0)	43 (4.2)	0.01
Two, n (%)	20 (10.5)	81 (7.8)	
Three, n (%)	26 (13.7)	144 (13.9)	
Four, n (%)	52 (27.4)	268 (25.8)	
Five (most deprived), n (%)	73 (38.4)	477 (46.0)	
Missing, n (%)	0 (0.0)	24 (2.3)	
Length of mechanical ventilation (days)			
Median (IQR), range	0 (0 - 3), 0 - 54	4 (0 - 18), 0 - 444	<0.001
Oxygen therapy at 36 weeks' corrected age			
Yes, n (%)	54 (28.4)	466 (44.9)	<0.001
No, n (%)	136 (71.6)	574 (55.1)	

Figure 5.2 Children recruited at each study site as a proportion of the total number of eligible infants born in 2008-2010 and discharged home from each site



5.2 NEURODEVELOPMENTAL OUTCOMES FROM RESEARCH ASSESSMENTS

The mean (SD) corrected age of the children at assessment was 24.8 (2.2) months. The research assessment took place at a median (inter-quartile range (IQR)) interval of 8 (0 - 27) days after the children received their NHS assessment, with a range between 89 days before and 82 days after the NHS assessment. This wide range in interval was due to difficulty in scheduling appointments and missed appointments by the children.

Based on information given by the parents, 30 (15.8%) children had a visual defect including reduced visual acuity and/or squints, although only 11 (5.8%) required glasses. 16 (8.4%) children had hearing impairment of whom 3 (1.6%) wore hearing aids.

The children performed significantly worse than the normative population in which Bayley-III scores were standardised in all domains other than fine motor (table 5.2).

Table 5.2 Mean Bayley-III scores (scaled scores and composite scores) of study population

		Mean (SD) Bayley-III scores	<i>p</i> -value*
Cognitive domain	Cognitive composite score (n = 189)	92.65 (12.8)	<0.001
	Language domain		
Language domain	Receptive communication scaled score (n=187)	8.0 (2.4)	<0.001
	Expressive communication scaled score (n=189)	7.5 (2.5)	<0.001
	Language composite score (n=187)	87.0 (13.6)	<0.001
Motor domain	Fine motor scaled score (n=190)	10.2 (2.5)	0.23
	Gross motor scaled score (n=189)	8.6 (2.3)	<0.001
	Motor composite score (n=189)	96.7 (12.7)	<0.001

**p*-value from Student's t-test comparing the study population mean scores to the Bayley-III standardised scaled score mean of 10 and standardised composite score mean of 100.

5.2.1 Categorisation of neurodevelopmental status using Bayley-III scores

All children were classified onto one of 3 SD score groups according to their Bayley-III scores (methods outlined in section 4.8.1). Based on the worst score achieved in the cognitive, language and motor domains, 114 (61.3%) children were classified as having scores higher than -1 SD from the standardised mean, 42 (22.6%) had scores between -1 and -2 SD and 30 (16.1%) had scores lower than -2 SD from the standardised mean.

Cognitive domain

In the cognitive domain, 156 (82.1%) children obtained Bayley-III scores higher than -1 SD, 26 (13.7%) had scores between -1 and -2 SD and seven (3.7%) had scores lower than -2 SD from the standardised mean.

Language domains

19 (10.0%) children had specific expressive communication impairment (scaled score <7), with no impairment in receptive communication. Five of these children would have been classified as having no impairment based on the Bayley-III language composite score alone (language composite score higher than -1 SD, i.e. ≥ 85), due to compensation from the receptive communication subtest. Based on the worst SD score category from the receptive and expressive communication subtests and the language composite score, 120 (63.2%) children achieved scores higher than -1 SD, 42 (22.1%) had scores between -1 and -2 SD and 25 (13.2%) had scores lower than -2 SD from the standardised mean.

Motor domains

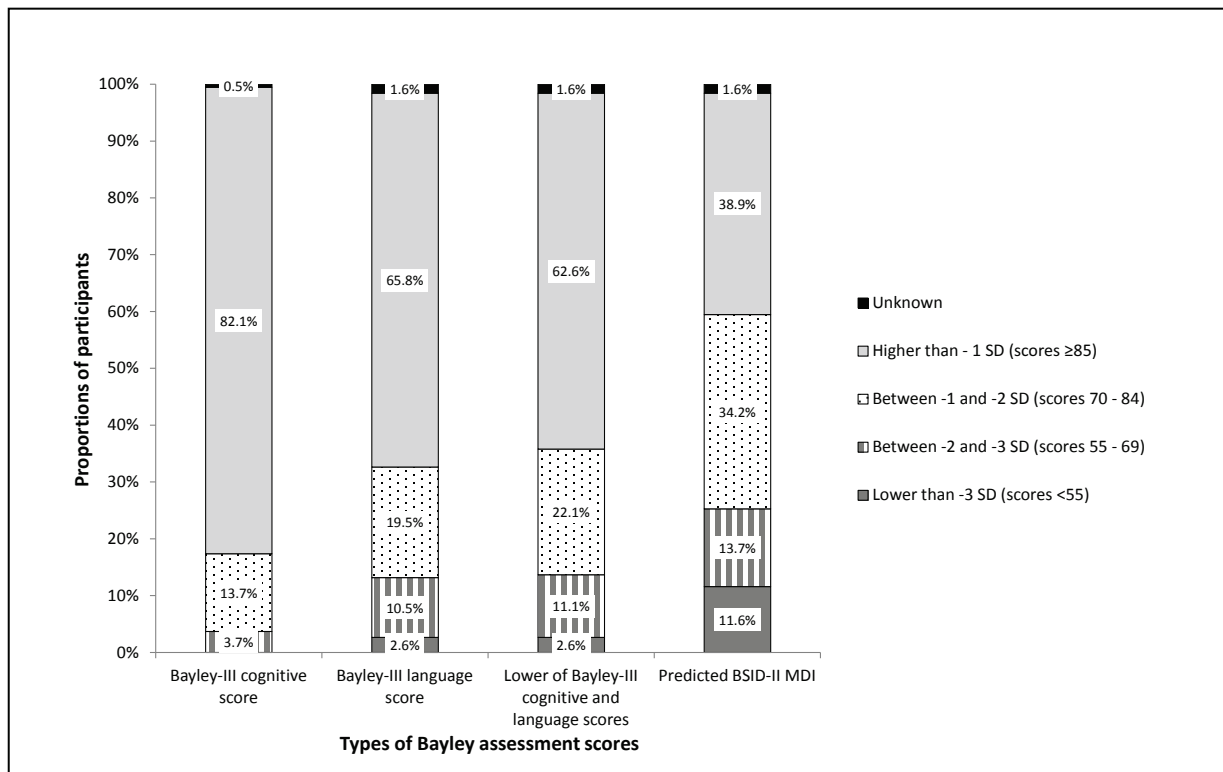
Motor function was generally intact among the children, with only 11 (5.8%) receiving scores between -1 and -2 SD and 11 (5.8%) having scores lower than -2 SD from the standardised mean.

Comparison of Bayley-III and the predicted BSID-II MDI scores

The predicted BSID-II MDI was estimated for each participant using the algorithm published by Moore et al ((Moore 2012c); section 4.7.2). The mean (SD) predicted BSID-II MDI for the study population was 77.9 (20.5) and was significantly lower than the mean Bayley-III cognitive and language composite scores ($p < 0.001$ for both). Figure 5.3 shows the proportions of children with scores higher than -1 SD (≥ 85), between -1 and -2 SD (70 - 84), between -2 and -3 SD (55 - 69) and lower than -3 SD (< 55) from the standardised mean for the Bayley-III cognitive and language composite scores individually, when the lower of the two Bayley-III scores was used and for the predicted BSID-II MDI score. From the figure, I considered whether a cut-off score of < 85 (-1 SD from the standardised mean) on the Bayley-III would be equivalent to a predicted BSID-II MDI score of < 70 (-2 SD from the standardised mean) for the diagnosis of impairment. Hence, I performed a post-hoc analysis using McNemar's test to compare the proportions diagnosed with impairment using these

cut-off scores. The proportions of children classified with impairment using a cut-off of <85 on the Bayley-III cognitive score (17.4%), language score (32.6%) and the lower of the cognitive and language scores (35.8%) were statistically dissimilar to the proportions with predicted BSID-II MDI <70 (25.3%) (p -values <0.001). However, the proportions of children with predicted BSID-II MDI <55 (11.6%) were similar to the proportions who scored <70 (classified with severe impairment) on the Bayley-III language composite score (13.1%; p -value 0.26) and the lower of the cognitive and language scores (13.7%; p -value 0.16).

Figure 5.3 Neurodevelopmental status by Bayley-III scores and the predicted BSID-II MDI



5.2.2 Classification of impairment using NPEU/Oxford criteria

Children were also classified into levels of functional impairment in the communication and motor domains as per the NPEU/Oxford criteria (methods described in section 4.8.2).

Communication

In the communication domain, the assignment of outcome was heavily influenced by the presence of specific expressive communication impairment, as most children with expressive communication difficulties had no or mild receptive communication problems. 107 (56.3%) children were classified to have no communication impairment, 55 (28.9%) had mild-moderate communication impairment and 27 (14.2%) had severe communication impairment.

Motor function

13 (6.8%) children had isolated gross motor impairment with no fine motor difficulties; only one (0.5%) child had specific fine motor impairment. The combined motor outcome was normal in 172 (90.5%) children. Nine (4.7%) were classed to have mild-moderate motor impairment and nine (4.7%) had severe motor impairment.

5.2.3 Concordance between Bayley-III and NPEU/Oxford criteria

I constructed cross-tabulations to evaluate the concordance in the classification of neurodevelopmental status by Bayley-III scores and by the NPEU/Oxford criteria, for the communication and motor domains (table 5.3). The definition of impairment based on Bayley-III scores were as follow: 'none' if scores were higher than -1 SD below the standardised mean, 'mild-moderate' if scores were between -1 and -2 below the standardised mean and 'severe' if scores were lower than -2 SD below the standardised mean. Of the 187 children tested for their communication skills, 144 (77.0%) were classified in the same category; of the other children, none differed by more than one category. The weighted kappa coefficient (κ) was 0.59 (95% CI 0.49 - 0.69), indicating moderate agreement between the two criteria for the communication outcome. For the motor domain, 180 out of 189 (95.2%) were classified in the same category. Classification differed by one category for eight children. One child who was assessed as having severe motor impairment

from the Bayley-III was classified as having ‘no impairment’ on the NPEU/Oxford criteria. The weighted κ for concordance between the two methods in the motor domain was 0.76 (0.62 - 0.93), representing substantial agreement.

Table 5.3 Cross-tabulation of classification of impairment by Bayley-III scores and the NPEU/Oxford criteria with agreement between the two methods measured by kappa coefficient (κ)*

NPEU/ Oxford criteria	Impairment in communication (n=190)				Impairment in motor ability (n=190)			
	Bayley-III SD score groups				Bayley-III SD score groups			
	> -1 SD	-1 to -2 SD	< -2 SD	Unknown	> -1 SD	-1 to -2 SD	< -2 SD	Unknown
None	101	4	0	2	165	5	1	1
Mild- moderate	19	27	9	0	2	6	1	0
Severe	0	11	16	0	0	0	9	0
Unknown	0	0	0	1	0	0	0	0
	Unweighted κ (95% CI) = 0.59 (0.49 - 0.68)				Unweighted κ (95% CI) = 0.76 (0.58 - 0.87)			
	Weighted κ (95% CI) = 0.59 (0.49 - 0.69)				Weighted κ (95% CI) = 0.76 (0.62 - 0.93)			

*Shaded cells indicate concordance in category of impairment between the 2 classification methods.

5.2.4 Reliability of research assessments

During the course of the study, I attended two sessions (one midway and one at the end of the assessment phase of the study) with Ms Hutchon (Bayley-III trainer), to revalidate my assessment techniques and to evaluate the inter-rater reliability between our scores. The percent agreement across all assessed items was 97.2% (69 out of 71 items in agreement) in the first session and 98.6% (69 out of 70 items in agreement) in the second session.

5.3 NEURODEVELOPMENTAL OUTCOMES FROM ROUTINE NHS DATA

The two-year outcome data recorded during the NHS appointments were extracted from the NNRD on 1st November 2012. A further check on the NNRD on 15th July 2013 retrieved data that was

missing in the first extraction for 12 children. Children attended their NHS follow-up assessment at a mean (SD) corrected age of 24.4 (2.3) months. Data were entered on the electronic 'two-year outcome' form by clinical consultants in 111 (58.4%) cases (36 (19.0%) by consultant neonatologists, 42 (22.1%) by hospital paediatrics consultants, 33 (17.4%) by community paediatrics consultants), trainee doctors in 15 (7.9%) cases, staff grade doctors in 58 (30.5%) cases and administrative staff in 6 (3.2%) cases. Sixty-seven (35.3%) children received standardised neurodevelopmental assessment or screening tests during their NHS appointment (19 (10.0%) using the Schedule of Growing Scales, 44 (23.2%) using the Griffiths Mental Development Scales and 4 (2.1%) using the Alberta Infant Motor Scale). Table 5.4 shows the responses to each question on the electronic form and the classification of impairment for the developmental domains. The classification of overall neurodevelopmental outcome was possible in 181 children, of whom 124 (68.5%) had no impairments, 38 (21.0%) had mild-moderate impairments and 19 (10.5%) had severe impairments.

Table 5.4 Responses to questions on the electronic ‘two-year outcome’ form and classification of impairment based on NHS data

Domain	Question	Response, n (%)			Classification of impairment, n (%)
		No	Yes	Missing	
Cognitive	D1	154 (81.1)	35 (18.4)	1 (0.5)	None: 141 (74.2) Mild-moderate: 42 (22.1) Severe: 6 (3.2) Unknown: 1 (0.5)
	D2	171 (90.0)	19 (10.0)	0 (0.0)	
	D3	183 (96.3)	6 (3.2)	1 (0.5)	
Receptive communication	RC1	174 (91.6)	13 (6.8)	3 (1.6)	None: 174 (91.6) Mild-moderate: 8 (4.2) Severe: 5 (2.6) Unknown: 3 (1.6)
	RC2	183 (96.3)	5 (2.6)	2 (1.1)	
Expressive communication	EC1	149 (78.4)	40 (21.1)	1 (0.5)	None: 143 (75.3) Mild-moderate: 32 (16.8) Severe: 13 (6.8) Unknown: 2 (1.1)
	EC2	153 (80.5)	36 (18.9)	1 (0.5)	
	EC3	176 (92.6)	13 (6.8)	1 (0.5)	
Combined communication*					None: 141 (74.2) Mild-moderate: 30 (15.8) Severe: 14 (7.4) Unknown: 5 (2.6)
Fine motor	FM1	188 (98.9)	2 (1.1)	0 (0.0)	None: 186 (97.9) Mild-moderate: 2 (1.1) Severe: 2 (1.1) Unknown: 0 (0.0)
	FM2	189 (99.5)	1 (0.5)	0 (0.0)	
	FM3	188 (98.9)	2 (1.1)	0 (0.0)	
Gross motor	GM1	178 (93.7)	12 (6.3)	0 (0.0)	None: 174 (91.6) Mild-moderate: 5 (2.6) Severe: 8 (4.2) Unknown: 3 (1.6)
	GM2	177 (93.2)	9 (4.7)	4 (2.1)	
	GM3	182 (95.8)	8 (4.2)	0 (0.0)	
	GM4	186 (97.9)	4 (2.1)	0 (0.0)	
	GM5	188 (98.9)	2 (1.1)	0 (0.0)	
Combined motor*					None: 173 (91.1) Mild-moderate: 6 (3.2) Severe: 8 (4.2) Unknown: 3 (1.6)
Overall*					None: 124 (65.3) Mild-moderate: 38 (20.0) Severe: 19 (10.0) Unknown: 9 (4.7)

*Combined communication impairment was judged as the worst category of outcome from receptive communication and expressive communication; combined motor impairment was judged as the worst category of outcome from fine motor and gross motor. Overall impairment was based on the worst category of outcome from the cognitive, combined communication and combined motor domains.

5.4 AGREEMENT IN THE CLASSIFICATION OF OUTCOMES BETWEEN RESEARCH AND ROUTINE NHS ASSESSMENTS

5.4.1 Validity of NHS assessment against research assessment

In figures 5.4 to 5.7, I display the proportions of children classified into each category of impairment by the research assessment (using Bayley-III scores and NPEU/Oxford criteria) and by the NHS assessment. Once again, the classification of impairment by Bayley-III scores were 'none' for scores higher than -1 SD, 'mild-moderate' for scores between -1 and -2 SD and 'severe' for scores lower than -2 SD below the standardised mean.

Figure 5.4 Classification of the severity of cognitive impairment of the children by research and NHS assessments

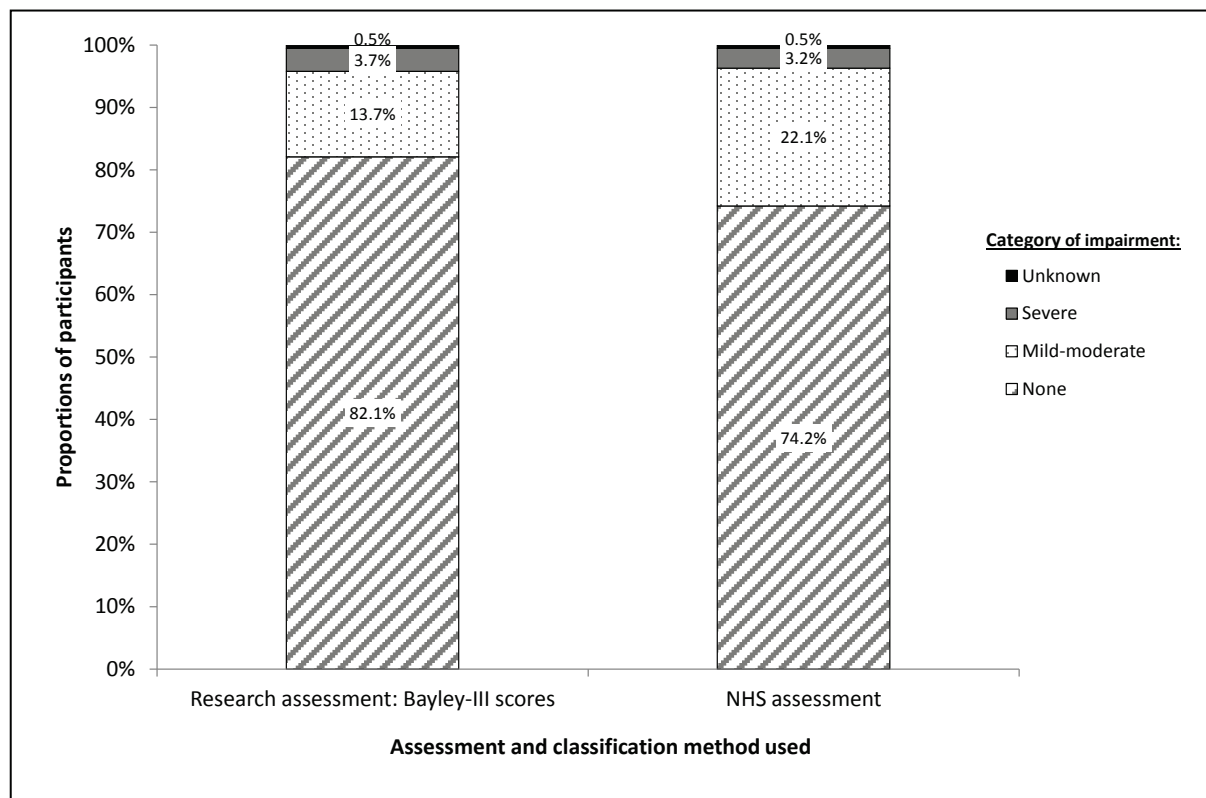


Figure 5.5 Classification of the severity of receptive communication, expressive communication and overall communication impairment of the participants based on the Bayley-III and NPEU/Oxford classification by research assessment, and by NHS assessments

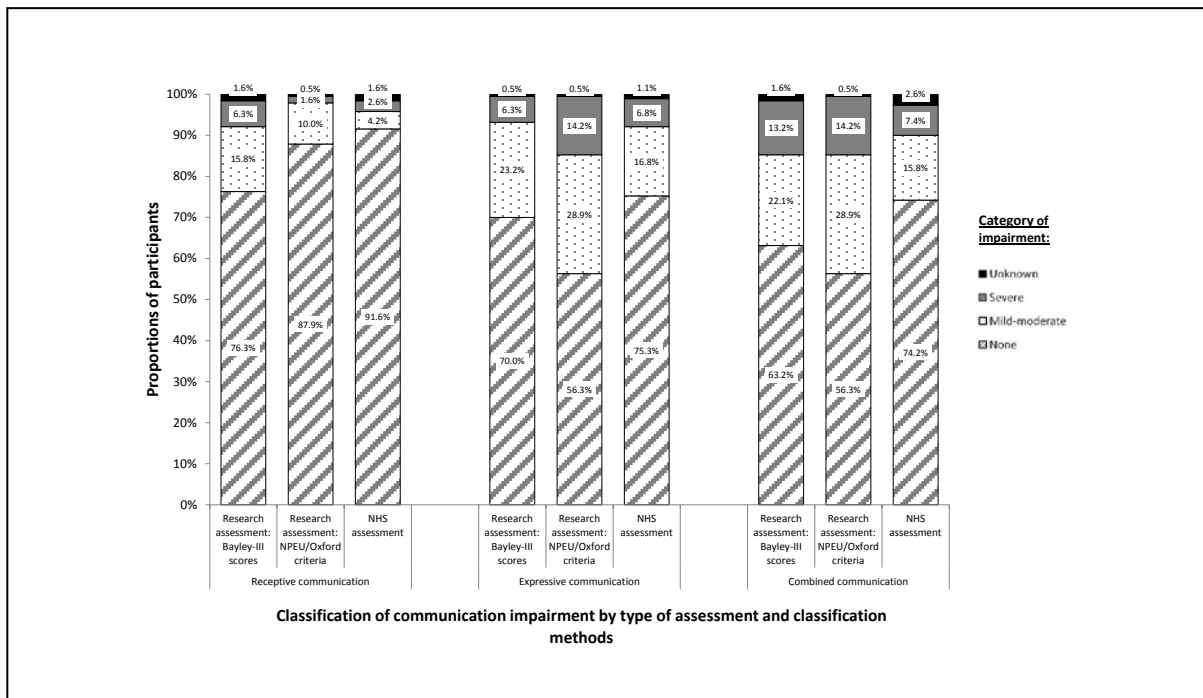


Figure 5.6 Classification of the severity of fine motor, gross motor and overall motor impairment of the participants based on the Bayley-III and NPEU/Oxford classification by research assessment, and by NHS assessments

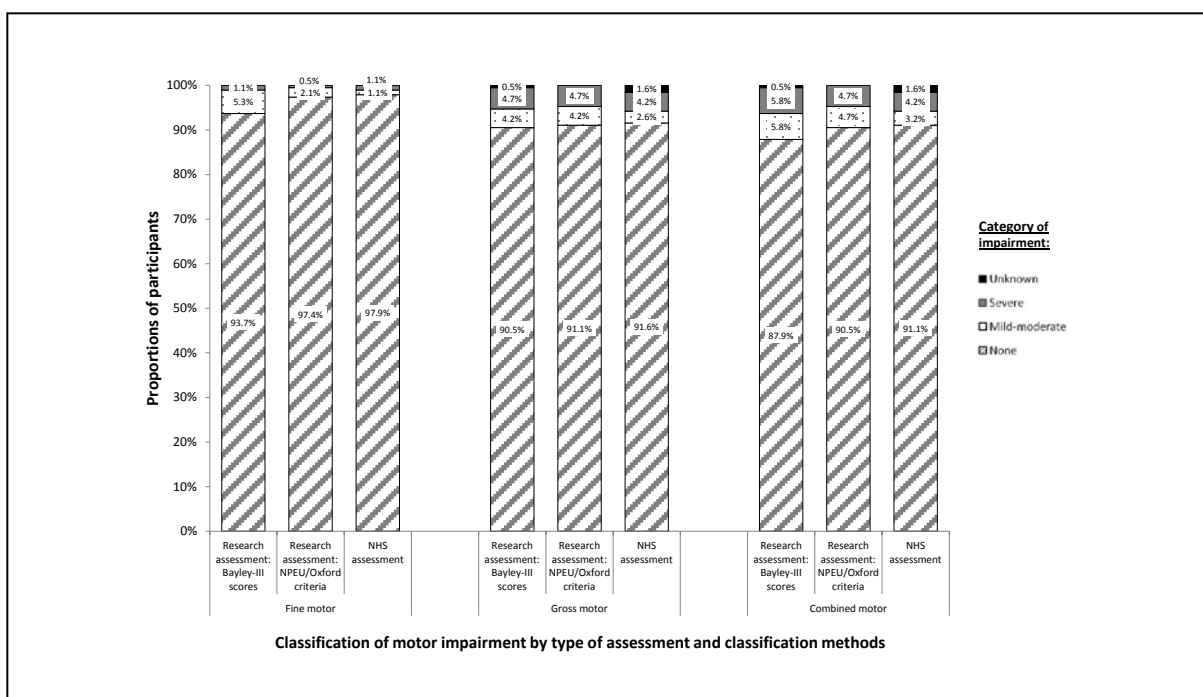
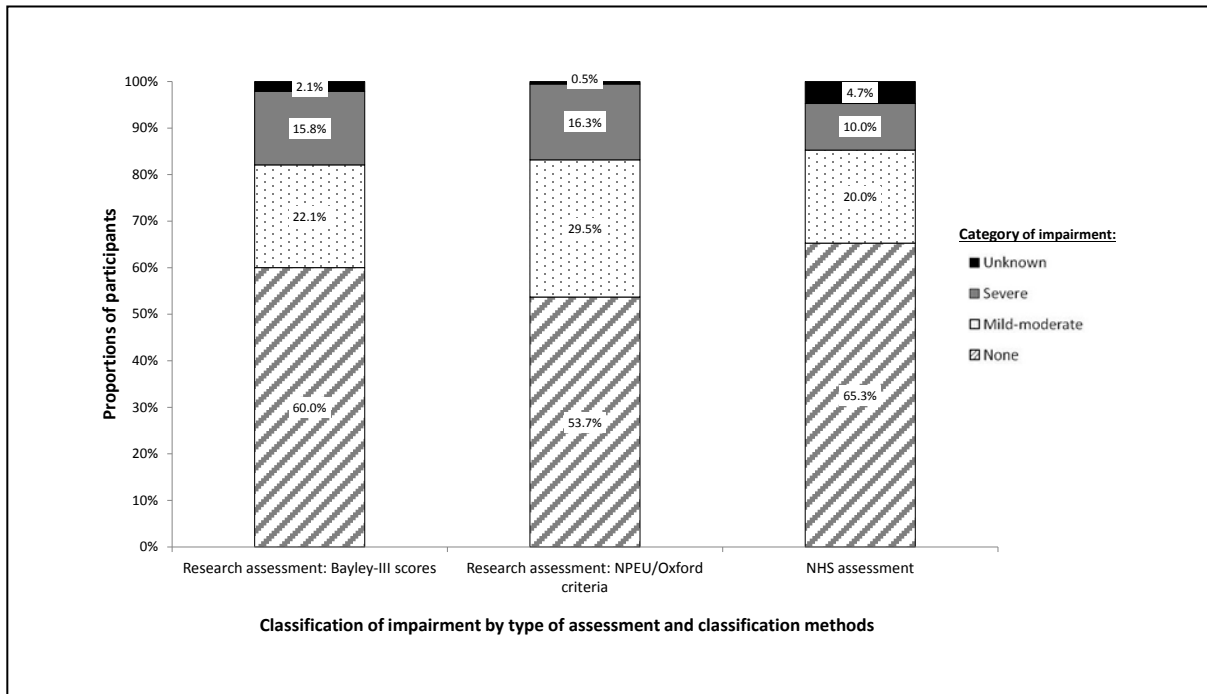


Figure 5.7 Classification of the neurodevelopmental outcome of participants by the severity of the worst impairment in the cognitive, communication and motor domains through research and NHS assessments



Cross-tabulations to compare the agreement between the research and the NHS assessments for the classification of ‘any impairment’ and ‘severe impairment’ are displayed in tables 5.5 and 5.6. Taking the research assessment as the ‘gold-standard’ with 100% sensitivity and specificity in the identification and classification of impairment, the estimated sensitivities and specificities for NHS assessment in each developmental domain are presented. The confidence intervals for sensitivities and specificities were calculated using robust standard errors to account for clustering of data by study sites. Sensitivity analyses revealed that potential correlated outcomes from siblings did not affect the results. Therefore, the results presented included data from all participating children. I present the findings from the sensitivity analyses on all singleton births and only one randomly selected child from each multiple birth set in Appendix 7.

The validity of the NHS assessments in identifying children with no impairments was high, with estimated specificities ranging between 83.9% - 100.0% for 'any impairment' and between 96.6% - 100.0% for 'severe impairment'. However, the validity of the NHS and the research assessment in identifying and categorising children with impairments was variable. The sensitivities for identifying gross motor impairment were high, particularly when the impairment was severe. In the cognitive domain, the sensitivity for the identification of any impairment was 69.7% (95% CI 55.1% - 84.3%) but dropped to only 28.6% (5.0% - 52.2%) for the identification of severe impairment. Of the 7 children diagnosed with severe cognitive impairment through the research assessment, 2 were also classified in the 'severe' category in the NHS dataset; 4 were classified as 'mild-moderate' and one was classified as 'no' impairment. Hence, the disagreement in the classification occurred mainly between the 'mild-moderate' and 'severe' categories. Agreement between NHS and research assessments was worst in the communication domain, especially in receptive communication where the sensitivity in identifying the presence of any impairment was only 23.1% (6.7% - 39.5%). In the combined communication domain, the 21 'false-negatives' for severe impairment, based on Bayley-III classification, were evenly classified between the impairment categories in the NHS data (9 'severe', 6 'mild-moderate', 6 'no' impairments). I note that the sensitivities were estimated with low precision (wide confidence intervals), particularly in the motor domains and with severe impairments where the prevalence of impairment was low.

Although the sensitivities appeared considerably higher in the receptive communication and fine motor domains when impairment was assigned using the NPEU/Oxford criteria compared with Bayley-III scores for the research assessment, this was driven by the small numbers in the 'false-negative' cells and the estimated sensitivities were associated with wide and overlapping confidence intervals, suggesting that the differences in sensitivities may not be statistically significant.

The sensitivities and specificities of NHS assessment in identifying cognitive deficit were 69.7% (55.1 - 84.3) and 83.9% (75.6 - 92.1) for the presence of any impairment and 28.6% (5.0 - 52.2) and 97.8% (95.1 - 100.0) for severe impairments. I repeated the analyses using the predicted BSID-II MDI scores as the 'gold-standard'. Using the cut-off of MDI <85 (-1 SD) to define mild-moderate impairment and <70 (-2 SD) to define severe impairment, there was a reduction in the sensitivities and an increment in the corresponding specificities (sensitivity 39.3% (30.2 - 49.0) and specificity 94.6 (86.7 - 98.5) for any cognitive impairment; sensitivity 12.5% (4.7 - 25.2) and specificity 100% (97.4 - 100) for severe cognitive impairment). However, if thresholds MDI <70 (-2 SD) for mild-moderate and <55 (-3 SD) for severe impairments were used, the results were similar to the reported findings using the Bayley-III (sensitivity 64.6% (49.5% - 77.8%), specificity 87.7% (81.0% - 92.7%) for any impairment; sensitivity 18.2% (5.1% - 40.3%), specificity 98.8% (95.7% - 99.9%) for severe impairment).

Table 5.5 Results of cross-tabulations comparing the NHS and research categorisation of impairment and the sensitivities and specificities of the NHS assessment in identifying children with *any* impairment against the ‘gold-standard’ research assessment

Domain of development*	Method of classification of impairment for research assessment	Identification of impairment by NHS assessment against the ‘gold-standard’ research assessment					
		True-positives, n (%)	False-negatives, n (%)	False-positives, n (%)	True-negatives, n (%)	Sensitivity, % (95% CI)	Specificity, % (95% CI)
Cognitive	Bayley-III scores (n=188)	23 (12.2)	10 (5.3)	25 (13.3)	130 (69.1)	69.7 (55.1 - 84.3)	83.9 (75.6 - 92.1)
Receptive communication	Bayley-III scores (n=184)	9 (4.9)	30 (16.3)	4 (2.2)	141 (76.6)	23.1 (6.7 - 39.5)	97.2 (94.6 - 99.9)
	NPEU/Oxford (n=186)	8 (4.3)	13 (7.0)	5 (2.7)	160 (86.0)	38.1 (10.7 - 65.5)	97.0 (94.8 - 99.2)
Expressive communication	Bayley-III scores (n=187)	32 (17.1)	22 (11.8)	13 (7.0)	120 (64.2)	59.3 (46.5 - 72.0)	90.2 (82.2 - 98.3)
	NPEU/Oxford (n=187)	39 (20.9)	41 (21.9)	6 (3.2)	101 (54.0)	48.8 (33.9 - 63.6)	94.4 (88.9 - 99.9)
Combined communication	Bayley-III scores (n=182)	33 (18.1)	29 (15.9)	11 (6.0)	109 (59.9)	53.2 (42.0 - 64.5)	90.8 (83.5 - 98.2)
	NPEU/Oxford (n=184)	38 (20.7)	39 (21.2)	6 (3.3)	101 (54.9)	49.4 (34.7 - 64.0)	94.4 (88.9 - 99.9)
Fine motor	Bayley-III scores (n=190)	3 (1.6)	9 (4.7)	1 (0.5)	177 (93.2)	25.0 (0.0 - 59.7)	99.4 (98.3 - 100.0)
	NPEU/Oxford (n=190)	4 (2.1)	1 (0.5)	0 (0.0)	185 (97.4)	80.0 (28.4 - 99.5)	100.0 (98.0 - 100.0)
Gross motor	Bayley-III scores (n=186)	12 (6.5)	4 (2.2)	1 (0.5)	169 (90.9)	75.0 (49.9 - 100.0)	99.4 (98.1 - 100.0)
	NPEU/Oxford (n=187)	11 (5.9)	5 (2.7)	2 (1.1)	169 (90.4)	68.8 (45.5 - 92.0)	98.8 (97.1 - 100.0)
Combined motor	Bayley-III scores (n=186)	13 (7.0)	8 (4.3)	1 (0.5)	164 (88.2)	61.9 (32.9 - 90.9)	99.4 (98.1 - 100.0)
	NPEU/Oxford (n=187)	12 (6.4)	5 (2.7)	2 (1.1)	168 (89.8)	70.6 (48.8 - 92.4)	98.8 (97.0 - 100.0)
Overall	Bayley-III scores (n=177)	40 (22.6)	25 (14.1)	16 (9.0)	96 (54.2)	61.5 (52.5 - 70.6)	85.7 (77.4 - 94.0)

*Combined communication impairment was judged as the worst category of outcome from receptive communication and expressive communication; combined motor impairment was judged as the worst category of outcome from fine motor and gross motor. Overall impairment was based on the worst category of outcome from the cognitive, communication and motor domains.

Table 5.6 Results of cross-tabulations comparing the NHS and research categorisation of impairment and the sensitivities and specificities of the NHS assessment in identifying children with *severe* impairment against the ‘gold-standard’ research assessment

Domain of development*	Method of classification of impairment for research assessment	Identification of severe impairment by NHS assessment against the ‘gold-standard’ research assessment					
		True-positives, n (%)	False-negatives, n (%)	False-positives, n (%)	True-negatives, n (%)	Sensitivity, % (95% CI)	Specificity, % (95% CI)
Cognitive	Bayley-III scores (n=188)	2 (1.1)	5 (2.7)	4 (2.1)	177 (94.1)	28.6 (5.0 - 52.2)	97.8 (95.1 - 100.0)
Receptive communication	Bayley-III scores (n=184)	3 (1.6)	8 (4.3)	2 (1.1)	171 (92.9)	27.3 (0.0 - 62.9)	98.8 (97.3 - 100.0)
	NPEU/Oxford (n=186)	2 (1.1)	1 (0.5)	3 (1.6)	180 (96.8)	66.7 (4.9 - 100.0)	98.4 (95.3 - 99.7)
Expressive communication	Bayley-III scores (n=187)	7 (3.7)	5 (2.7)	6 (3.2)	169 (90.4)	58.3 (36.6 - 80.0)	96.6 (92.7 - 98.7)
	NPEU/Oxford (n=187)	9 (4.8)	18 (9.6)	4 (2.1)	156 (83.4)	33.3 (12.0 - 54.7)	97.5 (93.7 - 99.3)
Combined communication	Bayley-III scores (n=182)	9 (4.9)	12 (6.6)	5 (2.7)	156 (85.7)	42.9 (14.2 - 71.5)	96.9 (92.9 - 99.0)
	NPEU/Oxford (n=184)	9 (4.9)	15 (8.2)	5 (2.7)	155 (84.2)	37.5 (15.4 - 59.6)	96.9 (92.9 - 99.0)
Fine motor	Bayley-III scores (n=190)	1 (0.5)	1 (0.5)	1 (0.5)	187 (98.4)	50.0 (0.0 - 100.0)	99.5 (97.1 - 100.0)
	NPEU/Oxford (n=190)	1 (0.5)	0 (0.0)	1 (0.5)	188 (98.9)	100.0 (2.5 - 100.0)	99.5 (97.1 - 100.0)
Gross motor	Bayley-III scores (n=186)	8 (4.3)	1 (0.5)	0 (0.0)	177 (95.2)	88.9 (51.8 - 99.7)	100.0 (97.9 - 100.0)
	NPEU/Oxford (n=187)	8 (4.3)	1 (0.5)	0 (0.0)	178 (95.2)	88.9 (51.8 - 99.7)	100.0 (97.9 - 100.0)
Combined motor	Bayley-III scores (n=186)	8 (4.3)	2 (1.1)	0 (0.0)	176 (94.6)	80.0 (44.4 - 97.5)	100.0 (97.9 - 100.0)
	NPEU/Oxford (n=187)	8 (4.3)	1 (0.5)	0 (0.0)	178 (95.2)	88.9 (51.8 - 99.7)	100.0 (97.9 - 100.0)
Overall	Bayley-III scores (n=177)	13 (7.3)	12 (6.8)	5 (2.8)	147 (83.1)	52.0 (23.8 - 80.2)	96.7 (92.5 - 99.9)

*Combined communication impairment was judged as the worst category of outcome from receptive communication and expressive communication; combined motor impairment was judged as the worst category of outcome from fine motor and gross motor. Overall impairment was based on the worst category of outcome from the cognitive, communication and motor domains.

5.4.2 Concordance in the assignment of the category of outcome between research and NHS assessments

The concordance of the research and the NHS assessments as measured by κ was consistent with the findings from the estimated sensitivities and specificities. I present both the unweighted and weighted κ in table 5.7, although the addition of weighting for partial agreement between mild-moderate and severe impairment did not lead to significant increase in the weighted κ statistic from the unweighted form. The agreement between NHS and research assessment was substantial in the motor domain with weighted $\kappa > 0.6$. In the cognitive and communication domains, agreement was moderate at best.

Table 5.7 Concordance in the assignment of the category of impairment between research and NHS assessments as measured by unweighted and weighted kappa coefficients (κ)

Domain of development*	Method of classification of impairment for research assessment	Unweighted κ statistic (95% CI)	Weighted κ statistic (95% CI)
Cognitive	Bayley-III scores (n=188)	0.37 (0.21 - 0.52)	0.41 (0.22 - 0.51)
Receptive communication	Bayley-III scores (n=184)	0.23 (0.10 - 0.45)	0.25 (0.10 - 0.39)
	NPEU/Oxford (n=186)	0.34 (0.11 - 0.52)	0.38 (0.16 - 0.55)
Expressive communication	Bayley-III scores (n=187)	0.44 (0.31 - 0.61)	0.48 (0.37 - 0.60)
	NPEU/Oxford (n=187)	0.35 (0.23 - 0.46)	0.40 (0.30 - 0.53)
Combined communication	Bayley-III scores (n=182)	0.40 (0.27 - 0.50)	0.43 (0.25 - 0.54)
	NPEU/Oxford (n=184)	0.36 (0.28 - 0.50)	0.41 (0.32 - 0.53)
Fine motor	Bayley-III scores (n=190)	0.30 (0.04 - 0.55)	0.33 (0.12 - 0.58)
	NPEU/Oxford (n=190)	0.77 (0.50 - 1.00)	0.83 (0.55 - 1.00)
Gross motor	Bayley-III scores (n=186)	0.78 (0.61 - 0.92)	0.80 (0.58 - 0.91)
	NPEU/Oxford (n=187)	0.71 (0.57 - 0.95)	0.72 (0.55 - 0.85)
Combined motor	Bayley-III scores (n=186)	0.69 (0.50 - 0.84)	0.71 (0.54 - 0.86)
	NPEU/Oxford (n=187)	0.41 (0.30 - 0.52)	0.74 (0.59 - 0.90)
Overall	Bayley-III scores (n=177)	0.41 (0.30 - 0.50)	0.44 (0.29 - 0.56)

*Combined communication impairment was judged as the worst category of outcome from receptive communication and expressive communication; combined motor impairment was judged as the worst category of outcome from fine motor and gross motor. Overall impairment was based on the worst category of outcome from the cognitive, communication and motor domains.

5.4.3 Post-hoc analysis of the validity of NHS assessments using a different question set to identify ‘moderate-severe’ impairment

The purpose of this post-hoc analysis was to assess whether by applying a broader criteria to define ‘moderate-severe’ impairment, the validity of the NHS data in identifying children with Bayley-III scores lower than -2 SD could be improved. Children were re-classified as having moderate-severe impairment if they meet the broader criteria in the NHS data (methods described in section 5.13.3). The results are displayed in table 5.8. Comparing this to the results shown in table 5.6, it was clear that the use of a broader category of moderate-severe impairment improved sensitivity of the NHS data, although this was at a cost of a small reduction in specificity. The biggest increase in sensitivity was observed in the cognitive and expressive communication domains.

Table 5.8 Sensitivities and specificities of the NHS data using a broader ‘moderate-severe’ impairment category in identifying participants with Bayley-III scores lower than -2 SD below mean

Domain of development*	Sensitivity (95% CI)	Specificity (95% CI)
Cognitive	71.4 (14.4 - 100.0)	91.7 (85.6 - 97.8)
Receptive communication	36.4 (0 - 79.8)	94.8 (93.1 - 96.5)
Expressive communication	91.7 (72.9 - 100.0)	85.2 (78.4 - 92.1)
Combined communication	66.7 (39.7 - 93.6)	85.8 (78.2 - 93.4)
Fine motor	50.0 (30.2 - 100.0)	99.5 (98.4 - 100.0)
Gross motor	88.9 (63.8 - 100.0)	98.3 (95.6 - 100.0)
Combined motor	80.0 (51.3 - 100.0)	98.3 (95.6 - 100.0)
Overall	72.0 (47.9 - 96.1)	86.3 (77.8 - 94.8)

5.4.4 Variables affecting the validity of the NHS assessments

As the diagnostic validity of the NHS assessment did not differ between the use of Bayley-III scores and NPEU/Oxford criteria for classifying impairment for the research assessment, subgroup analyses were performed using only the results from Bayley-III assessments. Results are presented in the

series of figures 5.8 to 5.18. The clustered bar charts show the observed prevalence of impairment based on the research assessment and the corresponding sensitivity and specificity of the NHS assessment, stratified by the factor under study. The prevalence of severe impairment and impairment in the receptive communication and fine motor domains was too low for meaningful subgroup analyses, hence only results from the cognitive, combined communication and combined motor domains and overall outcome are presented.

I observed lower prevalence of impairment with higher gestational age at birth across all domains, with apparent reduction in sensitivity but increased specificity of NHS assessment in identifying overall impairment with increasing gestational age (figure 5.8). The sensitivity in identifying cognitive impairment seemed higher if a standardised neurodevelopmental test was used during NHS assessment (figure 5.14). The accuracy in identifying impairment also appeared higher across all domains with increasing postnatal age at assessment (figure 5.13). However, as the confidence intervals for the estimated sensitivities and specificities overlapped widely, the observed effect of these factors on the diagnostic validity of NHS assessment can be conservatively deemed to be statistically insignificant (Schenker 2001). Similarly, there was no clear effect of the exposure to English language, the grade of the NHS assessor, IMD and the time interval between NHS and research assessments on the validity of NHS assessment.

5.4.5 Behaviour during assessments and the effect on study findings

15 (7.9%) children were highlighted as being 'difficult to test' during their NHS assessment. These children received significantly lower examiner rated behavioural score during the research assessment (median (IQR) score 23 (21 - 24) compared with 26 (23 - 26) for children who were not 'difficult to test'; $p < 0.001$), suggesting that the demonstration of challenging behaviour was consistent between assessments. However, the parents of children who behaved less positively

during the research assessment also reported that the children were demonstrating behaviour that were not typical of their usual self (median parent rated 'typical behaviour' score of 23 (21 - 26) for children scoring ≤ 22 on the examiner rated score; 26 (25 - 26) for children scoring ≥ 23 on the examiner rated score; $p < 0.001$).

The prevalence of impairment was significantly higher among children who were deemed difficult to assess during the NHS assessment (86.7% versus 31.7% for impairment in any domain; $p < 0.001$) or had received lower examiner rated behaviour scores (less positive behaviour) (73.8% for impairment in any domain among children with behaviour score ≤ 22 versus 28.5% among children with behaviour score > 22 ; $p < 0.001$). However, challenging behaviour demonstrated during assessments did not appear to affect the test validity of the NHS assessment against the research assessment (figures 5.17 and 5.18). The prevalence of impairment, sensitivity and specificity of NHS assessment did not differ by parent rated behaviour scores.

Figure 5.8 The prevalence of impairment and the estimated sensitivity and specificity of NHS assessment in identifying impairment in different neurodevelopmental domains, stratified by gestation at birth

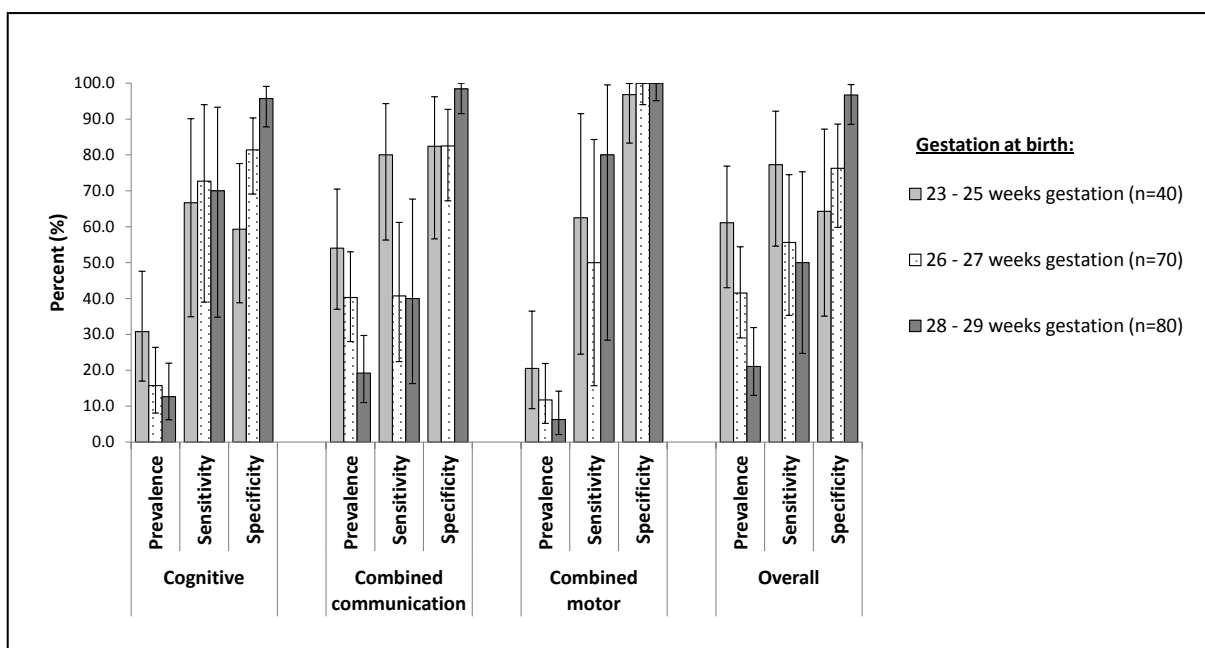


Figure 5.9 The prevalence of impairment and the estimated sensitivity and specificity of NHS assessment in identifying impairment in different neurodevelopmental domains, stratified by *sex of participants*

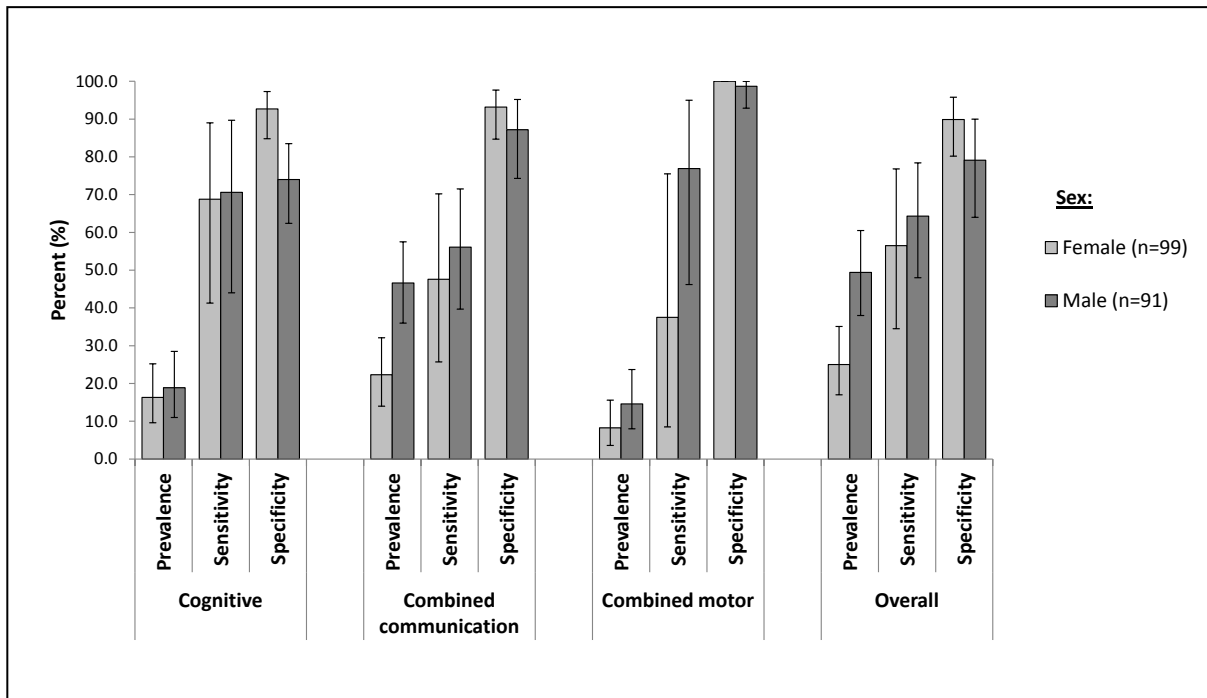


Figure 5.10 The prevalence of impairment and the estimated sensitivity and specificity of NHS assessment in identifying impairment in different neurodevelopmental domains, stratified by *the requirement for supplemental oxygen at 36 weeks corrected gestational age*

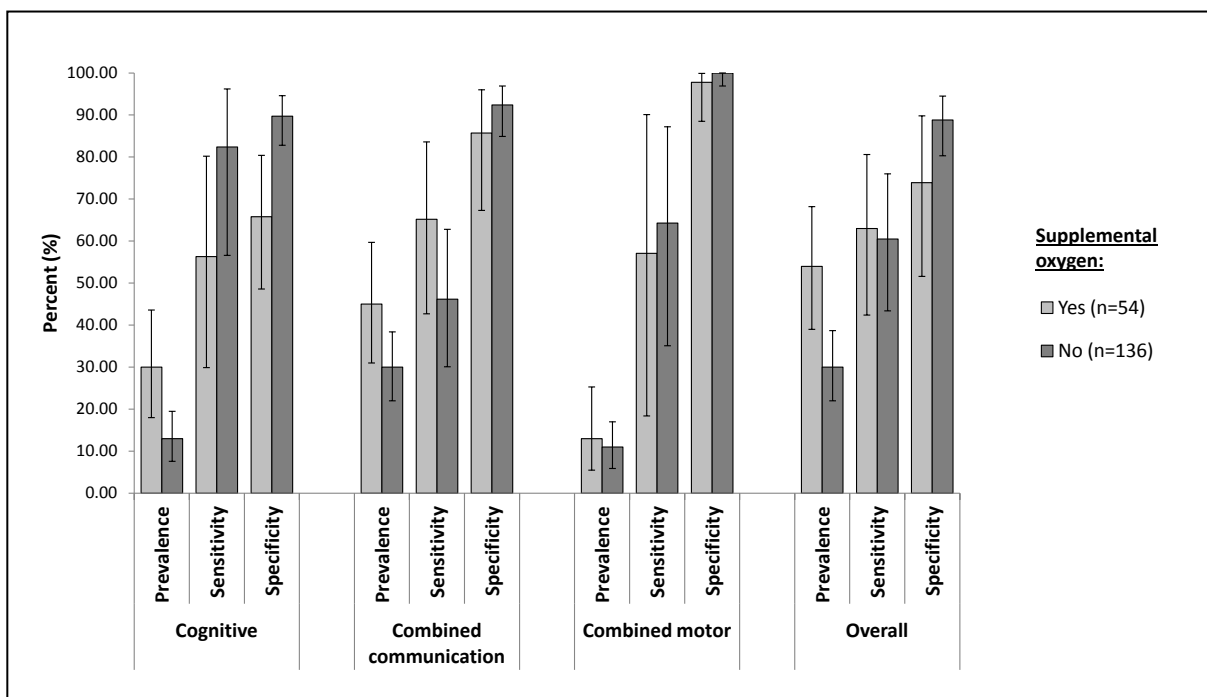


Figure 5.11 The prevalence of impairment and the estimated sensitivity and specificity of NHS assessment in identifying impairment in different neurodevelopmental domains, stratified by *the IMD quintile of residence at the time of assessment*

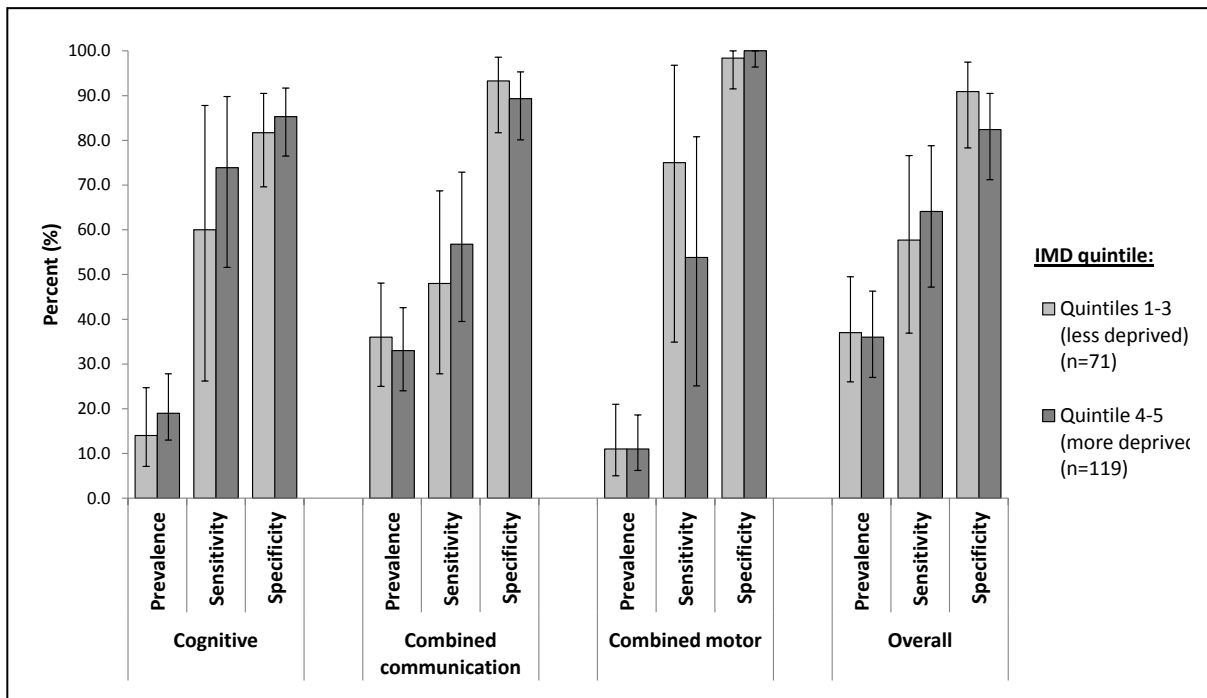


Figure 5.12 The prevalence of impairment and the estimated sensitivity and specificity of NHS assessment in identifying impairment in different neurodevelopmental domains, stratified by *whether English was the only language spoken at home*

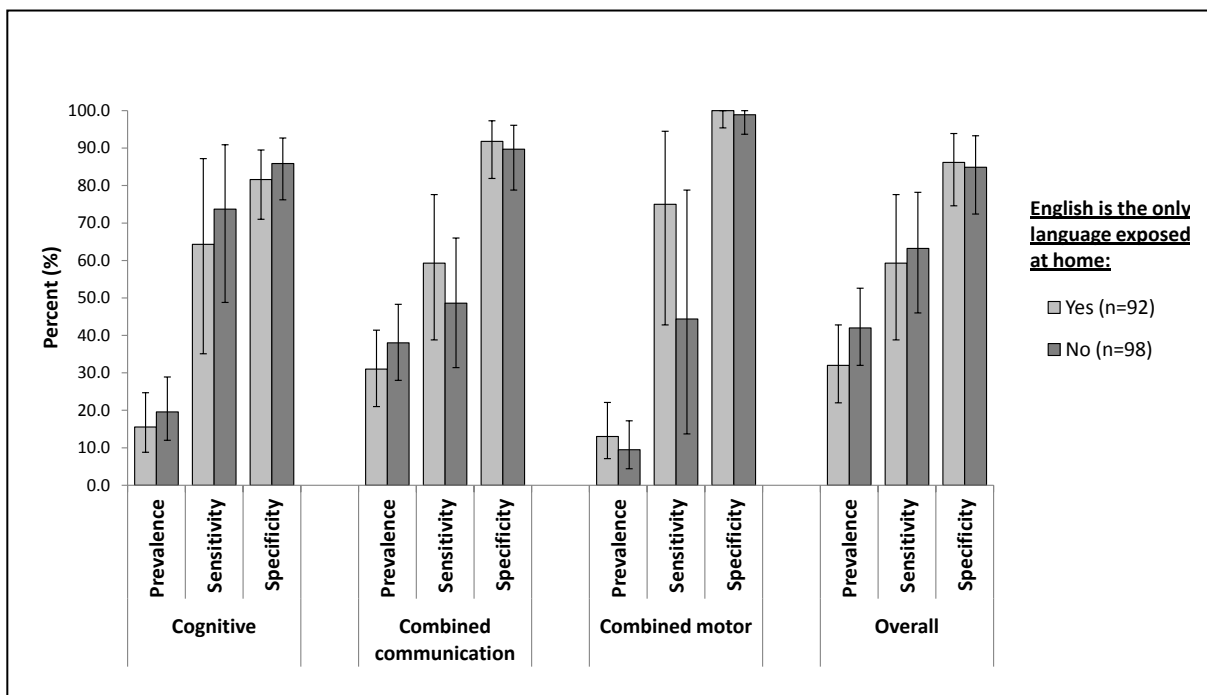


Figure 5.13 The prevalence of impairment and the estimated sensitivity and specificity of NHS assessment in identifying impairment in different neurodevelopmental domains, stratified by *the corrected age of participants at the time of the NHS assessment*

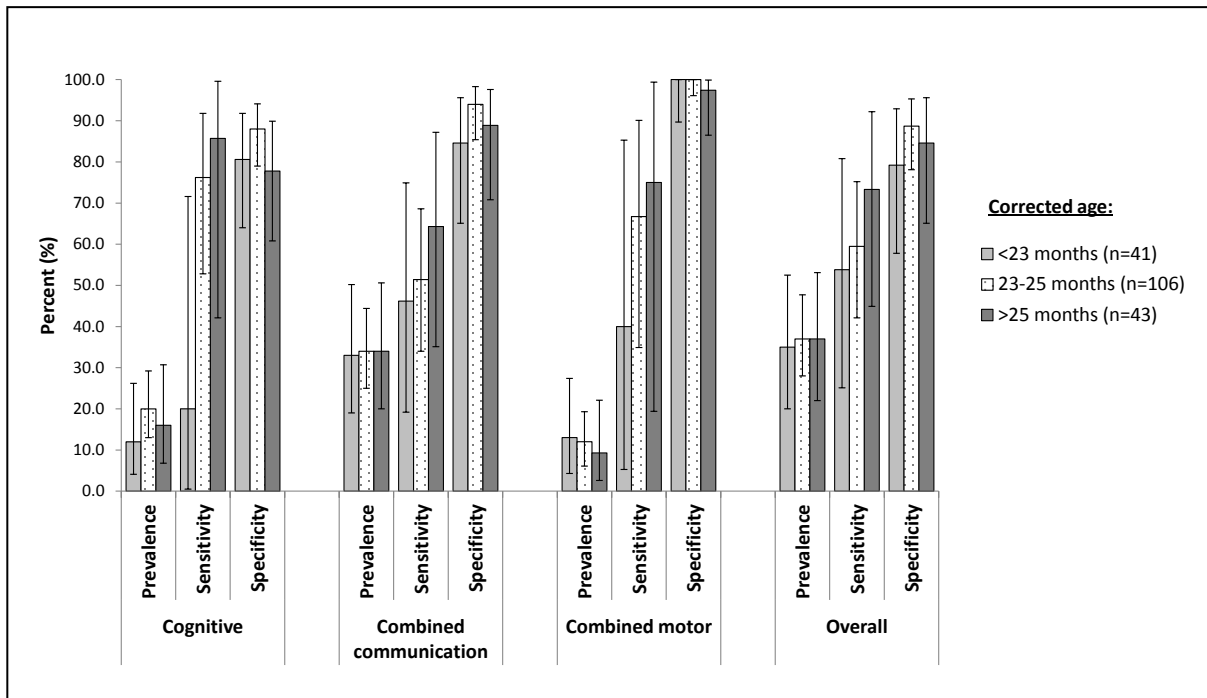


Figure 5.14 The prevalence of impairment and the estimated sensitivity and specificity of NHS assessment in identifying impairment in different neurodevelopmental domains, stratified by *whether a standardised neurodevelopmental test was used during the NHS assessment*

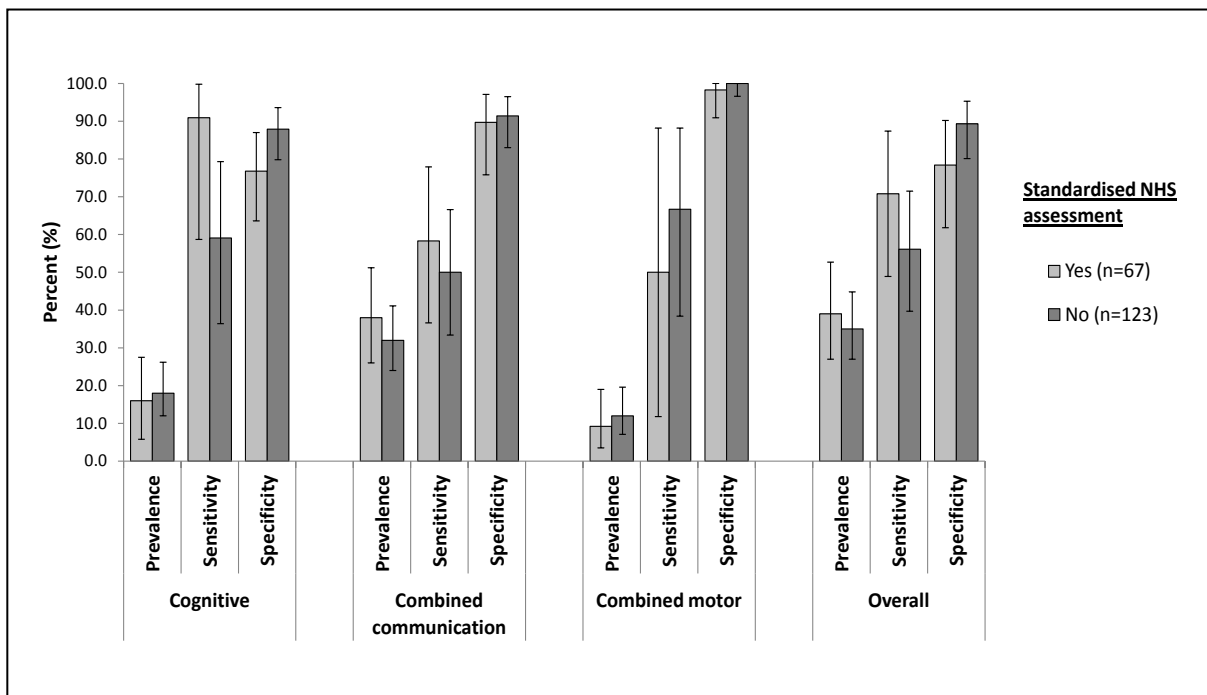


Figure 5.15 The prevalence of impairment and the estimated sensitivity and specificity of NHS assessment in identifying impairment in different neurodevelopmental domains, stratified by *the grade of the assessor who conducted the NHS assessment*

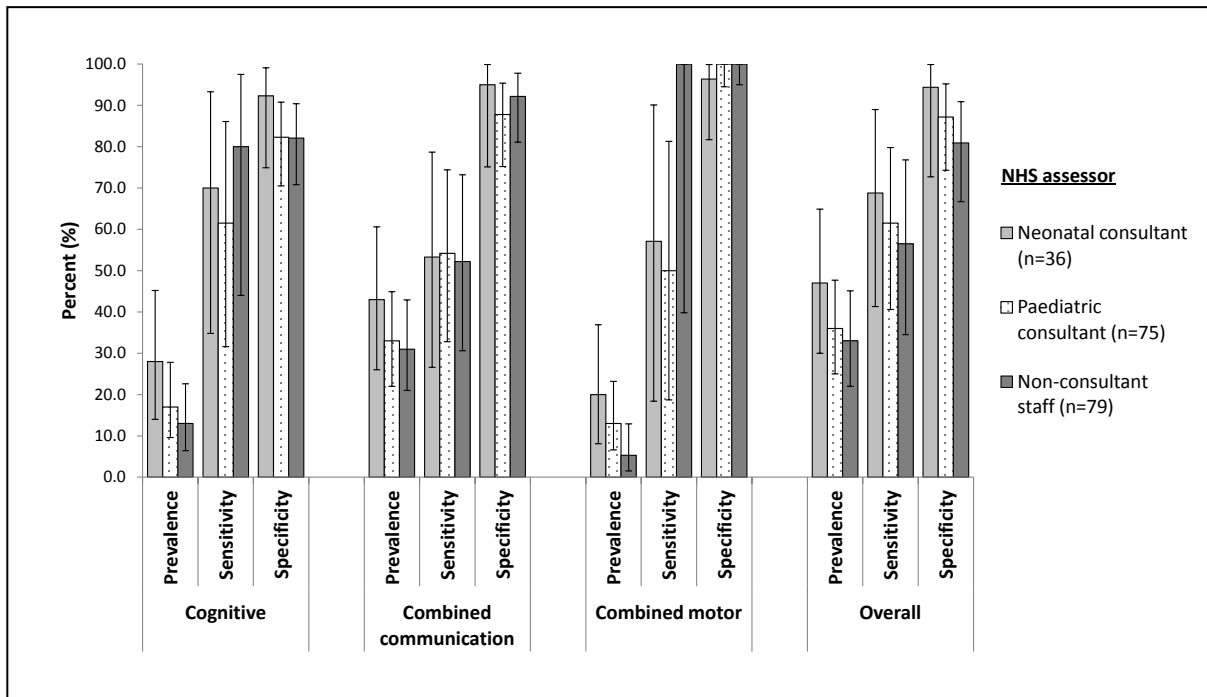


Figure 5.16 The prevalence of impairment and the estimated sensitivity and specificity of NHS assessment in identifying impairment in different neurodevelopmental domains, stratified by *the time interval between NHS and research assessments*

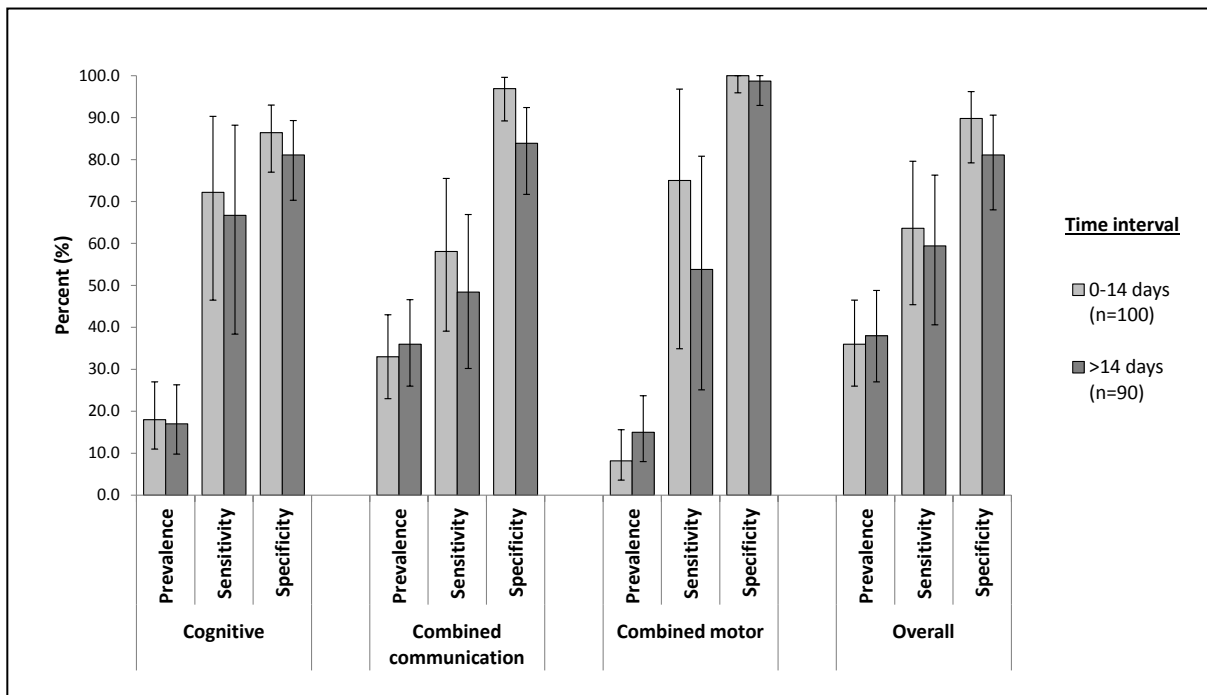


Figure 5.17 The prevalence of impairment and the estimated sensitivity and specificity of NHS assessment in identifying impairment in different neurodevelopmental domains, stratified by *the examiner rated behavioural score*

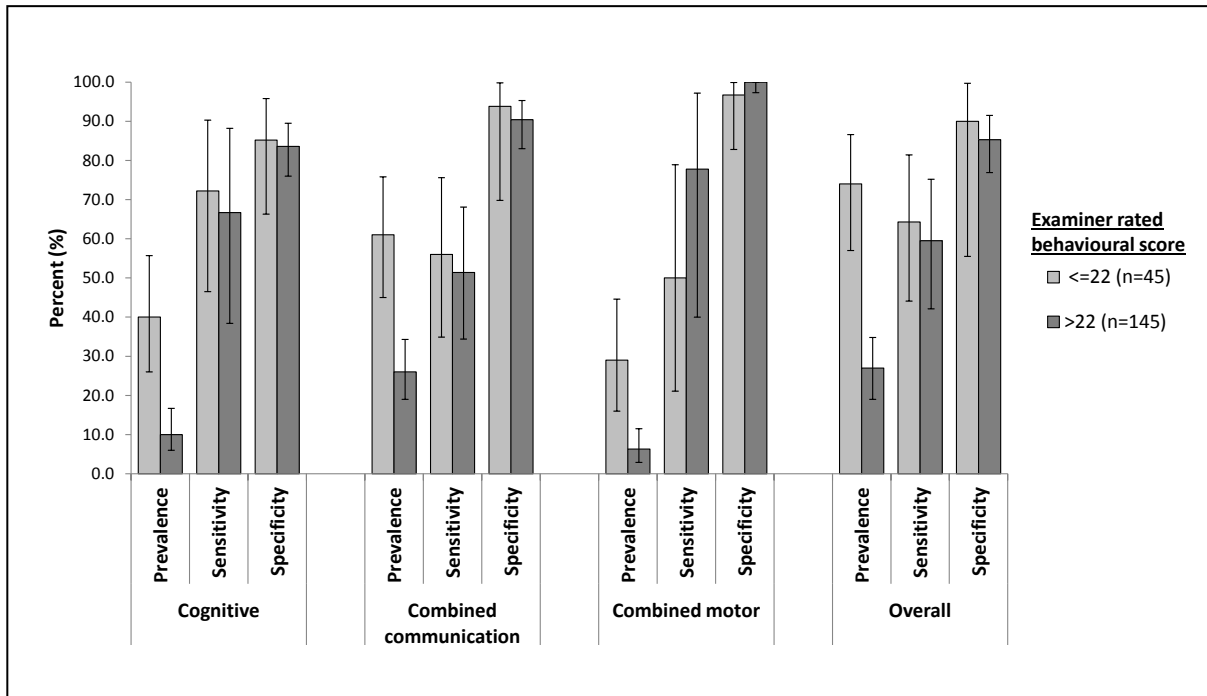
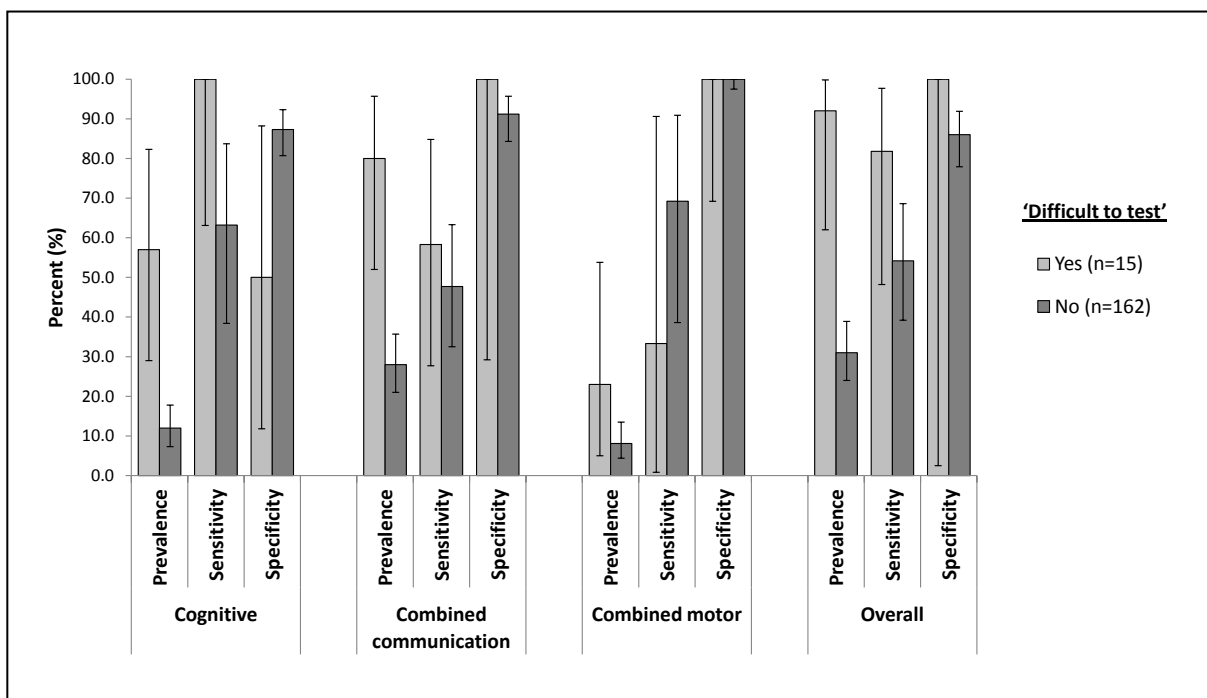


Figure 5.18 The prevalence of impairment and the estimated sensitivity and specificity of NHS assessment in identifying impairment in different neurodevelopmental domains, stratified by *whether the participant was considered 'difficult to test' during the NHS assessment*



5.5 RESULTS FROM THE HAMMERSMITH INFANT NEUROLOGICAL EXAMINATION (HINE) AND DIAGNOSIS OF CEREBRAL PALSY

Forty-seven (24.7%) children had a suboptimal global score (<73 of 78) on the HINE. In general, in the preterm population although scores below 73 are sub-optimal, those with scores above 64 will walk independently by two years, those below 64 but above 52 will sit independently and those below 52 will not be able to do either. The proportions of participants who achieved suboptimal score in each subsection were as follow: 30 (15.8%) for cranial nerve function, 39 (20.5%) for posture, 16 (8.4%) for movement, 36 (18.9%) for tone and 8 (4.2%) for reflexes.

Nine (4.7%) children were assessed to have cerebral palsy during the research assessment. The HINE scores for these children (median 53 (IQR 38.5 - 59.5), range 32 - 63) were significantly lower than those without cerebral palsy (78 (74 - 78), 65 - 78) ($p < 0.001$) and consistent with published data (Frisone 2002). Two children had spastic quadriplegia, five had spastic diplegia, one had 3-limb involvement and one had dyskinetic cerebral palsy. The gross motor function varied from GMFCS level 1 (walks without limitations) for one child with spastic diplegia to GMFCS level 5 (transported in manual wheelchair) for the child with dyskinetic cerebral palsy and one of the children with spastic quadriplegia. Most children with spastic diplegic cerebral palsy functioned at GMFCS level 2 (walks with limitations).

Two children with spastic diplegia were not identified to have cerebral palsy in the NHS data. The topographic classifications entered in the NHS data for all other children identified to have CP were in agreement with the research assessment.

5.6 DISCUSSION

5.6.1 Agreement between routine NHS data and research standard data

In this study, I found that the agreement in classifying neurodevelopmental status between data recorded during routine NHS assessments and that obtained through a research assessment was strong in the absence of neurodevelopmental impairment (high specificity; low rates of false-positive diagnoses). However, routine NHS assessments lack satisfactory sensitivity for identifying children with impairment, particularly in the cognitive and communication domains. Using the Bayley-III scores as the 'gold-standard' (this is discussed in section 5.6.4), I found that approximately 30% of children with at least mild cognitive impairments and nearly 50% with at least mild communication impairments were falsely classified as having no impairment in the NHS dataset. In the next section, I would like to consider the possible explanations for the results obtained.

5.6.2 Explanatory factors that affected the concordance of NHS and research data

I found good agreement between Bayley-III scores and the NPEU/Oxford criteria for classifying outcomes from the research assessment. Also, the discordance between the NHS and research data remained irrespective of the criteria used to categorise outcomes measured at the research assessment. This implied that the structural and content differences between the classification methods could not account for the discordance between the NHS and the research assessment.

I believe that inter-rater (assessor) variability in outcome assignment was one of the main reasons for the disagreement between NHS and research data. Clinical judgment is inevitably influenced by the assessor's knowledge, experience, beliefs and preconceptions. Studies on behavioural psychology have shown that people tend to rely on judgmental heuristics e.g. intuition, which are by nature unreliable, to simplify the complex task of assessing probabilities and predicting values in order to provide reasoning on the outcome of an event, say the diagnosis of disability in a child

(Tversky 1974). The use of standardised assessment tools improves inter-rater agreement by establishing objective measures. Without using a standardised assessment, the judgment and interpretation of clinical findings may be highly variable. It is therefore unsurprising that in a study comparing the diagnosis of cognitive impairment made using an intelligence test versus judgments by paediatricians, the agreement was only fair (κ 0.39) (Sondaar 2008). Even if standardised assessments were used, the agreement between the different tools in classifying impairment is unknown. Chaudhary et al reported that at 22 months, children obtained 5 points higher on the BSID-II MDI than the Griffiths Scales developmental quotient (Chaudhary 2013). Furthermore, the interpretation and translation of the standardised assessment scores into the NPEU/Oxford classification instruction can still be inconsistent and subject to biases and errors.

Another difference between the conduct of the NHS and research assessments that can account for the results was the reliance of routine assessments on parents' report on their child's ability, particularly for communication and cognitive skills. Parents are a valuable source of information in a time-restrained appointment, especially if the child was not engaging in the assessment, and can provide detailed insight into their child's level of functioning. However, studies investigating the level of agreement of neurodevelopmental status between parent's evaluation and paediatricians' assessments had reported variable results (Kim 1996, Bortolus 2002, Pritchard 2005, Johnson 2004) and parents' perception can be affected by cultural and socio-demographic influences.

Intra-subject (participant) variability in performance between assessments could also contribute to the discordance between NHS and research assessments. There are multiple factors such as mood and ease of engagement of the participant, time of the day (meals/ snacks or nap times) and environment that can influence the children's performance. Preterm children have been shown to be at risk of inattention/hyperactivity (Brogan 2014, Wilson-Ching 2013), and social-emotional delays (Boyd 2013) which could manifest as inability to complete a task. The testing time is also

generally longer for research assessments and it was not unusual for children to become tired during testing. These issues may not have been taken into account by the assessors and specifically, the objective scoring of a standardised assessment would not have made allowance for underperformance due to these factors.

I found that routine NHS assessments had higher sensitivities for diagnosing motor impairment compared with cognitive or communication impairments. There are several possible reasons for this. Important motor developmental milestones (such as sitting and walking) are reached at a relatively young age and parents and health professionals place great emphasis on checking that children achieve these milestones. Cerebral palsy is the most commonly quoted morbidity of preterm birth; therefore motor assessments are regularly performed at follow-up appointments. Cognitive and communication skills can be difficult to ascertain in a single setting, particularly without the use of standardised assessment tools, and can be affected by the issues of judgment and reporting bias discussed above. Furthermore, in the NPEU/Oxford classification, the categorising of cognitive impairment by 'number of months behind corrected age' introduced another level of variability. In addition, in the electronic 'two-year outcome' form, the term 'development' was used as the heading for the cognitive domain. As a result, there was misinterpretation amongst the NHS assessors that the questions in that category applied to 'overall development in all domains' rather than being specific to cognitive function, potentially leading to misclassification.

The impact of inter-rater and intra-subject variability would be exacerbated by the classification of neurodevelopment skills, a continuous trait, into categories of ability or impairment. Levels of abilities or skills near the 'cut-off' between categories are more difficult to discriminate and are at risk of being misclassified into higher or lower impairment categories. I would assume that the effect of misclassification had, to a certain degree, affected the agreement between the NHS and research data.

Using subgroup analyses, I investigated whether the validity of the NHS assessment could be affected by measurable neonatal and socio-demographic factors as well as factors related to the conduct of the assessments. However, given that the numbers of children with impairment ('true-positives') within each subgroup were small, it was likely that subgroup analyses were underpowered. Therefore, I cannot rule out the possibility that the negative findings were a reflection of type 2 errors ('false-negative'). In addition, as the number of children diagnosed with cerebral palsy was so low, I could not reliably compare the validity of NHS assessments in diagnosing cerebral palsy.

In the following sections, I will consider the methodological strengths and limitations of the study that may influence the validity of the results.

5.6.3 Strengths of study

One of the key strengths of the study is the involvement of thirteen different hospitals sites with a wide catchment area for recruitment. This meant I was able to study different NHS post-discharge follow-up practices and incorporate children from a wide range of background, thereby minimising selection bias. Another major strength of the study is that I was the only assessor performing the research assessment, which eliminated information (rater) bias due to poor agreement between assessors. I ensured that I was fully trained in the Bayley-III assessment techniques prior to commencing the study. I also maintain consistent standards in administering the assessments and underwent 'reliability checks' with Ms Betty Hutchon (Bayley-III trainer). During the assessments, both the NHS assessors and I were blinded to the results of the comparative assessment. In the analyses, I took into account the likelihood of data clustering by study sites and secondary to multiple births. I also considered the concerns raised in the literature about over-estimation of

abilities by the Bayley-III assessment and analysed how this could have affected the categorisation of impairment of the children.

5.6.4 Limitations affecting the internal validity of the study

An important limitation of the study was that the targeted sample size was not achieved. When I planned the study, I considered the intended sample size to be feasible as it was represented 50% of the cohort discharged from the participating hospitals. The final sample was equivalent to 18.3% of the discharged population. The recruitment of participants occurred at a steady rate over the two-year recruitment phase and there was no period of under-recruitment. The study was conducted in London where the population mobility is high and I believe a large proportion of children were lost to NHS follow-up. It was also possible that the local collaborators missed potential participants, particularly if their follow-up assessment was conducted by health professionals who were unaware of the study.

The sample size target was calculated with the desire to estimate the sensitivity of NHS assessment in correctly classifying children with severe impairment to a high precision, achieving a narrow 95% confidence interval with half-widths within 10%. In addition to sample size, the precision was also dependent on the actual value of the sensitivity estimate; the lower the sensitivity, the wider the confidence interval. Since the estimated sensitivity for diagnosing severe overall impairment was low (52.0%), based on the 14.1% prevalence of severe impairment among the study participants, a sample of at least 680 children providing independent (unclustered) observations would be necessary to achieved the intended precision. The increment in precision with increasing sample size followed the law of diminishing returns. It was therefore difficult to justify the continued provision of additional time and resources required to achieve the desired precision for the sensitivity estimate.

The study population differed from the population of very preterm children discharged from the participating hospitals as it consisted of proportionally more white children who were less likely to have been mechanically ventilated, diagnosed with chronic lung disease (broncho-pulmonary dysplasia) and/or were living in less deprived areas. It was conceivable, therefore, that the study population was at lower risk for neurodevelopmental impairment. The selection bias was introduced by attrition of children from routine NHS follow-up and the non-random recruitment method. I was not able to determine the proportion of children who did not attend their scheduled NHS appointment. In the literature, the proportions of very low birth weight children who were lost to follow-up or reviewed with difficulty in regional follow-up programmes were reported to be around 11% - 27% at 2 years (Tin 1998, Campbell 1993, Catlett 1993) and 25% at 5 years (Callanan 2001). Characteristics associated with dropping out from follow-up included non-white ethnicity, young maternal age and low socioeconomic and maternal educational levels (Catlett 1993, Callanan 2001, Campbell 1993, Wolke 1995). Ideally, a random sample of participants selected from a known sampling frame (e.g. list of all children with scheduled follow-up appointments) would provide the most representative study cohort. The study population, however, were formed through selection by the local collaborators (who invited the parents to participate) and following positive response to the invitation from parents who were motivated to participate. I know that some local collaborators omit recruitment of certain groups of children if they had preconceived beliefs that the child would not cooperate with the research assessment or that the parents would not agree to participate. This referral filter further augmented the selection bias.

The presence of selection bias predominantly affected the accuracy of the estimated prevalence of impairment in the population (i.e. impairment rates being higher or lower than the true rates in the target population). Traditionally, sensitivity and specificity were considered to be independent of disease prevalence (Gordis 2014). Consequently, the adverse effect of selection bias on the validity of this study can be regarded as minimal. However, a number of studies have shown that variation in

prevalence can result in either clinical or artefactual variation in test accuracy (Leeflang 2009, Brenner 1997). In this study, for example, the study population may be at lower risk for impairment than the target population. Since it is probably easier to diagnose impairment in severe cases compared with mild-moderate cases, a study population with lower spectrum of impairment could potentially have more false-negative or false-positive results.

Since the Bayley-III was published in 2006, several studies have raised concerns that when compared with the BSID-II, the Bayley-III was underestimating neurodevelopmental impairment (Anderson 2010, Lowe 2012, Moore 2012c, Vohr 2012). When the same group of preterm children were concurrently tested using both editions, the mean Bayley-III cognitive and language scores were between 7 and 18 points higher than the mean BSID-II MDI scores (Vohr 2012, Moore 2012c, Lowe 2012). These findings were inconsistent with the Flynn effect (Flynn 1999) and call into question the previous findings of morbidity using the older versions of the Bayley Scales as well as the validity of the current Bayley-III scales as the 'gold-standard' research tool. It is however, reassuring that validation studies of the Bayley-III showed that the scores are consistent with other revised ability tests such as the Wechsler Preschool and Primary Scale of Intelligence - Third Edition (Wechsler 2002) and the Preschool Language Scale - Fourth Edition (Zimmerman 2002) (Bayley 2006b). At the moment, the causes of the discrepancies between the two editions are unclear. Possible explanations suggested by the Bayley-III development team include differences in the demographic characteristics between the BSID-II and Bayley-III normative samples and the improved precision in norming methodology since the development of BSID-II (Bayley 2008). In this study, more children were classified as being impaired using the predicted BSID-II MDI scores than Bayley-III scores if the same threshold were applied. Therefore, as expected, the sensitivities of the NHS assessment dropped when the predicted BSID-II MDI were used as the 'gold-standard' instead of Bayley-III.

Another issue that needs to be considered is the impact of administering the English-based Bayley-III assessment on cognitive and language scores in children whose primary language was not English. Although I excluded families who require interpretation for English, 52% of the study population was living in bilingual environment; some of whom had limited exposure to English. This was a reflection of the multicultural population living in London. As the Bayley-III was designed to be administered in English, cautious scoring of a non-English-speaking child would result in a low score for the language domain. This could, in part, explain the lower mean language composite scores obtained in the study population compared with the composite scores for the cognitive and motor domains. Lowe et al compared the Bayley-III scores of 752 children from English-speaking homes with 98 children from Spanish-speaking homes (Lowe 2013). For Spanish-speaking children, the Bayley-III was administered in Spanish by either a bilingual examiner or an examiner with an interpreter. They found that, even after adjustment for medical and socio-economic factors, the language scores in Spanish-speaking children were 5.0 points lower than in English-speaking children. There was no significant difference between the two groups for cognitive scores. Studies that examined the effect of bilingualism on language acquisition had provided conflicting evidence (Paez 2007, Bialystok 2004, Walch 2009). I am therefore, unable to rule out testing bias, particularly in the language domain. However, for a child who was functioning in the 'severe impairment' category, communications skills were assessed by the observation of gestures and the production of consonant-vowel sounds, which are not language-specific. Hence, the overall effect of over-cautious scoring of non-English speaking children would be over-estimation of 'mild-moderate' communication impairment (leading to lower estimated sensitivities of NHS assessments) but the judgment of children having severe communication impairment was likely to be valid. It should be noted, that testing bias could also occur in the NHS assessments. Furthermore, NHS assessors were more likely to rely on parental reporting in these situations, which in turn lead to reporting/ recall bias.

5.6.5 Limitations affecting external validity (generalisability)

The fact that the study cohort consisted of a large proportion of children living in bilingual environment and in more deprived areas may limit the generalisability of the findings to other populations. Assuming that non-English speaking children and those from lower socio-economic groups were more challenging to test, hypothetically, the validity of NHS assessments in a different population could be better; however, this was difficult to predict.

5.7 CONCLUSIONS

The agreement in identifying preterm children with neurodevelopmental impairment and classifying the severity of impairment between data collected through routine NHS follow-up assessment and a research assessment using the Bayley-III was poor. This was likely to be due to inter-rater variability and systematic misclassification during the categorisation of a continuous outcome into levels of severity. I recognised bias in the study that could affect the validity of the results. The poor validity of NHS data collected routinely through the electronic 'two-year outcome' form limits this as a source for population outcome measures. There remains a need for other quick and cost-effective neonatal outcome reporting systems.

CHAPTER 6

STUDY 2 RESULTS: EARLY CHILDHOOD SOCIAL-COMMUNICATION DIFFICULTIES IN CHILDREN BORN PRETERM

6.1 CHARACTERISTICS OF STUDY POPULATION

The Q-CHAT questionnaire was sent to the parents of all 208 children who attended the research assessment in study 1. Ten children were assessed to have major functional impairments (nine with cerebral palsy; one with severe hearing impairment) and were ineligible for this study. The parents of three children who declined to participate in study 1 agreed to join this study and completed the Q-CHAT and Bayley-III SE questionnaires. A total of 150 questionnaires, including eight from children who were ineligible, were returned. One questionnaire with seven missing responses was treated as a non-respondent and excluded, leaving 141 participants (70.1% of eligible participants) for the analyses.

In table 6.1, I compared the neonatal and sociodemographic characteristics between respondents, non-respondents and the 'baseline population' (all infants born between 1 January 2008 and 31 December 2010, at gestational ages below 30 weeks, and discharged from the participating study sites, as identified on the NNRD). Non-respondents were more likely to be parents of girls (66.7%, $p = 0.02$). Nonetheless, both boys and girls were equally represented in the respondent group. Similar to my observation in study 1, compared with the 'baseline' population, the respondent group consisted of an over-representation of children of white ethnicity, born to mothers living in less deprived IMD quintiles, had significantly shorter duration of mechanical ventilation and were less likely to have required supplemental oxygen therapy at 36 weeks' corrected age.

The mean corrected age of the respondents was 24.7 (SD 2.6, range 18.5 to 35.6) months at the time of completion of the questionnaires. The mean (SD) Bayley-III composite score of the 138

respondents who completed the assessment was 94.6 (13.0) for the cognitive scale, 87.7 (13.0) for the language scale and 98.0 (10.1) for the motor scale.

Table 6.1 Comparing the characteristics of respondents, non-respondents and non-participants born <30 weeks gestation in 2008-2010 and discharged from the participating study sites

Characteristics	Respondents (n=141)	Non-respondents (n=60)	'Baseline population' (n=1,037)	<i>p</i> -value	
				Respondents vs non-respondents	Respondents vs 'baseline' population
Gestation					
(completed weeks)	27 (26 - 29),	27 (26 - 28),	27 (26 - 29),		
Median (IQR), range	23 - 29	23 - 29	22 - 29	0.15	0.58
Birth weight (g)					
	958	920	1000		
Median (IQR), range	(810 - 1167), 490 - 1720	(740 - 1082), 560 - 1400	(812 - 1200), 455 - 1990	0.07	0.44
Sex					
Female, n (%)	68 (48.2)	40 (66.7)	444 (42.8)		
Male, n (%)	73 (51.8)	20 (33.3)	503 (48.5)	0.02	0.09
Missing, n (%)	0 (0.0)	0 (0.0)	90 (8.7)		
Ethnicity					
White, n (%)	66 (46.8)	21 (35.0)	364 (35.1)		
Non-white	75 (52.2)	39 (65.0)	611 (58.9)	0.12	0.03
Missing, n (%)	0 (0.0)	0 (0.0)	62 (6.0)		
Pregnancy					
Singleton, n (%)	110 (78.0)	48 (80.0)	690 (66.5)		
Multiples, n (%)	31 (22.0)	12 (20.0)	250 (24.1)	0.67	0.25
Missing, n (%)	0 (0.0)	0 (0.0)	97 (9.4)		
Mode of delivery					
Vaginal, n (%)	61 (43.3)	22 (36.7)	475 (45.8)		
Caesarean, n (%)	71 (50.4)	32 (53.3)	540 (52.1)	0.34	0.84
Missing, n (%)	9 (6.4)	6 (10.0)	22 (2.1)		
Maternal age (years)					
Mean (SD)	31.5 (6.0)	31.9 (7.8)	31.0 (6.4)	0.68	0.35
IMD quintile at birth					
One, n (%)	13 (9.2)	6 (10.0)	43 (4.2)		
Two, n (%)	13 (9.2)	5 (8.3)	81 (7.8)		
Three, n (%)	20 (14.2)	5 (8.3)	144 (13.9)	0.77	0.05
Four, n (%)	42 (29.8)	17 (28.3)	268 (25.8)		
Five, n (%)	53 (37.6)	27(45.0)	477 (46.0)		
Missing, n (%)	0 (0.0)	0 (0.0)	24 (2.3)		
Length of mechanical ventilation (days)					
Median (IQR), range	1 (0 - 3), 0 - 54	1 (0 - 7), 0 - 61	4 (0 - 18), 0 - 444	0.13	<0.001
Oxygen therapy at 36 weeks' corrected age					
Yes, n (%)	38 (27.0)	23 (38.3)	466 (44.9)		
No, n (%)	103 (73.1)	37 (61.7)	571 (55.1)	0.11	<0.001

SD = standard deviation; IMD = index of multiple deprivation

6.2 Q-CHAT SCORES OF THE PRETERM POPULATION

The Q-CHAT scores of the preterm study population (mean 33.7, SD 8.3, range 15 to 55) were normally distributed and significantly higher (less favourable) than the published general population scores (difference in means = 7.0 (95% CI 5.6 to 8.3); $p < 0.001$) (Figure 6.1). The mean Q-CHAT scores were 33.8 (SD 7.8, range 15 to 55) for preterm boys and 33.5 (SD 8.8, range 15 to 54) for preterm girls. When compared with the general population, sex-specific scores in both preterm boys and girls were significantly higher (mean difference 6.3 (95% CI 4.5 to 8.1) for boys, 7.7 (95% CI 5.6 to 9.8) for girls; $p < 0.001$ for both sexes). In contrast to the higher scores described in boys in the general population, no sex difference in Q-CHAT scores were observed in the preterm population ($p = 0.85$). There was no correlation between Q-CHAT scores and the corrected ($p = 0.21$) and uncorrected ($p = 0.36$) ages at assessments. There was no difference in the mean Q-CHAT scores between respondents whose parents had completed the questionnaire before and those who had completed the questionnaire after their observation of the Bayley-III assessment ($p = 0.84$).

The distribution of item-specific scores by category of autistic behaviour is displayed in Table 6.2. The distribution of scores between the preterm study cohort and the general population differed significantly in 17 items. In all these items, there were greater proportions of preterm children receiving higher scores, indicating greater social-communication difficulties and autistic behaviour characteristics. The differences were most predominant in the categories of restricted, repetitive, stereotyped behaviour (7 of 9 items differ significantly), communication (3 of 4 items) and sensory abnormalities (all 3 items). Only 4 out of the 9 items exploring social-relatedness were scored differently in the preterm population.

Figure 6.1 Histogram of Q-CHAT scores of the preterm study population with superimposed distribution of published Q-CHAT scores of unselected toddlers (general population)

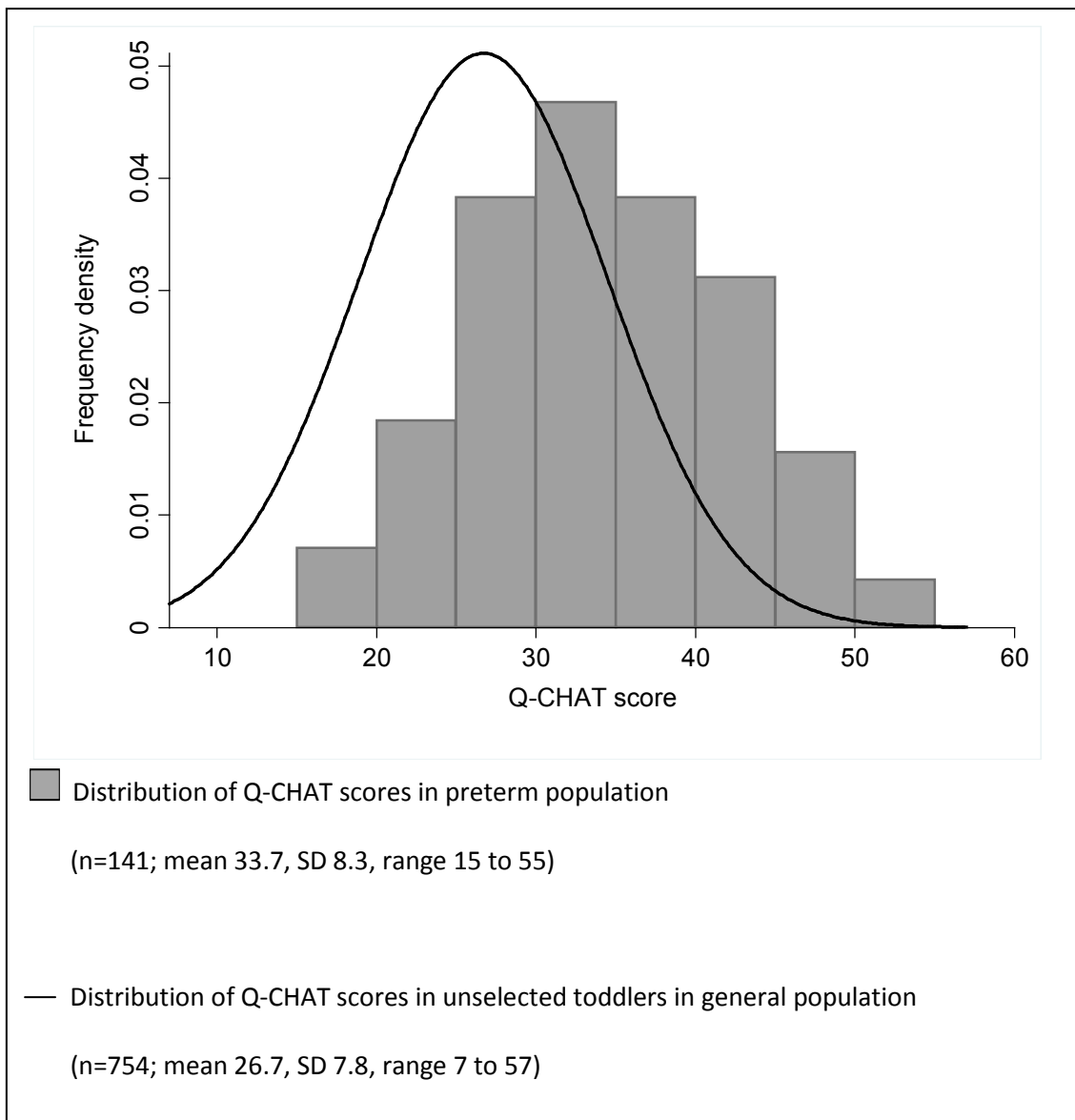


Table 6.2 Item-specific distribution of Q-CHAT scores

Q-CHAT item	Score (% of responses)					Difference in distribution compared with general population (<i>p</i> -value)
	0	1	2	3	4	
Items exploring social-relatedness:						
1. Look when name is called*	49.6	44.7	5.0	0.0	0.7	0.13
2. Eye contact*	52.5	44.0	2.8	0.7	0.0	0.004
3. Protoimperative pointing*	68.8	25.5	3.5	0.7	1.4	0.58
4. Protodeclarative pointing*	61.7	24.1	11.3	0.7	2.1	0.56
5. Pretend play*	64.5	22.0	7.8	2.1	3.5	0.02
6. Follow a gaze*	53.6	37.1	7.1	0.7	1.4	0.45
7. Offer comfort	34.8	32.6	24.6	4.3	3.6	0.04
8. Use simple gestures*	76.6	18.4	3.5	0.7	0.7	0.28
9. Check reaction	33.3	32.6	22.0	9.9	2.1	<0.001
Restricted, repetitive, stereotyped behaviour:						
10. Line objects up†	7.1	10.6	40.4	24.1	17.7	<0.001
11. Interest maintained by spinning object†	16.8	43.1	26.3	6.6	7.3	<0.001
12. Adapt to change in routine*	23.4	58.2	17.0	1.4	0.0	<0.001
13. Do the same thing over and over again	8.5	7.8	13.5	23.4	46.8	<0.001
14. Echolalia	2.8	3.5	19.1	25.5	48.9	0.06
15. Unusual finger movement†	53.2	10.1	12.9	13.7	10.1	<0.001
16. Maintenance of interest†	23.9	24.6	30.4	13.0	8.0	<0.001
17. Twiddle objects repetitively†	33.1	14.4	17.3	24.5	10.8	<0.001
18. Stare at nothing with no purpose*	59.6	22.7	9.2	5.0	3.5	0.15
Communication abnormalities:						
19. Understand child's speech	18.4	36.9	31.2	12.1	1.4	<0.001
20. Number of words	17.9	17.9	41.4	19.3	3.6	0.69
21. Typicality of first words*	56.2	34.3	4.4	1.5	3.6	0.02
22. Use of hand as tool	7.8	5.0	10.6	36.2	40.4	<0.001
Sensory abnormalities:						
23. Sniff or lick unusual objects	15.9	14.5	20.3	29.0	20.3	<0.001
24. Walk on tiptoe†	13.5	22.0	46.8	9.9	7.8	<0.001
25. Oversensitive to noise†	23.4	36.9	21.3	9.2	9.2	<0.001

*Chi-square test was performed by combining proportions with scores 2, 3 and 4.

†Chi-square test was performed by combining proportions with scores 3 and 4.

6.3 ASSOCIATION OF Q-CHAT SCORES WITH BAYLEY-III COGNITIVE, LANGUAGE AND MOTOR SCORES

With univariable analyses, Q-CHAT scores were significantly associated with Bayley-III cognitive, language and motor composite scores. When the variables were input into a multivariable regression model, the effect of cognitive and motor function on Q-CHAT scores was no longer significant ($p = 0.18$ for cognitive scores; $p = 0.67$ for motor scores). Bayley-III language composite scores independently predicted Q-CHAT scores in a linear fashion (correlation coefficient -0.51 ; $p = 0.001$) and accounted for 24.5% of the variance in Q-CHAT scores. The observed relationship between language and Q-CHAT scores was entirely due to expressive communication ability. The regression coefficient between Bayley-III expressive communication subscale scores and Q-CHAT scores was -1.35 (95% CI -1.96 to -0.74 ; correlation coefficient -0.43 ; $p < 0.001$). There was no association between receptive communication ability and Q-CHAT scores ($p = 0.22$). Apart from the expected association between Bayley-III language scores and sub-categorical Q-CHAT scores from items exploring communication skills, no significant association was found between cognitive, language and motor scores and other sub-categorical Q-CHAT scores.

6.4 NEONATAL AND SOCIODEMOGRAPHIC PREDICTORS OF Q-CHAT SCORES

The results of the univariable analysis for all the neonatal and sociodemographic variables examined are listed in table 6.3. Non-white ethnicity and living in deprived areas were found to be associated with higher Q-CHAT scores in univariable analyses. Although non-white children were more likely to live in areas of higher deprivation (test for trend $p < 0.001$, data not shown), there was no interaction between ethnicity and IMD in the association with Q-CHAT scores ($p = 0.72$). As lower Bayley-III language scores were observed among non-white children (mean difference 7.31 (95% CI 3.07 to 11.5 ; $p < 0.001$)) and children living in more deprived areas (mean decrease of 1.89 (95% CI 0.24 to 3.55 ; $p = 0.03$) points per IMD quintile increase in deprivation), language ability was considered to be a potential confounder in the relationship between ethnicity, IMD and Q-CHAT scores. There was no interaction between Bayley-III language scores and IMD quintiles ($p = 0.88$) or ethnicity ($p = 0.51$).

The final multivariable regression model included all variables found to be statistically significant during univariable analysis (Bayley-III language composite score, ethnicity and IMD) and is displayed in table 6.4.

Table 6.3 Univariable association of neonatal and sociodemographic factors with Q-CHAT scores

Variable	<i>n</i>	Q-CHAT score coefficient (95% CI)	<i>p</i> -value
Gestation (per completed week)	141	-0.77 (-1.61 to 0.06)	0.07
Birthweight z-score (per point increase)	126	0.07 (-1.41 to 1.55)	0.09
Male sex	141	0.27 (-2.54 to 3.07)	0.85
Singleton pregnancy	141	3.80 (-0.42 to 8.01)	0.08
White ethnicity	141	-7.55 (-10.2 to -4.86)	<0.001
Maternal age (per year)	141	-0.17 (-0.40 to 0.07)	0.17
Caesarean section delivery	132	-2.22 (-5.38 to 0.93)	0.17
Length of mechanical ventilation (per day)	139	0.10 (-0.01 to 0.20)	0.07
Supplemental oxygen requirement at 36 weeks post-menstrual age	141	1.30 (-2.06 to 4.67)	0.45
IMD quintile (per quintile increase in deprivation)	141	2.07 (1.04 to 3.11)	<0.001

Table 6.4 Final multivariable model of factors associated with Q-CHAT scores

Variable	Q-CHAT score coefficient (95% CI)	<i>p</i> -value
Bayley-III language composite score (per point)	-0.23 (-0.33 to -1.39)	<0.001
White ethnicity	-5.30 (-7.92 to -2.67)	<0.001
IMD quintile (per quintile increase in deprivation)	0.96 (-2.00 to 0.08)	0.07

n = 136 r^2 = 0.38

6.5 COMPARISON OF Q-CHAT AND BAYLEY-III SOCIAL-EMOTIONAL SCORES

The Bayley-III SE questionnaire was completed in 140 out of the 141 eligible respondents to the Q-CHAT questionnaires. These 140 participants form the cohort for the analysis to compare the two questionnaires. The Bayley-III SE score distribution of the preterm population (mean 97.8, SD 17.2, range 55 to 145) did not differ significantly from the standardised norm of mean 100 and SD 15 ($p=0.12$) (figure 6.2). Figure 6.3 is the scatterplot that shows the relationship in the distribution of Q-CHAT and Bayley-III SE scores. It is important to note that the scales for the two scores are different and whilst for the Q-CHAT, higher scores represent higher frequency of autism symptoms, for the Bayley SE, lower scores reflect greater social-emotional immaturity. The correlation between the two scores was relatively weak (correlation coefficient = -0.38).

Figure 6.2 Histogram of Bayley-III Social-Emotional composite scores of the preterm population with superimposed distribution of the standardised norm scores (mean 100, SD 15)

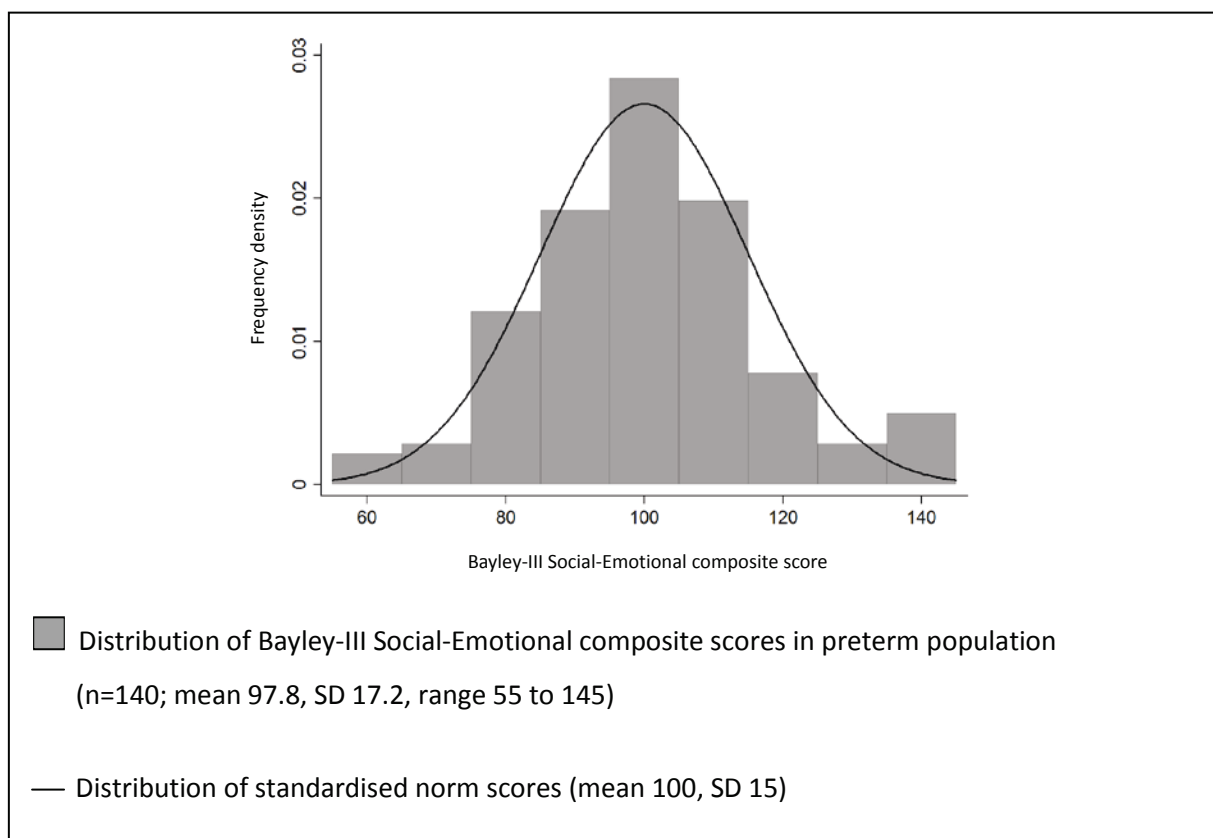
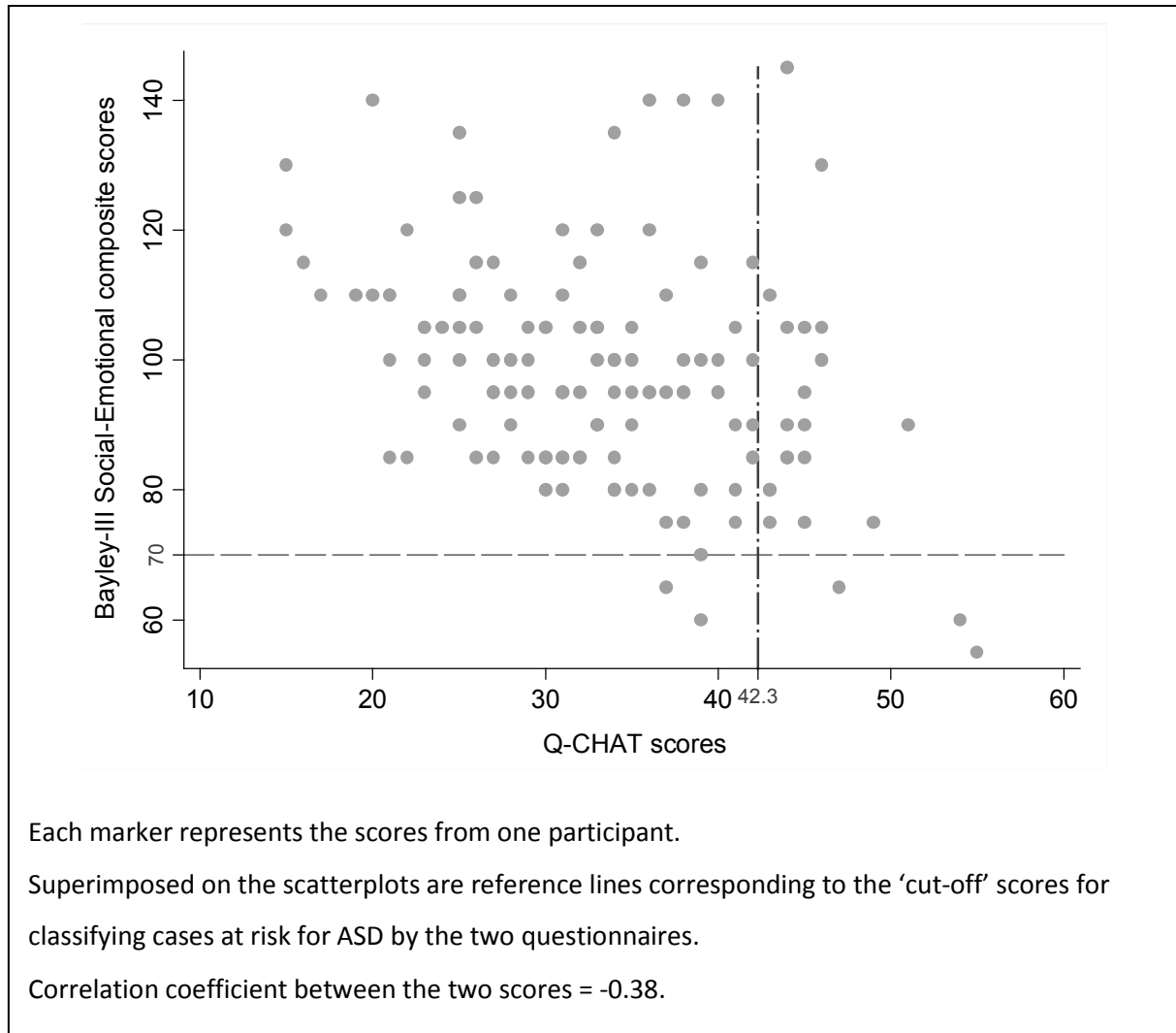


Figure 6.3 Scatterplot showing the relationship in the distribution of Q-CHAT and Bayley-III Social-Emotional composite scores of the study population



Twenty-three (16.4%) children had Q-CHAT scores higher than two SD above the general population mean (i.e. higher than 42.3). Only five (3.6%) children scored lower than two SD below the standardised mean (i.e. lower than 70) for the Bayley-III SE scale. If these children were classed to be 'at risk' for ASD, table 6.5 is a cross-tabulation that revealed the agreement in the classification of ASD risk between the two questionnaires. There was poor concordance between the questionnaires with only three children (2.1%) classed to be 'at risk' for ASD by both questionnaires and a resulting Cohen's κ coefficient of 0.17 (0 - 0.36). As can be seen from the scatterplot, the Bayley-III SE scores for the children who were classified as 'at risk' for ASD by Q-CHAT scores ranged widely from 55

(lowest possible score) to 145 (highest possible score). Similarly, children who scored <70 (lowest functioning) on the Bayley-III SE did not represent the children with the highest (most autism symptoms) Q-CHAT scores.

Table 6.5 Cross-tabulations of the number of children classified to be ‘at risk’ for ASD by the Q-CHAT and the Bayley-III SE questionnaires

		Bayley-III SE		
		‘At risk’	Not ‘at risk’	Total
Q-CHAT	‘At risk’	3	20	23
	Not ‘at risk’	2	115	117
	Total	5	135	140

6.6 DISCUSSION

6.6.1 Social-communication skills of children born very preterm

In this study, I demonstrated that at 24 months corrected age, children born <30 weeks gestation were rated by their parents as having greater social-communication difficulties and autistic traits compared with the general population. Utilizing properties of the Q-CHAT as quantitative measures of autistic features, I showed that Q-CHAT scores in the preterm population, although following a normal distribution with similar variability as the general population, were shifted to the right, yielding higher, less favourable, scores. These findings corroborate the report by Johnson et al of a similar right-shift in frequency distribution of ASD symptoms in preterm children compared to term-born classmate controls as measured by the Social Communication Questionnaire at age 11 years (Johnson 2010a).

The study sample was comprised of children without major functional disability. Previous studies have reported significantly higher odds of positive autism screening on the M-CHAT in children with

motor, visual, hearing and cognitive impairments (Kuban 2009, Moore 2012d). Some questions on the Q-CHAT also depended on intact neurosensory and motor functions and children with disabilities in these areas could be expected to receive higher Q-CHAT scores on these questions. Therefore, it is likely that the distribution of Q-CHAT scores in this very preterm population would be even higher if children with cerebral palsy and severe neurosensory disabilities were included.

In the general population, there was a 1.7 point sex difference in mean scores with the scores for boys being significantly higher than the scores for girls. I did not find a sex difference in the study population. This may be due to insufficient statistical power given that a sample size of 24,000 children would be required to detect the 0.3 point sex difference in Q-CHAT scores as significantly different. Nevertheless, it has been suggested that the autistic phenotype seen in preterm children resembles more closely that of children with syndromic or medically-explained autism where the sex ratio is closer to 1:1 than those with idiopathic autism (Kuban 2009), supporting the hypothesis that autism in preterm children, rather than being a primary deficit, represents part of a 'preterm phenotype' with different aetiology.

By analysing the differences in the responses given on each questionnaire item between the preterm and the general populations, I found that preterm children experience difficulties across all aspects of autistic behaviour but particularly in the categories of restricted, repetitive, stereotyped behaviour, communication and sensory abnormalities. The presence of reduced language abilities among children born preterm is well-described (Barre 2011, van Noort-van der Spek 2012). Dysfunction in sensory modulation in preterm children, characterized by either hyposensitivity or hypersensitivity to sensory input, is a problem anecdotally recognized by parents and clinicians. There is, however, a paucity of studies in this area. It is hypothesized that the premature exposure to the stressful and multisensorily enhanced environment of the neonatal intensive care unit at a critical period of brain development in the third trimester interferes with the normal maturation of

the sensory system (Als 1986, Bar-Shalita 2008). Sensory modulation dysfunction is thought to be negatively associated with emotional development and can affect social-interactive capabilities (Bart 2011).

There is some evidence that restricted and repetitive behaviours are associated with cognitive status (Bishop 2006, Ozonoff 2008). EPICure study investigators also concluded that cognitive deficits in their extremely preterm cohort accounted for the excess of repetitive and stereotyped behaviour compared to term controls (Johnson 2009). Although I did not demonstrate a correlation between cognitive scores and sub-categorical Q-CHAT scores in the restricted and repetitive behaviour domain, as the mean cognitive score of the preterm population was lower than would be expected in the general population, the potential association between cognition and restricted and repetitive behaviour could in part explain the higher Q-CHAT scores obtained by the participants in this category.

There were fewer differences between preterm children and the general population in response to items exploring social-relatedness. Previous work on early autism screening highlighted the absence of pretend play and joint-attention as strong predictors for later diagnosis of autism (Baron-Cohen 1992). I would speculate that Q-CHAT items exploring social-relatedness may provide a higher degree of specificity for differentiating early autistic features from concurrent developmental delay in children without severe physical and neurosensory impairment compared with items in the other categories. Although parents reported a lower frequency of pretend play among the preterm children, development in joint-attention (elucidated by questions on protodeclarative pointing and following a gaze) were similar to the general population. Focusing on elucidating social-relatedness for autism screening in the preterm population may reduce the 'false-positive' screening rate associated with currently available screening tools.

The significant association observed between language ability at age two years and Q-CHAT scores was unsurprising as four items on the Q-CHAT specifically examined language development. Furthermore, language ability was closely related to cognitive function which could in turn, influenced performance on other Q-CHAT items. Previous studies examining the M-CHAT in preterm children had utilised composite developmental scores such as BSID-II MDI (Bayley 1993) and the Parent Report Composite score from the PARCA-R (Johnson 2008), which encompassed assessments in both language and general cognition, to describe the association between cognitive ability and autistic symptoms (Kuban 2009, Moore 2012d). Both studies reported that cognitive impairment conferred a 3- to 4-fold increased risk for screening positive on the M-CHAT. More recent studies had utilised the Bayley-III, which provides separate cognitive and language scores, to examine the association between cognitive and language ability with positive M-CHAT screen (Gray 2015, Stephens 2012). Whilst Stephens et al reported that both language and cognitive impairment were independently associated with positive M-CHAT screen, Gray et al found that neither language nor cognitive scores affected results on the M-CHAT. In my study, I found that language ability confounded the association observed between cognitive scores and Q-CHAT scores in the univariable analysis. No independent association between cognitive and Q-CHAT scores were observed in the multivariable regression model that included scores from all Bayley-III domains. This observation may have arisen due to the emphasis of the Q-CHAT on identifying language delay as a core feature of ASD and insufficient weight being placed on elucidating cognitive impairment as a comorbidity in ASD.

This study also highlights the inter-relationship between ethnicity, area deprivation, language skills and Q-CHAT scores. Despite the obvious limitations of this approach, as ethnicity is correlated with area deprivation, it is often used as a proxy measure for socioeconomic status. The relationships between Q-CHAT scores, ethnicity and socioeconomic status in the general population are unknown. My findings suggest the possibility of an environmental impact of socioeconomic disadvantage on

early social-communication development. It could also represent ethnic and cultural differences, language abilities and other socioeconomic influences on parental reporting on the Q-CHAT. I lack information on individual socioeconomic status and there could be ecologic fallacy in using IMD. The reliability of the Q-CHAT questionnaire across diverse ethnic and socioeconomic groups will need to be further determined.

6.6.2 Early screening for autism spectrum disorders

The Q-CHAT and the Bayley-III SE are developmental surveillance tools designed to identify toddlers at risk for developing ASD, with the aim of implementing timely intervention strategies to achieve better outcomes for these children. The validity of the Q-CHAT for ASD screening has not been established. The Bayley-III SE questionnaire, using a scaled score of 6, reportedly had a sensitivity of 87.0% and specificity of 90.0% for the identification of ASD in the general population (Casenhiser 2007). In this study, I found that whilst the preterm population achieved higher (more unfavourable) Q-CHAT scores compared with the general population, the distribution of Bayley-III SE scores in the preterm population did not differ significantly from standardised norms. The result is surprising and to my knowledge, the Bayley-III SE is the only ASD screening tool that had been applied to a preterm population that did not reveal higher frequency of ASD symptoms in the population. This raises concerns that the Bayley-III SE may have poor psychometric properties in identifying preterm children with ASD.

I used an arbitrary cut-off of scores two SD from the general population or standardised mean to classify children 'at risk' for ASD by the Q-CHAT and the Bayley-III SE. The resulting 'positive screen' rates in my study population are 16.4% by the Q-CHAT and 3.6% by the Bayley-III SE. The positive screen rates observed on the Q-CHAT is comparable to that achieved using the M-CHAT on cohorts born ≤ 30 weeks gestation (13.4%) (Gray 2015) and VLBW (17.8%) (Dudova 2014). Although previous studies using the M-CHAT have reported higher prevalence of positive screens between 21%-41%

(Kuban 2009, Limperopoulos 2008, Moore 2012d), the populations in Kuban et al and Moore et al were extremely preterm infants and Limperopoulos et al included children with functional disabilities in the study cohort. In contrast, the positive screen rate of 3.6% by the Bayley-III SE is similar to the reported M-CHAT positive screen rate in term-born children (Gray 2015).

The poor agreement between the Q-CHAT and the Bayley-III SE may reflect a difference in psychometric properties or symptom coverage. In the Bayley-III SE questionnaire, there was greater attention placed on sensory-integration and joint-attention whereas disordered communication and stereotypical behaviour were emphasized more in the Q-CHAT.

It must be emphasised that the predictive validity of these screening tools had not been investigated through diagnostic assessment for ASD. As discussed earlier, the co-existence of cognitive, language, motor, neurosensory and behavioural deficits pose a major methodological challenge for ASD-specific screening in the preterm population. Furthermore, there is little understanding of the differences in properties of the available screening tools. Oosterling et al compared four different instruments – the Early Screening of Autistic Traits Questionnaire (Dietz 2006, Swinkels 2006), the Social Communication Questionnaire (Rutter 2003), the Communication and Symbolic Behavior Scales-Developmental Profile, Infant-Toddler Checklist (Wetherby 2002) and key items of the Checklist for Autism in Toddlers (Baron-Cohen 2000) and found that no particular screening tool showed superior discriminating power for distinguishing children with ASD from those without (Oosterling 2009). Stephens et al (2012) and Dudova et al (2014) had also described a lack of overlap in the cases screened positive for ASD using different screening tools among the preterm population (see section 2.4.4). It is possible that each individual screening tool has low positive predictive value and that the simultaneous use of several screening tools may improve the identification of infants at highest risk of ASD to be targeted for early assessment and intervention. However, this approach

may prove time-consuming and burdensome. Larger longitudinal studies are needed to understand the association between early positive ASD screen and future behavioural or learning disabilities.

6.6.3 Strengths of the study

The methodological strengths of this study included prospective collection of both neonatal and two-year data which precluded recall bias and the stringent administration of the Bayley-III neurodevelopmental assessment. Information bias was further minimised as I was blinded to the medical history of the participants. In addition, the recruitment of participants, based on a gestational age rather than a birth weight criterion that was not limited to extreme prematurity, allows broader generalisability of the results.

6.6.4 Limitations of the study

I also recognise several limitations. The parents of eligible participants self-selected to attend the routine clinical follow-up appointment and agreed to participate in study 1 (to evaluate the validity of NHS assessments). The Q-CHAT response rate of 70%, although high for a questionnaire, may have introduced additional selection bias. I was also unable to assess children from non-English speaking families, thus limiting the broader applicability of my findings to other populations, particularly in those with higher proportions of non-English speakers. The lack of a contemporaneous control group meant that Q-CHAT scores could only be compared with the general population estimates. The proportion of preterm children in the unselected population from which the published estimates were based is unknown. In addition, there were higher proportions of children in the preterm study population living in more deprived areas. The preterm children were at a slightly older age when the Q-CHAT questionnaires were completed by their parents compared with the unselected general population. Nevertheless, no correlation between Q-CHAT score and age was found in both the general population and in this study cohort. Around one-quarter of the

parents had completed the Q-CHAT after their child had received the Bayley-III assessment. Although the Q-CHAT scores of these children did not differ from those whose parents had completed questionnaire before the Bayley-III assessment, knowledge of the results from the developmental assessment might have influenced parents' responses on the questionnaire and resulted in reporting bias.

6.7 CONCLUSIONS

Preterm children display greater early childhood social-communication difficulties and autistic behaviour than the general population as measured by their parents on the Q-CHAT. The higher frequency in autistic traits were observed mainly in the areas of restricted, repetitive, stereotyped behaviour, communication and sensory abnormalities. Further studies using diagnostic assessment is needed to evaluate the validity of early ASD screening tools and determine the true rates of ASD in very preterm children.

CHAPTER 7

STUDY 3 RESULTS: PREDICTIVE VALIDITY OF EARLY DEVELOPMENTAL ASSESSMENTS IN IDENTIFYING SCHOOL-AGE COGNITIVE DEFICITS IN CHILDREN BORN PRETERM OR VERY LOW BIRTH WEIGHT

7.1 RESULTS OF LITERATURE SEARCH

Figure 7.1 is the PRISMA flow diagram that depicts the flow of articles through the literature search and selection process. The electronic literature search yielded 3600 unique citations (1 duplicate). Application of search limits excluded 343 non-English articles and 413 articles published before 1990. The number of articles retrieved at each stage of the search process is detailed in Appendix 6. Two additional studies were identified through manual search and author correspondence. Sixty-eight studies were selected for full text evaluation from the title/abstract screen and 44 met the eligibility criteria. By matching 375 articles that reported the conduct of early developmental assessments with 323 articles that reported school-age assessments, 10 additional studies (in 23 articles) were identified. Data required for the review and meta-analysis were extractable directly from six articles. The authors of 18 of the remaining 48 studies contributed unpublished data for this review. The list of included studies is in Appendix 8. 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 all tables and figures.

7.2 DESCRIPTION OF INCLUDED STUDIES

The characteristics of the 24 studies included in the review are displayed in table 7.1. The studies were conducted in Europe (12 studies), the USA (7 studies), Australia (3 studies), New Zealand (1 study) and Israel (1 study). Sample sizes ranged from 11 to 313 participants. Most studies restricted the recruitment of participants to a single institution (15 studies); three studies were multi-centre and six studies adopted a geographical population-based sampling method. The inclusion criterion

was wholly based on birth weight in nine studies, gestational age in five studies and both birth weight and gestational age in five studies. For the other studies, additional inclusion criteria applied, including intra-uterine growth restriction (Bassan 2011), spastic diplegia (Fedrizzi 1993), specific neonatal diagnoses (McGrath 2000, Gray 1995) and low parental socioeconomic status (Smith 2006). The participants in six studies consisted of children born at >32 weeks gestation and with birth weight >1500g but the authors were able to provide relevant data limited to the subgroup that meet the criteria for this review. The different tools used by each study for early developmental and school-age cognitive assessments as well as the ages at which the tools were applied are also listed in table 7.1. As the studies spanned a period of more than 30 years, different editions of the same assessment tool were recorded.

7.2.1 Study populations

From these 24 studies, there were a total of 3133 children born at ≤ 32 weeks gestation and/or <1500g received both early and school-age assessments. 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. 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 48% of the final sample) excluded children who were unable to complete the assessments as a result of their physical disabilities (Bassan 2011, Bowen 1996, Bruggink 2010, Charkaluk 2011, Cohen 1995, Gray 2006, Gray 1995, Potharst 2012, Reuss 1996, Skranes 1998, Smith 2006, Tommiska 2003, Veelken 1991). The actual number of children excluded from the analysis for this reason is unknown as not all studies provided this information. In the studies by Class (2011), Fedrizzi (1993) and Vermeulen (2001), no child was unable to complete the assessment due to the presence of physical disability. In studies that included participants that were 'too physically disabled

to be tested' (Hack 2005, Kilbride 1990, McGrath 2000, Munck 2012, Orchinik 2011, Roberts 2010, Wolke 1999), these children were assigned a nominal score that was equivalent to being more than 2-4 SD below the population mean.

Figure 7.1 PRISMA flow diagram depicting the literature search process

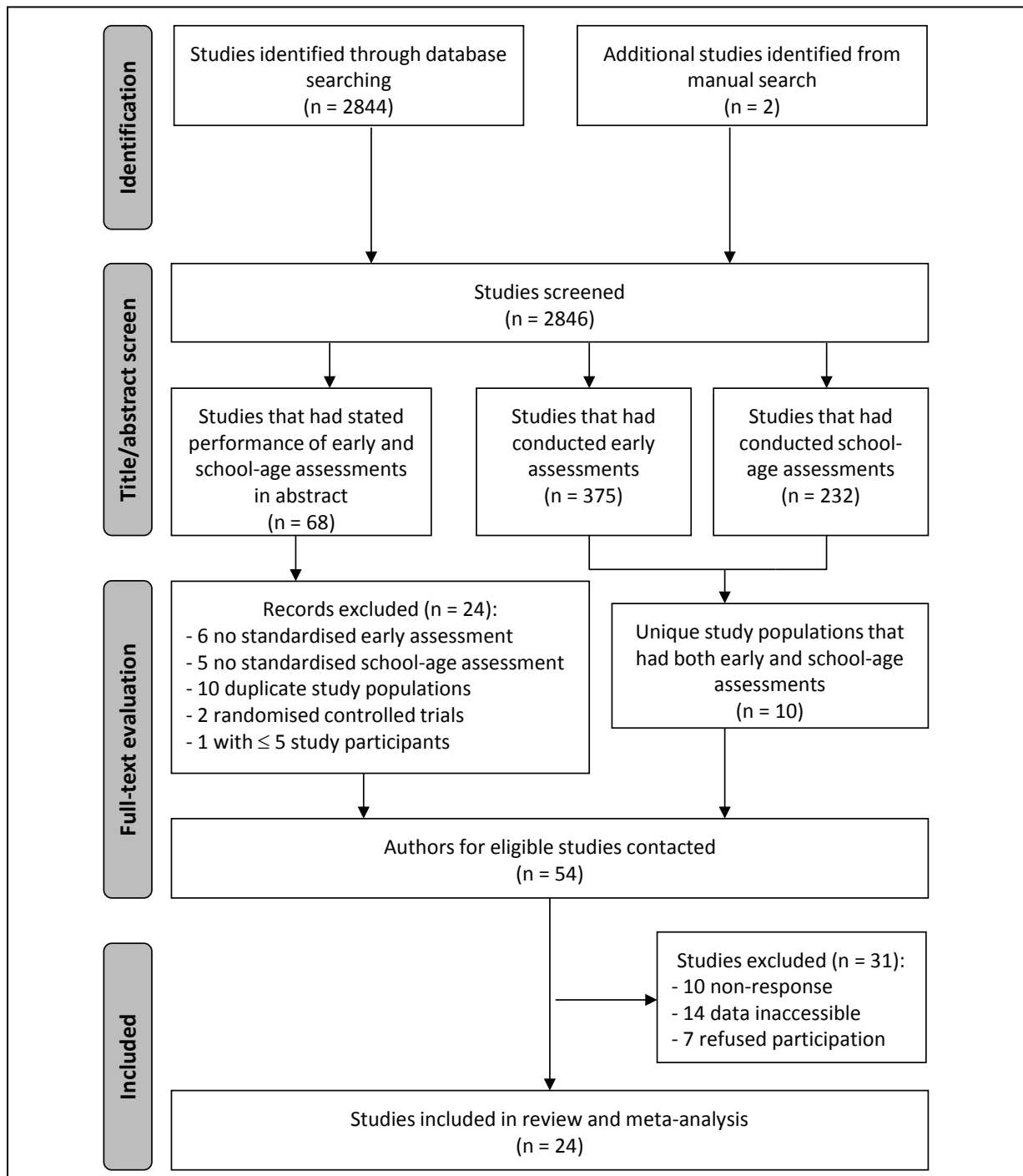


Table 7.1 Characteristics of studies included in review

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 <10 th 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)
Marlow, 2005	UK	GA <26 weeks	1995	PB	212	25.0 (0.7)	748 (116)	Early:	30m	BSID-II	81.7 (14.5)
								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.

7.2.2 Developmental and cognitive assessments

Ten studies reported the results of developmental assessments conducted between 12-24 months corrected age and 11 studies reported the results at 24 months corrected age. In three of these studies (Bowen 1996, Kilbride 1990, Potharst 2012), a repeat assessment was conducted at age 3 years. Marlow (2005) reported results at 30 months corrected age, Fedrizzi (1993) reported results at 3 years and Smith (2006) reported results at 3.5 years chronological age.

Results of school-age cognitive assessment was available at the ages of 5-6 years in 17 studies, 7-10 years in 11 studies and >10 years in 3 studies. Cohen (1995), Reuss (1996), Marlow (2005), Smith (2006) and Wolke (1999) had conducted multiple school-age assessments at different time points for their study populations.

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% (Bruggink 2010) to 67.0% (Hack 2005). The reported prevalence of school-age cognitive deficit was between 5.0% (Cohen 1995) and 67.4% (Marlow 2005) for mild-moderate (1-2 SD below mean) and 0.0% (Cohen 1995, Fedrizzi 1993) and 37.8% (Marlow 2005) for severe impairment (>2 SD below mean). In six studies (Gray 2006, Marlow 2005, McGrath 2000, Orchinik 2011, Roberts 2010, Smith 2006), the categorisation of outcomes was based on the mean and SD of the scores achieved by concurrently recruited term-born controls. Wolke (1999) 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. It should be noted that the study population in Smith (2006) was from middle- to low- socioeconomic groups and the mean test scores achieved by the control group was about 0.5 SD below the standardised mean. Using the results from the control group in this case could lead to an underestimation 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.

7.2.3 Quality of included studies: results of QUADAS-2 appraisal

Table 7.2 shows the details of the quality of each included study based on the QUADAS-2 appraisal tool and in figure 7.2, the proportions of studies that were considered ‘low’, ‘high’ and ‘unclear’ risk for bias and applicability in each domain were displayed. 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. Risk of information bias is low but may be introduced in three studies (Bowen 1996, Claas 2011, Fedrizzi 1993) due to the lack of blinding of assessors performing the school-age assessments to the results of the early developmental tests. It was unclear if blinding occurred in the studies by Roberts (2010), Reuss (1996), Smith (2006) and Tommiska (2003). Whilst the overall risk of bias was low, there is high concern for the applicability of the results from the studies to our current population in more than 50% of the studies. This is because many of the included studies were conducted more than 20 years ago; the characteristics of the study populations would be different and the assessment tools had been superseded by newer versions.

Figure 7.2 Proportions of studies with low, high or unclear risk of bias and concerns regarding applicability

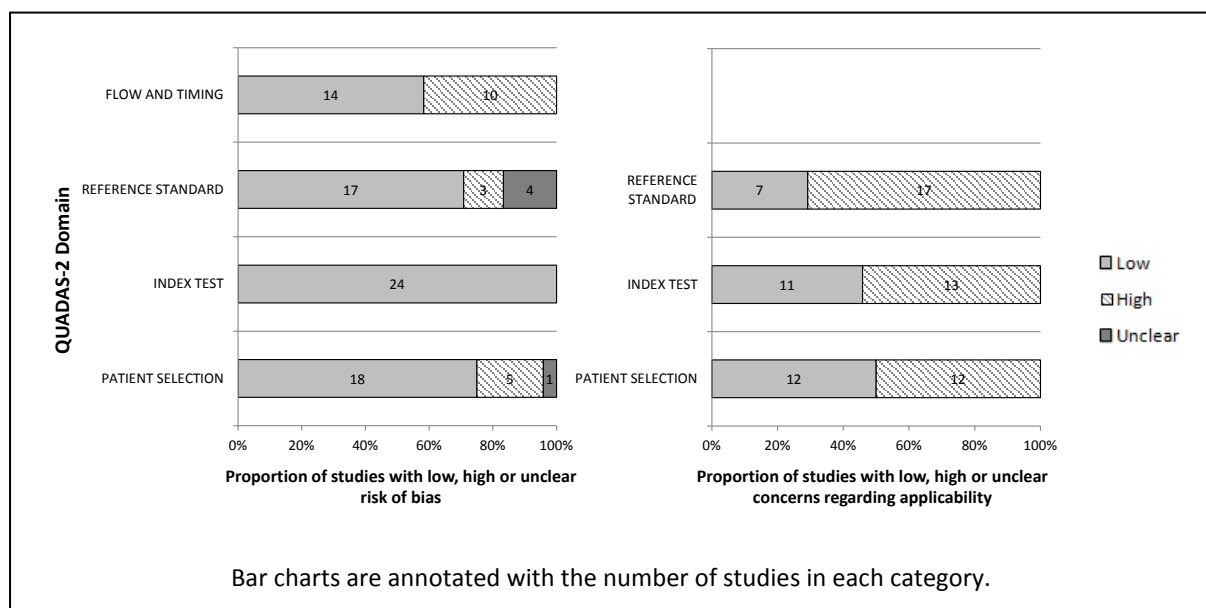


Table 7.2 Quality assessment of included studies using the QUADAS-2 appraisal tool

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 10 th percentile for gestational age. [4] Final cohort represents <30% of eligible population. [5] Study population restricted to children with birth weight below 10 th 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 <30% of eligible population.
Charkaluk 2011	↔	↔	↔	↑	↔	↑	↑	[4] Final cohort represents <30% of eligible population. [6] Non-universal assessment tool (Brunet-Lezine Revised, a French psychometric test) used. [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 <30% 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).
Kilbride 1990	↔	↔	↔	↔	↑	↑	↑	[6] Assessment tool developed before 1990 (BSID, 1969). [7] Assessment tool developed before 1990 (S-B-III, 1973).
Marlow 2005	↔	↔	↔	↔	↔	↔	↑	[7] Outdated assessment tool (KABC, 1983) used.
McGrath 2000	↑	↔	↔	↑	↑	↔	↔	[1] Inclusion criteria: meets <i>a priori</i> medical criterion (not specified). [4] Final cohort represents <30% 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 <30% of eligible population.
Reuss 1996	↔	↔	?	↑	↑	↑	↔	[3] Blinding of assessors not stated. [4] Final cohort represents <30% 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 <30% 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. [3] Blinding of assessors not stated. [4] Final cohort represents <30% 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 <30% 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) used. [7] Assessment tool developed before 1990 (WISC-R, 1974) used.
Wolke 1999	↔	↔	↔	↔	↑	↑	↑	[5] Study population was born before 1990. [6] Assessment tool developed before 1990 (GMDS, 1970). [7] Assessment tool developed before 1990 (KABC, 1983).

↔ = 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

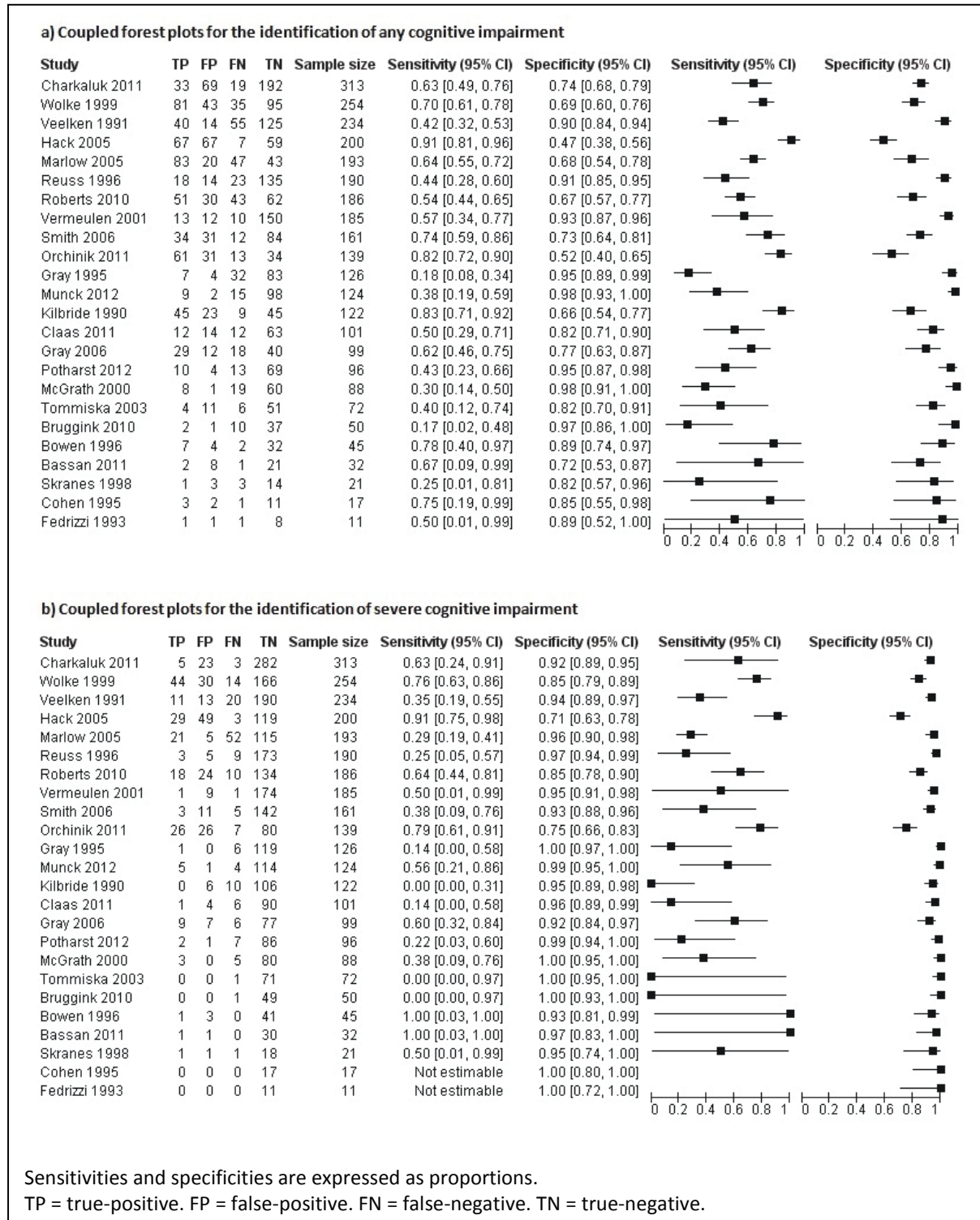
WISC-R = Wechsler Intelligence Scale for Children-Revised, 1974

WPPSI = Wechsler Pre-school and Primary Scale of Intelligence-Revised, 1989

7.3 PREDICTIVE VALIDITY OF EARLY DEVELOPMENTAL ASSESSMENT

In figure 7.3a, I present the results of the cross-tabulations and the estimated sensitivities and specificities of early assessments for identifying any cognitive deficit for each study, in the form of coupled forest plots ordered by the sample size of the study. Figure 7.3b shows the same information for the diagnosis of severe cognitive impairment. In studies where participants were examined at different time points, 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% (Bruggink) to 90.5% (Hack) and the corresponding estimated specificities ranged from 46.8% (Hack) to 98.4% (McGrath). For the diagnosis of severe impairment, the range of sensitivities was 0.0% (Bruggink 2010, Cohen 1995, Kilbride 1990) to 100.0% (Bassan 2011, Bowen 1996) and the range of specificities was 70.8% (Hack 2005) to 100.0% (Bruggink 2010, Cohen 1995, Fedrizzi 1993, Gray 1995, McGrath 2000, Tommiska 2003). The sensitivity of detecting severe impairment could not be estimated in the studies by Cohen (1995) and Fedrizzi (1993) as no participant had severe impairment. 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 more likely due estimates of sensitivity being based on smaller samples than estimates of specificity. The estimated sensitivity of 0.0% for severe impairment was based on a denominator of one (Bowen 1996, Bruggink 2010, Tommiska 2003) and 10 (Kilbride 1990) diagnosed cases at school-age assessments. In general, the larger the sample size, the more precise (smaller 95% CI) the sensitivity estimates. The precisions of specificity estimates appear to be high with the CI half-widths in 11 studies being less than 10.0% (Charkaluk 2011, Gray 1995, Hack 2005, Marlow 2005, McGrath 2000, Munck 2012, Potharst 2012, Reuss 1996, Smith 2006, Veelken, 1991, Wolke 1999).

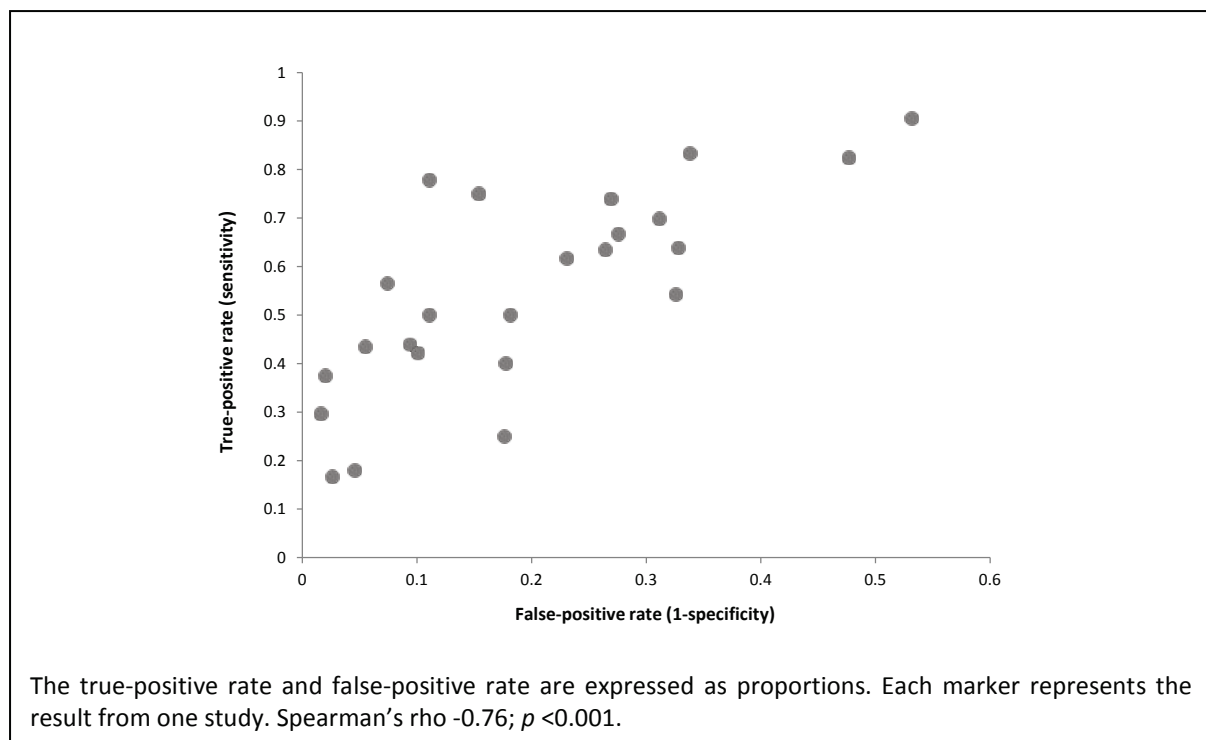
Figures 7.3 Results of cross-tabulations and coupled forest plots of the estimated sensitivities and specificities of early developmental assessments in identifying the presence of (a) any cognitive impairment and (b) severe cognitive impairment



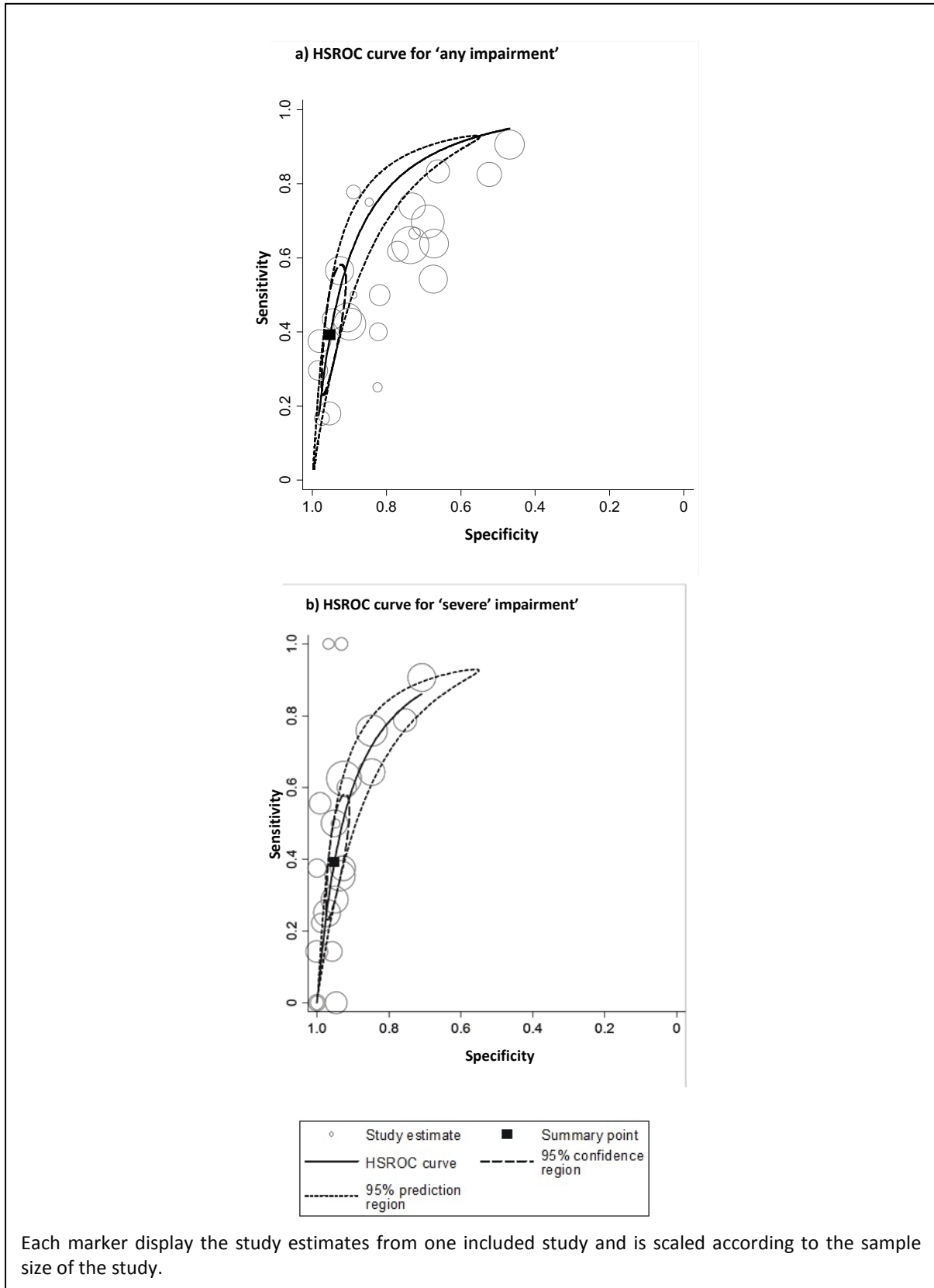
7.3.1 Meta-analytic pooled estimates of sensitivity and specificity

There was significant correlation between estimated sensitivities and specificities (figure 7.4; Spearman's rho -0.76; $p < 0.001$). Therefore, I did not compute the weighted averages of sensitivities and specificities separately. The pooled measures were estimated from the Rutter and Gatsonis HSROC curves that are presented in figure 7.5a for the presence of any impairment and figure 7.5b for severe impairments. The summary points and 95% confidence interval regions are mapped out in the figures as well as the 95% prediction regions which provide a forecast of the true sensitivity and specificity in a future study. 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%).

Figure 7.4 Scatterplot of the true-positive rate (sensitivity) against the false-positive rate (1-specificity)



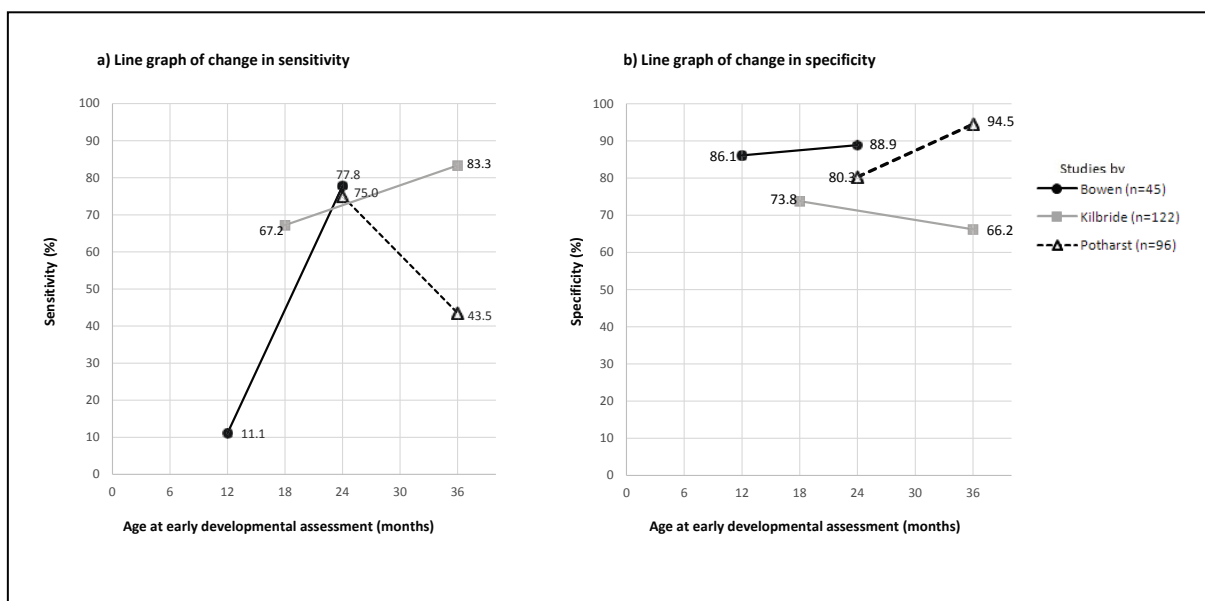
Figures 7.5a and 7.5b Hierarchical summary receiver operator characteristic (HSROC) curves for the pooled sensitivity and specificity of early developmental assessment in identifying (a) any impairment and (b) severe impairment



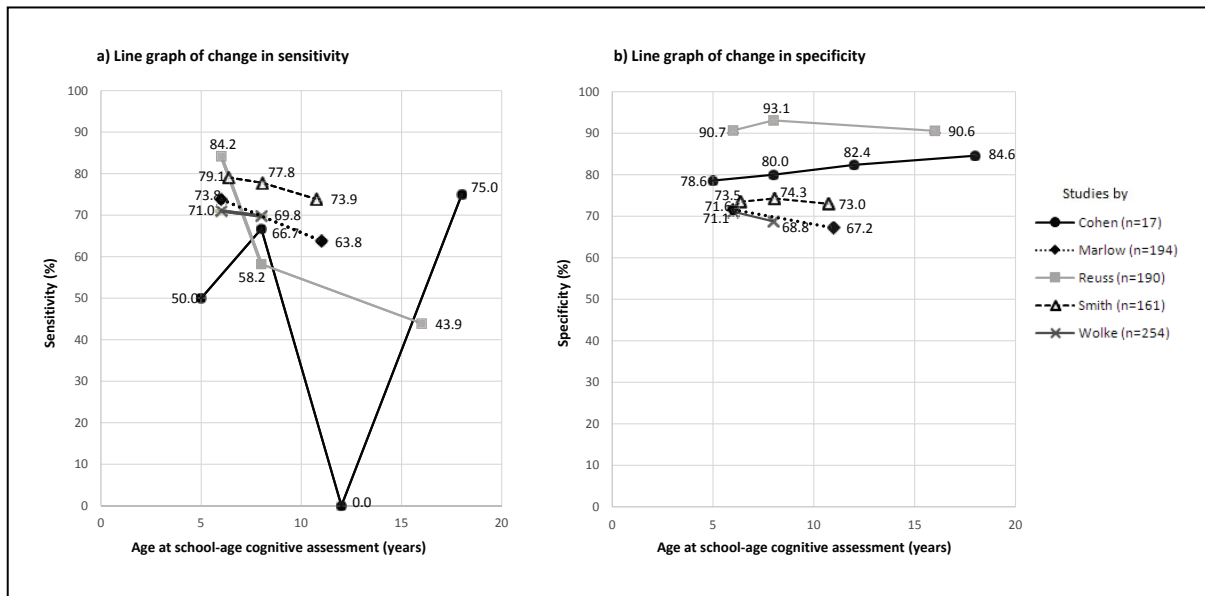
7.3.2 Validity of early assessment assessed at different time points

In three studies (Bowen 1996, Kilbride 1990, Potharst 2012), participants were assessed at two different time points for the early developmental assessments. In the five studies by Cohen (1995), Marlow (2005), Smith (2006), Wolke (1999) and Reuss (1996), participants received school-age cognitive assessments more than once. In figures 7.6a and 7.6b, I plotted the sensitivity and specificity for the identification of any impairment over the age at developmental assessment for the three studies that had examined early assessment at two different time points. Figures 7.7a and 7.7b are similar plots for the results obtained at serial school-age assessments in the five studies. It would appear, from these graphical displays, that the specificity of early assessment in excluding cognitive deficit remains relatively stable over time whereas no real correlation between sensitivity and age at assessment was apparent.

Figures 7.6a and 7.6b Line graphs demonstrating the change in (a) sensitivities and (b) specificities when early developmental assessments were repeated at different ages in three studies



Figures 7.7a and 7.7b Line graphs demonstrating the change in (a) sensitivities and (b) specificities when school-age cognitive assessments were repeated at different ages in four studies



7.3.3 Meta-regression: association of study-level variables with diagnostic validity

The odds ratios (OR) and 95% CI, together with the corresponding p -values, for the association of study-level variables with sensitivity and specificity of identifying cognitive deficit by early developmental assessment are presented in table 7.3. In Appendix 9, I display the scatterplots of the sensitivity and specificity from each study against the study-level variables under examination. There was reduction in specificity with increased observed prevalence of impairment in the study population (Figure S7b in Appendix 9). For each 1% increase in the prevalence of developmental impairment, the odds of identifying an additional case of ‘true-negative’ among those with no cognitive impairment reduced by 3% ($p=0.01$). The associations between mean gestational age and mean birth weight and specificity of identifying cognitive impairment reached borderline statistical significance (specificity increased with mean gestational age and mean birth weight of the study population). Post-hoc analysis revealed no association between the prevalence of impairment reported in each study and the mean gestational age ($p=0.55$) and mean birth weight ($p=0.95$) of the

study population, therefore excluding the speculation that the observed associations between specificity and mean gestational age and birth weight were mediated by the prevalence of impairment. Age at assessments, time interval between early and school-age assessments and the year of birth of the participants were not associated with sensitivity or specificity and therefore did not explain the heterogeneity present between studies.

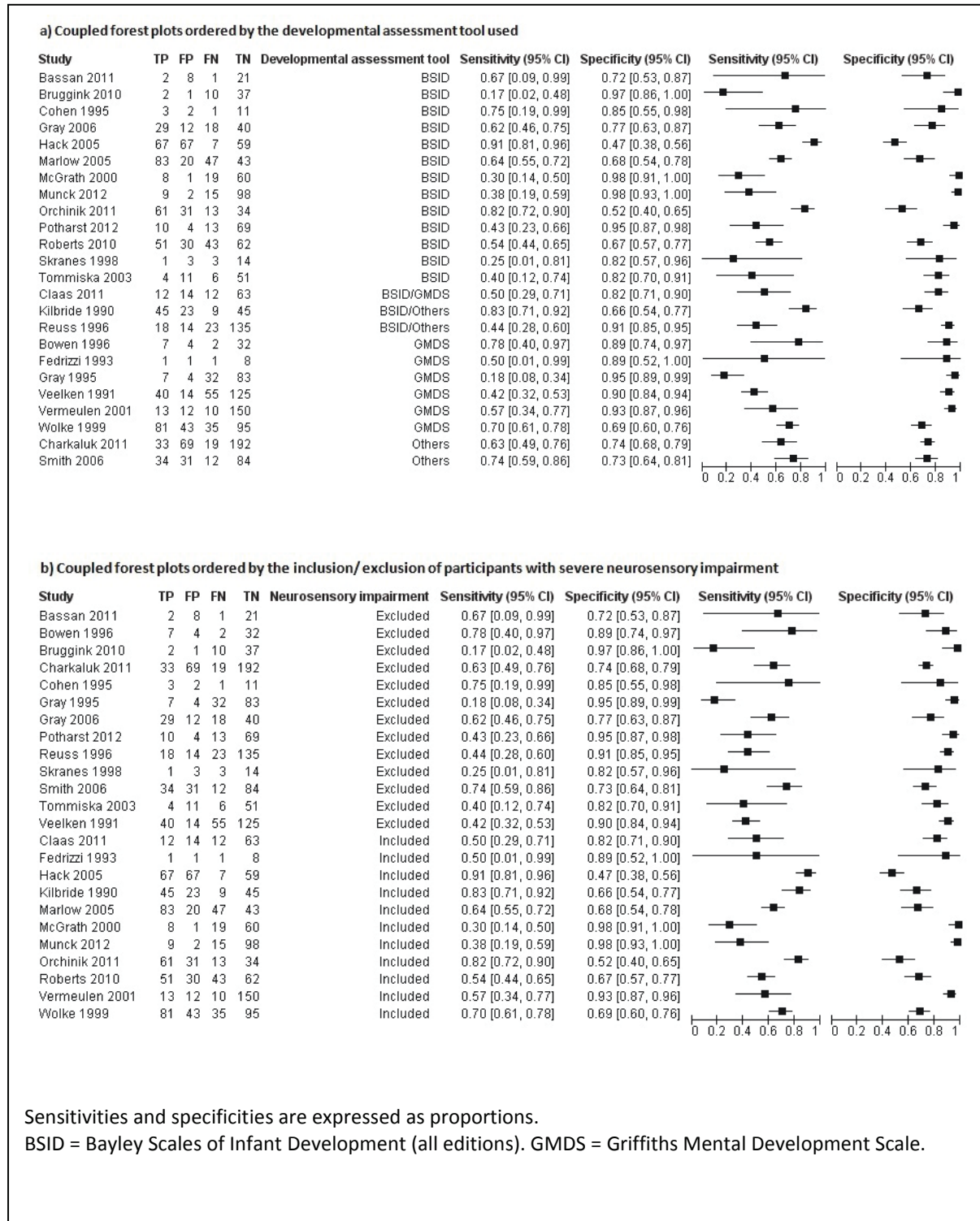
Figure 7.8a is a coupled forest plot of the sensitivity and specificity reported by the included studies, ordered by the early developmental assessment used. Studies were grouped into those that had used the Bayley Scales of Infant Developmental (all editions), Griffiths Mental Development Scale or other assessment tools. It is apparent visually from the forest plot that the type of assessment tool used did not change the variability of the reported sensitivities and specificities. As there were too many different types of school-age cognitive assessment tools used to be categorised into reasonably homogenous groups, the subgroup analysis on the association between cognitive assessment tools and sensitivity or specificity was not robust and therefore, not performed. The inclusion or exclusion of participants with severe neurosensory impairments also did not appear to affect the variability of the reported sensitivities or specificities of reported studies (Figure 7.8b).

Table 7.3 Association of study-level variables with estimated sensitivity and specificity

Study-level variable	Sensitivity		Specificity		<i>p</i> -value for joint test
	OR (95% CI)	<i>p</i> -value	OR (95% CI)	<i>p</i> -value	
Mean gestational age (per 1 week increase)	0.84 (0.68 - 1.04)	0.11	1.29 (0.98 - 1.61)	0.04	0.11
Mean birth weight (per 100g increase)	0.86 (0.72 - 1.03)	0.09	1.21 (1.00 - 1.48)	0.05	0.14
Mean age at early assessment (per 1 year increase)	1.51 (0.77 - 2.98)	0.22	0.79 (0.36 - 1.72)	0.54	0.35
Mean age at school-age assessment (per 1 year increase)	0.98 (0.86 - 1.11)	0.73	1.01 (0.88 - 1.17)	0.86	0.90
Mean time between assessments (per 1 year increase)	0.97 (0.86 - 1.10)	0.57	1.02 (0.89 - 1.17)	0.78	0.82
Year of birth (per 1 year increase)	0.99 (0.97 - 1.01)	0.291	0.99 (0.97 - 1.01)	0.23	0.82
Prevalence of impairment (per 1% increase)	1.02 (0.99 - 1.04)	0.16	0.97 (0.94 - 1.00)	0.01	0.02
Prevalence of severe impairment (per 1% increase)	1.05 (0.99 - 1.12)	0.12	0.93 (0.87 - 1.00)	0.02	0.03

OR =odds ratio; CI = confidence interval

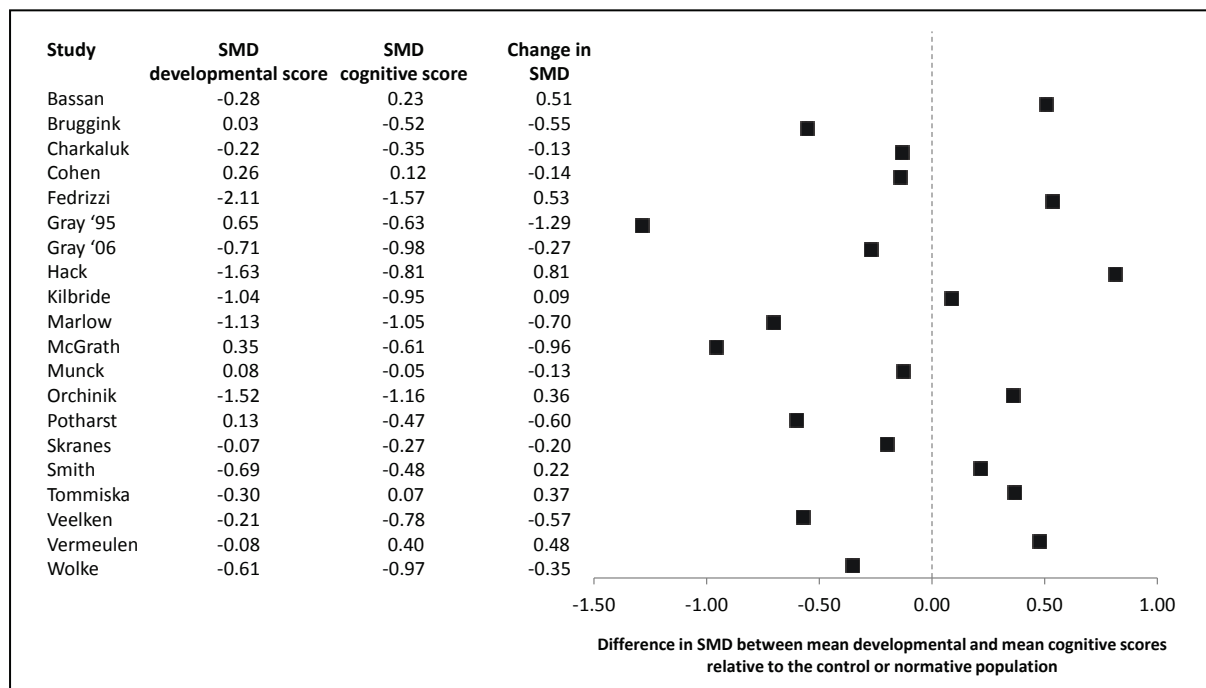
Figures 7.8a and 7.8b Coupled forest plots of the estimated sensitivities and specificities of included studies, ordered by (a) the developmental assessment tool used and (b) the inclusion/exclusion of participants with severe neurosensory impairment



7.3.4 Post-hoc analysis: changes in the standardised mean difference between mean developmental and cognitive scores

Figure 7.9 shows the SMD in the early developmental and school-age cognitive scores as well as the direction and magnitude in the change from early developmental score SMD to school-age cognitive score SMD over time for each study. The mean developmental and/or cognitive score for the study population was not available from the studies by Bowen (1996), Claas (2011), Reuss (1996) and Roberts (2010). The SMD was calculated using the mean and SD of the control population scores in the studies by Gray (2006), Marlow (2005), McGrath (2000), Orchinik (2011), Smith (2006) and Wolke (1999) and with the normative mean and SD appropriate for the assessment tool used in the other studies.

Figure 7.9 Change in the standardised mean difference in mean developmental and cognitive scores relative to normative or control populations over time

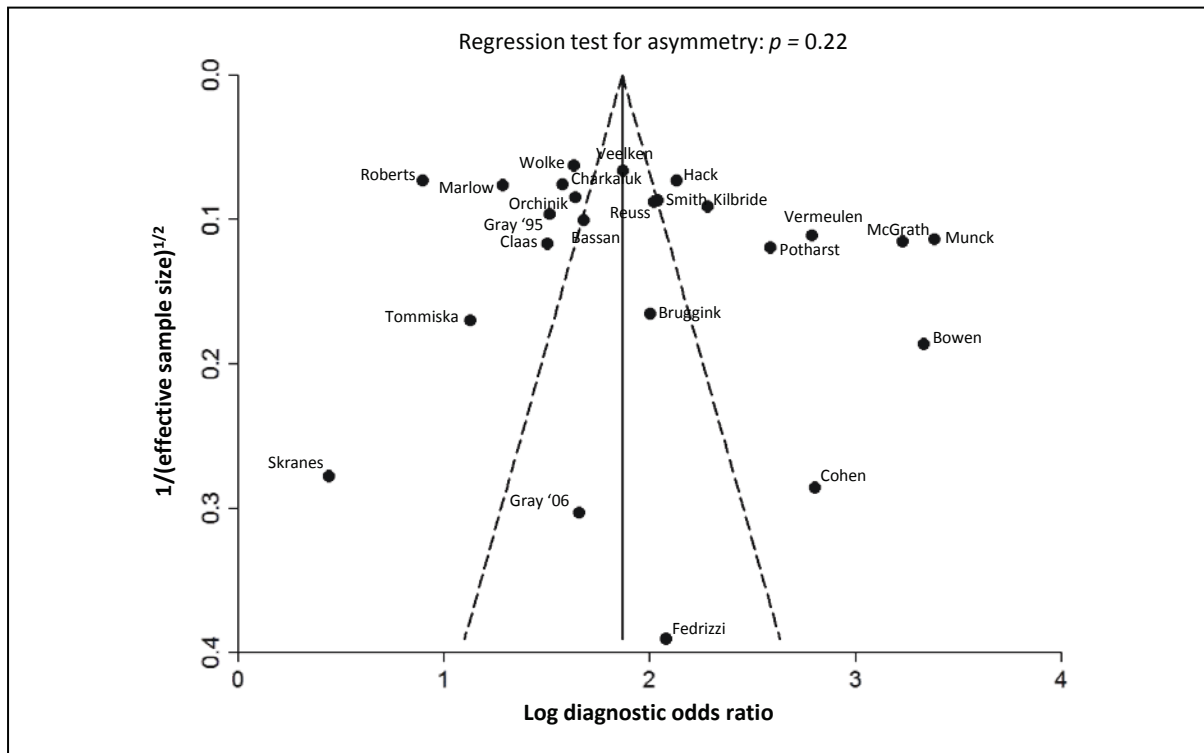


In the majority of studies, the mean developmental and cognitive score of the study populations were lower than the control/ normative mean (represented by a negative SMD). Overall, there is no clear direction in which the mean scores change relative to the control or normative populations over time. In 12 of the 20 studies in which the SMD was calculated, the mean cognitive score of the study populations deviated away (lower scores) from the control or normative mean by 0.1 to 1.3 SMD points compared with the mean early developmental score. In the other 8 studies, the mean cognitive scores of the study populations became 'closer' (higher scores) to the control or normative mean by 0.1 to 0.8 SMD points. The change in developmental and cognitive score SMD can be influenced by many factors such as the representativeness of the control populations and the assessment tools used. It is important to note that as this analysis was based on population-level scores, it would not be possible to extrapolate the findings and predict the change in scores for individual patient.

7.3.5 Funnel plot for sample size-related effects and publication bias

The funnel plot of the log DOR against the inverse of the square root of the effect sample size is presented in figure 7.10. Significance testing (effective sample size weighted regression test) confirmed that asymmetry was not present in the funnel plot ($p=0.22$), indicating the absence of sample size-related effects in the meta-analysis.

Figure 7.10 Funnel plot of the log diagnostic odds ratio against the inverse of the square root of the effective sample size, with pseudo 95% confidence limits



7.4 DISCUSSION

Through a systematic review of the literature, I found a substantial number of studies published in the past 20 years that had reported the early neurodevelopmental outcomes and later school-age cognitive abilities of children born preterm or VLBW. Whilst the reported specificities of early neurodevelopmental assessment in predicting later school-age cognitive outcomes were generally high, especially for severe cognitive impairment, the reported sensitivities were inconsistent. Meta-analysis of the data revealed that that early neurodevelopmental assessment had low sensitivity and high specificity for identifying school-age cognitive deficit. Such a test is considered useful for 'ruling in' the condition of interest if the test is positive (Akobeng 2006). This means that when a neurodevelopmental impairment was diagnosed at ages 1-3 years, the likelihood of having cognitive deficit at school-age was high (low false-positive rate; or 1-specificity). However, it would not be

possible to accurately exclude later cognitive deficit even when an early assessment demonstrated normal neurodevelopmental outcomes (high false-negative; or 1-specificity). My results suggest that almost half of the number of children thought to have normal neurodevelopmental function at ages 1-3 years would experience cognitive difficulties at school-age. Even for cases of severe cognitive deficit, the accuracy in early detection was low (meta-analytic sensitivity of 40.6%). This finding is not unexpected. Cognitive function in infancy has been shown to be poor predictors of later IQ in the general population (Aylward 2004). The poor predictability may reflect inherent 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 contents and psychometric properties of early neurodevelopmental and later cognitive assessment tools.

7.4.1 Internal validity

The internal validity of this study is influenced by the quality of the data from the included studies as well as the methods I had adopted. Data quality as appraised by the QUADAS-2 tool was good with most studies considered to be at low risk of bias. Nevertheless, the presence of missing data from participants lost to follow-up over time is a common problem affecting these longitudinal studies. Such incomplete outcome ascertainment can distort the result in either direction and it is difficult to speculate its effect on the meta-analysis. For example, if the children who did not attend were those with severe cognitive deficit that was apparent from early childhood (the 'true-positives'), the resulting sensitivity could be underestimated. On the other hand, if there were disproportionately higher attrition from children diagnosed with normal early neurodevelopment but had school-age cognitive deficits (the 'false-negatives'), then the estimated sensitivity would be exaggerated in the study.

Another source of missing data arose from the exclusion of children with severe neurosensory and motor impairment who were unable to complete the neurodevelopmental or cognitive assessments due to their physical disabilities. If one assumes that these children had stable diagnoses of severe neurodevelopmental and cognitive deficits throughout childhood, then the impact of excluding them from the study population would be an underestimation of the sensitivity of early neurodevelopmental assessments.

There are additional biases which could affect the accuracy of the data that were not identified through the QUADAS-2 appraisal. One such factor is the experience and qualification of the assessors ascertaining the neurodevelopmental or cognitive outcome. Although all the included studies had employed trained assessors who utilised standardised assessment tools, inter-observer differences were inevitable and could lead to variations in the results. This is an important consideration. Neurodevelopmental and cognitive abilities exist as a continuum but for the purpose of the study, participants were dichotomised using a 'cut-off' score into groups 'with impairment' and 'without impairment'. Inter-observer variations around the 'cut-off' score would result in misclassification of outcomes. The effect of differential misclassification on the study results is difficult to predict, and will depend on whether the misclassification was more likely to occur during the early or school-age assessment and whether the misclassification is biased towards or away from impairment. In general, we can expect any misclassification to have a bigger impact on the estimated sensitivity, which is calculated using a small number of 'positives' in this condition of relative low prevalence, than specificity, which is based on a large number of 'negatives'.

Another factor not taken into account in the QUADAS-2 appraisal was the sample size in each included study. It is apparent from figures 7.3a and 7.3b that the confidence intervals for the estimated sensitivities and specificities widened with decreasing sample sizes. Five of the included

studies had less than 50 participants and in four studies, the sensitivities were estimated based on less than 5 cases and hence, subjected to significant sampling error.

Participants were included in the review if they fulfilled either the gestational age or the birth weight inclusion criteria. The birth weight criterion was used in order to capture all relevant studies since it was common for neonatal studies to base the eligibility criteria on birth weight rather than gestational age. However, the methodological bias in using a birth weight criterion is the inclusion of more mature but growth-restricted children. Notably, in the study by Bassan et al (2011), all the participants were small for gestational age (birth weight <10th percentile for gestational age). Intra-uterine growth-restriction is a recognised risk factor for poor neurodevelopmental outcome (Gutbrod 2000). The consequence of this, as well as additional criteria applied by other studies, would be an alteration in the patient spectrum from the target population, accompanied by a bias in the results obtained.

7.4.2 External validity (generalisability)

The QUADAS-2 appraisal highlighted the lack of applicability of older study populations and assessment tools in more than half of the included studies to the current practice and hence raised the question on the wider generalisability of the study findings. This is, of course, a reflection of the nature of all longitudinal studies but it is a significant limitation, particularly in the context of a rapidly advancing neonatal specialty. 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 (Salt 2006, Wilson-Costello 2007, Doyle 2011) so we can expect that the characteristics of the current preterm population to be different those from past eras. Although I had attempted to focus the review on studies published since 1990, only twelve of the 22 included studies recruited participants born after 1990 and none were born in the last 10 years (i.e. after 2004).

More importantly, the assessment tools used in the included studies, although validated and contemporary at the time of each study, had mostly been superseded by newer editions. For example, the Bayley Scale of Infant Development is now in its third edition (Bayley 2006b). Recent studies have suggested that children achieved higher scores on the third edition of the Bayley Scales compared with the second edition when concurrently tested with both versions (Moore 2012c, Lowe 2012). Therefore, caution should be exercised when extrapolating the results based on earlier versions of the assessment tools to the current practice. I had chosen not to restrict the review to only studies using a particular assessment tool to increase the generalisability of the findings. 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.

The timing and setting of the assessments also played a part in determining the external validity of the study findings. The early neurodevelopmental assessments were performed between 12 and 36 months and the timing match common clinical practice. School-age assessments were mostly conducted between the ages 5 and 8 years, when children were at the primary stages of schooling. Only two studies reported cognitive assessment during adolescence, one of which had only 20 participants. Therefore, the validity of early assessment in diagnosing cognitive deficit extending into adulthood could not be estimated from this study, although one could speculate that the sensitivity might be even poorer. Furthermore, all the included studies were conducted in developed countries, restricting the generalisability of the findings to developing regions.

7.4.3 Strengths and limitations of review and meta-analysis

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, in addition to the issues mentioned before that would affect the internal and external validity of this study, other potential limitations exist, some of which are common to reviews and meta-analyses. The review was restricted to English-language literature due to the lack of translation support. There is concern that the English-language journals publish a skewed sample of studies that report positive and more noteworthy results (Morrison 2012). Similarly, it is common for articles with negative or inconclusive findings to remain unpublished. The exclusion of grey literature including abstracts and dissertations could have led to the omission of essential and more recent information. Despite demonstrating funnel plot symmetry which would suggest a lack of sample size-related publication bias, I could not completely rule out the effect of reporting bias from this review. Furthermore, the accuracy of the aggregated data provided by study investigators could not be verified.

I 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 (Bossuyt 2013). This was compounded by the narrow range of values available for each of the explanatory variables under evaluation. For example, the time difference between early and school-age assessment fell between the narrow range of 3 to 8 years in 19 of the 24 studies. Hence, a type II error could not be excluded. More importantly, meta-regression is subjected to ecological fallacy (or aggregation bias). This is the mistaken assumption that statistical between-study relationship based on aggregated data reflects within-study

relationship. Therefore, in order to reliably identify factors that influence the validity of early developmental assessments, it would be necessary to obtain individual patient-level data.

7.4.4 Comparison with other studies

The few studies that had specifically examined the stability of the early neurodevelopmental diagnosis in preterm infants had been included in this review. In 2013, a similar meta-analysis on the predictive value of the BSID on the development of very preterm and/or VLBW children was published by Luttikhuisen dos Santos et al from The Netherlands (Luttikhuisen dos Santos 2013). They reported a strong positive correlation between BSID Mental Developmental Index in the first three years of life 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 the Dutch meta-analysis and my study that could explain the different conclusions drawn. Firstly, only studies using the BSID were included in the Dutch meta-analysis and it included studies published before 1990. Secondly, 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 my 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.

7.5 CONCLUSIONS

Early neurodevelopmental assessment has high specificity but low sensitivity in identifying later school-age cognitive deficit. The inclusion of a large number of studies conducted on older populations and using outdated versions of assessment tools had reduced the generalisability of my findings. However, I am concerned that a significant number of older children and adolescents born very preterm or VLBW are experiencing difficulties in school, and that they could have potentially benefitted from earlier support and intervention if their cognitive deficits had been recognised. I would encourage future studies to research on the potential factors affecting the sensitivity and specificity of early neurodevelopmental assessments, in order to improve the follow-up care of these patients.

CHAPTER 8

GENERAL DISCUSSION AND CONCLUSIONS

8.1 EVALUATION OF EVIDENCE TO ORIGINAL HYPOTHESES

The main hypotheses for the three studies detailed in this thesis were set out in section 3.2. I would now like to consider the soundness of the hypotheses in light of my findings.

Study 1:

In study 1, the hypothesis was tested through calculations of the sensitivity and specificity of routinely collected NHS data in defining children with neurodevelopmental impairments using a standardised questionnaire, against a 'gold-standard' research assessment using the Bayley-III scales. Conventionally, the desired sensitivity of a developmental test is between 70% and 80%. Overall, I found that the sensitivity of routine NHS data in classifying children with any impairment was 61.5% (95% CI 52.5% - 70.6%) and the corresponding specificity was 85.7% (95% CI 77.4% - 94.0%). For the presence of severe impairment, the sensitivity was 52.0% (95% CI 23.8% - 80.2%) and specificity was 96.7 (95% CI 92.5% - 99.9%). Therefore, my results do not support the original hypothesis that the neurodevelopmental data determined through routine NHS assessment is of sufficient accuracy in identifying children with neurodevelopmental impairments.

Study 2:

In study 2, children in my study population (born at less than 30 weeks gestation and assessed at 24 months corrected age) received significantly higher Q-CHAT scores than those reported in the general population. The difference in mean scores between the preterm and the general population was 7 points, a magnitude of nearly one standard deviation. This supports the hypothesis that children born very preterm display greater social-communication difficulties than the general population at age two years.

Study 3:

In study 3, I used meta-analytical methods to combine all available and relevant data collected by studies published in the past 2 decades, identified through a comprehensive systematic review of the literature, in order to determine the predictive validity of early developmental assessment in identifying cognitive impairment at school age. The pooled sensitivity (55.0% (95% CI 45.7 - 63.9%)) and pooled specificity (84.1% (77.5 – 89.1%)) of early developmental assessment for the identification of any cognitive impairment were low. The results, therefore, would support my original hypothesis that early developmental assessments are poor at predicting the presence of cognitive impairment in the very preterm population. However, I recognise that my findings, obtained through aggregated data, are subjected to ecological fallacy and cannot be extrapolated to individual patients.

8.2 CLINICAL RELEVANCE OF RESULTS

8.2.1 Reliability of neurodevelopmental assessments and outcome data recorded during routine clinical follow-up

The results of study 1 highlight the low sensitivity of routine NHS assessments for identifying mild to moderate neurodevelopmental impairment. This has significant clinical implications. At an individual level, it raises the concern that children with impairment may be missed. At a population level, current documentation of two-year outcome data during routine NHS assessments, using the standardised electronic form in its present format, would underestimate the proportion of children with impairment, when compared with a research-standard assessment using the Bayley-III. Due to the time and resources required to assess and prospectively collect neurodevelopmental outcome data, there is a paucity of high-quality and up-to-date published evidence of outcomes from preterm birth. Many neonatal networks and individual units rely on local follow-up services to report on the impairment rates of their 'NICU graduates' and the NPEU/Oxford classification is widely used for

categorising the levels of functional outcomes. The accuracy of the 'local impairment rates' needs to be interpreted with caution in light of my findings.

8.2.2 Social-communication difficulties experienced by children born preterm

My findings of higher Q-CHAT scores in the preterm population in study 2 suggest that suboptimal development of social-communication skills in this population exists from early childhood. Since ASD exists as a continuum with autism representing the extreme end of the spectrum, the results also support the likelihood that a large proportion of preterm children experience clinically significant social-communication difficulties below the diagnostic threshold for ASD from a young age, when early intervention may be possible. My findings draw attention to the need for better understanding and potentially early assessment of social-communication skills in the preterm population.

8.2.3 Predicting school-age cognitive impairment

The results from study 3 confirmed that a significant proportion of children born very preterm who were assessed to have 'normal' neurodevelopmental outcome in early childhood go on to experience cognitive difficulties later in school. The implications of this finding on current clinical practice are considerable. 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, I found the false-positive rate for early diagnosis of impairment to be low. It is likely that children with more severe impairments, who would receive greater benefit from early intervention, 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 early interventions. Additionally, outcome data used during discussions with parents during the antenatal and neonatal periods are commonly based on neurodevelopmental outcomes

determined during early childhood. Given my findings, it is essential to discuss potential school difficulties children may face, even in the absence of obvious impairment or disability.

8.3 IMPLICATIONS AND RECOMMENDATIONS FOR FUTURE PRACTICE

8.3.1 Individual versus population outcome data

In the background chapter, I described that neonatal neurodevelopmental outcome data are collected for multiple purposes (section 2.1) and that an establishment of a coordinated neonatal follow-up programme in which eligible infants are assessed and outcome data are collected to serve these multiple purposes would be hugely advantageous (section 2.6.2). Through my research, it has become apparent that some fundamental differences in the principal requirements of individual patient-level and population-level outcome measures may have contributed to the challenges in creating a merged data source.

For an individual patient, current clinical follow-up focussed on identifying children who are experiencing difficulties that may require early interventions. For this 'screening' purpose, the use of a detailed standardised developmental assessment, which requires trained staff and is expensive and time-consuming, can be considered unnecessary. Furthermore, the derivative of a standardised assessment score may not be deemed particularly useful by parents, as they can be difficult to interpret and the individual trajectory from a single score is unpredictable. The classification of outcomes into categories of severity, particularly for children functioning near the margins of the 'cut-off' point is unhelpful. Hence, many services continue to assess children using clinical judgment alone or screening assessments that purely classify children into 'high' or 'low' risk.

The main purposes of collecting population-level data were to facilitate analyses of the trends and distributions of impairment prevalence, to aid the study of the determinants of various outcomes

and to assist in health provision planning. Therefore, the key attributes for a population-level neurodevelopmental outcome measure are that it should be (i) relevant and easily measurable, (ii) valid and reliable, (iii) measurable over time, and be (iv) sensitive to changes in factors that influence it (e.g. new interventions in neonatal practices). In the past few decades in developed countries, there had been significant reduction in the rates of neonatal mortality and major disabilities that these outcome measures cannot be considered sufficiently sensitive measures. In this respect, a standardised assessment tool that provides a numerical 'score' would be ideal. A score would provide characterisation of the population with a mean and standard deviation, and allow comparison between populations and monitoring of trends over time. The ideal assessment tool should encompass assessment of all neurodevelopmental domains of interest while at the same time be easy to administer in order to maximise acceptability by parents and health professionals.

The ideal assessment for the collection of patient-level and population-level outcome data appear in the first instance to be in conflict. In reality, any measure used to monitor population outcome, as long as it is deemed acceptable to the parents, can be used to assess an individual child's development and if necessary, as a basis for referral for additional support and to plan interventions. Although the process of developing a national follow-up programme with standardised collection of outcome data that meets the requisite for both patient-level and population-level requirements could be complex and challenging, I still believe that this endeavour will prove hugely beneficial for the neonatal community and should be pursued. In the next section, I provide recommendations for the development of the follow-up programme and data collection.

8.3.2 Considerations and recommendations for neonatal follow-up programmes and data recording

One of the advantages of the current practice for following-up preterm children is that the children are mostly assessed by health professionals who have already been involved in their neonatal care and with whom the parents have already built up rapport and relationship with. At an organisational level, the processes required to run the follow-up services can be made as accessible and as cost-efficient as possible, based on the set-up of the existing local paediatric services. However, this leads to regional variation in follow-up rates, reliability of assessments and quality of outcome data recording. The lack of consistency in the quality of service delivery throughout the country is a major flaw.

The Healthy Child Programme is a universal child health prevention and promotion programme in the UK (Shribman 2009). Under this programme, all children receive health visitor led developmental screening. There has been some interest in extending the roles of health visitors to capture developmental outcome data of children born preterm, assessed using developmental screening tools or through questions similar to those listed on the electronic '2-year outcome forms' (NPEU/Oxford criteria) (Amess 2010). Based on the findings of study 1, I would be cautious of this approach. I have shown that, even with the use of developmental screening tools, the false-negative rates (sensitivities) are unacceptably high. Other factors such as shortage of health visitors, requirement for further training and lack of universal uptake of the screening programme will further limit the successful implementation of this programme.

Choice of assessment tool and outcome measure

I have highlighted in chapter 5 (study 1) the issue of inter-rater variability on population data collection. Standardising the choice of neurodevelopmental tool used during follow-up assessment would be an obvious way of minimising this variability. Such a tool should have strong psychometric

properties, be user-friendly and ideally be adaptable for use in non-English-speaking patients. Misclassification occurs during categorisation of outcomes, hence the strategy of presenting outcome data in categories should also be further considered. Categorical outcomes are easy to interpret and communicated and mirror clinical practice e.g. referral of children below a certain threshold for further assessment or intervention. However, for an individual child, the labelling of 'outcome category' is unhelpful. Besides, as I have shown in study 3, the categories of outcomes do not remain stable over time. In this respect, a tool that allows the presentation of the distribution of standardised scores would be more valuable. In addition, the cognitive abilities of preschool children can be difficult to test and the recommended assessment tool must provide reliable assessment of cognitive skills. To monitor population outcomes effectively, high uptake of the assessment tool is required. The Bayley-III assessment is considered by many services to be too time-consuming and expensive to adopt into routine follow-up practice. Besides, the ongoing concern that it underestimates developmental delay is worrying. The choice of the 'right' recommended tool that is not only highly informative, but also acceptable to health professionals and parents, could be pivotal to the success of any follow-up programme. On this basis, I would recommend assessing the utility of other developmental tools, particularly parent-completed assessments such as the Parent Report of Children's Abilities (PARCA-R) (Johnson 2004, Johnson 2008) and the Ages and Stages Questionnaire (Skellern 2001) as population outcome measures. They can be used as a time-efficient and cheap method for collecting outcome data and as part of a comprehensive follow-up assessment in combination with other strategies such as a formal neurological assessment conducted by health professionals. The PARCA-R has already been shown to have strong psychometric properties when validated against the BSID-II (Johnson 2008) and the Bayley-III (Martin 2013) for use with very preterm children and had been employed for outcome evaluation in neonatal studies (Marlow 2006, Brocklehurst 2011). The ASQ was singled out in the review by the Policy Research Unit in the Health of Children, Young People and Families at the University College London Institute of Child Health as one that "best satisfy the requirements for a population measure of children's development"

(Bedford 2013) and since then, has been adopted as part of the health visitor led two year review in some parts of England. These parent-led tools have the advantage of engaging the parents in the evaluation, hence improving the acceptability of the follow-up programme. In addition, a parent-completed measure may provide a better overall assessment than a one-off test. However, the extension of their widespread use in a neonatal neurodevelopmental follow-up programme will require further evaluation (see section 8.4.2).

Organisation of follow-up programme

Based on my findings, I would suggest a centralised approach to the assessment and collection of 2-year outcome data for children born very preterm. Typically, very preterm children are offered post-discharge appointments every three to six months. I think that these visits should continue to occur at the local hospitals where allied health professional support e.g. dietetics, physiotherapy can be sought if necessary. However, I believe that there are advantages for the two-year neurodevelopmental assessment to be organised at the neonatal network level. This approach will ensure that each child receive an assessment by a dedicated team of health professionals who had been highly trained, thereby improving the reliability of the outcome information recorded and minimising inter-rater variability. In addition, the follow-up programme can include a dedicated coordinator and other administrative support aimed at tracing and contacting families to maximise follow-up rates. The results from the neurodevelopmental assessment could then be fed back to the local clinicians. There must be adequate funding to allow implementation of such a programme and this should be considered in line with the commissioning of neonatal critical care services.

8.3.3 Improvement in completeness of population outcome data

In addition to being of high quality, data recorded during clinical care should be complete to enable meaningful analysis. Currently, the utility of the routinely recorded electronic clinical data as a source of population-based outcome information is limited by poor data completeness. Based on

the NNAP, two-year outcome data was available from only 44% of all infants born at less than 30 weeks gestation in England and Wales between July 2010 and June 2011. It is unclear if the financial incentives introduced by the Commissioning for Quality and Innovation (CQUIN) payment framework to encourage post-discharge follow-up and data collection had been successful at improving data completeness. Strategies to reduce missing data must be aimed at clinician engagement and minimising loss to follow-up. Therefore, the barriers to uptake of follow-up programmes need to be understood (see section 8.4.1). There should be active gathering of information on children who did not attend appointments, either through the use of parent-completed questionnaires, telephone interviews or by health visitors. Ultimately, promoting a change in practice such that all health professionals conducting the neurodevelopmental assessments take responsibility for data entry may be necessary.

8.3.4 Intervention for social-communication difficulties

Since 2007, the American Academy of Pediatrics (AAP) has recommended ASD-specific screening for all children at 18 months to facilitate early diagnosis and to prevent delay in the initiation of early intervention (American Academy of Pediatrics 2006). This is in addition to a systematic developmental surveillance process that should occur during every visit to the primary care paediatrician during childhood. The purpose of the surveillance for ASD was to identify risk factors such as family history as well as parental concern that would prompt referral for early comprehensive ASD evaluation. The aim of screening is to use standardised tools to support and refine the risk for ASD (American Academy of Pediatrics 2006). These recommendations were made based on concerns that physician estimates of the developmental status of children are less accurate when only clinical impressions, rather than formal screening tools, are used (Sand 2005). Selected ASD-specific screening tools, classified into 'level 1' or 'level 2' tools were endorsed (Johnson 2007). The M-CHAT and the PDDST-II (Seigel 2004) are considered level 1 tools that are designed to differentiate children who are at risk of ASD from the general population. Level 2 tools, such as the

Childhood Autism Rating Scale (Schopler 2010) and the Social Communication Questionnaire (Rutter 2003) are more useful in differentiating children at risk of ASD from those at risk of other developmental disorders. These screening tools are age-specific and the major limitation of the recommended level 2 screening tools for early screening is that the suggested minimum age for testing starts from 3 years (Johnson 2007).

The UK National Screening Committee does not currently recommend universal screening on the basis that none of the available screening tools have sufficient reliability in identifying children at risk for ASD when applied to the general population (UK National Screening Committee 2009). As I discussed in sections 2.4.4 and 6.6.2, ASD screening among the preterm population is challenging due to co-existing cognitive, language, motor and neurosensory deficits. Although the putative behaviours identified on the Q-CHAT may represent precursors of a later ASD diagnosis, they may also reflect developmental delay or idiosyncratic manifestations of prematurity-related complications. The validity of the 'level 2' screening tools recommended by the American Academy of Pediatrics had not been formally tested in the preterm population. Regardless of an ASD diagnosis, I have illustrated that these preterm-born toddlers experience problems in current functioning that may interfere with adaptive exploration and social engagement. It is pertinent that as clinicians, we recognise these difficulties and the impact they have on families. Parents should be given information on the social-communication difficulties that preterm children experience, particularly as some of these behaviours may be amenable to specific interventions such as speech and language, occupation and sensory integration therapies as well as educational programmes targeted at enhancing communication, social skills instruction and reducing interfering maladaptive behaviours (Myers 2007).

8.4 AREAS FOR FUTURE RESEARCH

There is much to be learnt so that we can continue to improve and refine neonatal follow-up assessments and outcome data recording. In this section, I propose some considerations for future research.

8.4.1 Investigate barriers to uptake of follow-up programme

As I discussed in section 8.3.3, currently the NNAP was only able to collate outcome data from less than 50% of eligible infants born less than 30 weeks gestation in England and Wales. In my experience during the recruitment of children to study 1, a significant proportion of children are lost to follow-up before the age of 2 years although I am unable to provide exact quantification. In the Swiss national cohort study, the attrition rate at 2 years was still 19% despite a robust tracking process. It is obvious that achieving adequate follow-up rates, particularly in populations that are highly mobile, with diverse cultures, primary spoken languages and socioeconomic status would be the major challenge in the collection of outcome data through a standardised follow-up programme. It would be worthwhile to focus effort on understanding the barriers to participation in neonatal follow-up assessment. This could be achieved, in the first instance, by setting up a focus group to identify factors that influence non-attendance at follow-up appointments. The factors that need to be considered include the quality of communication to parents, their understanding of the importance of follow-up assessments and level of engagement, the continuity in service provision and tracking system following discharge from neonatal inpatient care, difficulties in accessing care (e.g. transportation problems) as well as language and literacy barriers. A Working Group of experts can be convened to review these factors and provide recommendations for improvement to current practices.

8.4.2 Improve validity of data collection form

In section 8.3.2, I recommended that parent-completed assessment questionnaires, specifically the PARCA-R and the ASQ, could be used as part of neonatal follow-up programmes to collect population-level outcome data. The utility of these tools for this purpose needs to be investigated. For example, whilst the PARCA-R had been validated for use in the preterm population, up-to-date norm scores for the general population or control term groups is unknown. This information will be helpful as a benchmark from which scores from different populations can be compared against. Similarly, there is no standardised ASQ norm scores for the UK population and so far, only a couple of small studies had reported the validity of the ASQ as a screening tool in the preterm population (Skellern 2001, Schonhaut 2013). Revalidation of the ASQ in a larger and more comprehensive population of very preterm infants is necessary. In addition, more work needs to be done around the acceptability of these questionnaires among parents and health professionals, especially as the 'burden' of assessment would be placed on the parents, and their use by parents with potential language barriers, cultural differences and with literacy problems.

In the meantime, in the absence of the standardised use of a single assessment tool to allow comparison of outcomes between centres, we could seek to improve the validity of routine data recording. I have shown that the low sensitivities for identifying impairment through routine NHS assessments exist mainly in the cognitive and language domains. On the electronic '2-year outcome' form, the documentation of outcomes in these domains is more subjective than in the motor domains. It is possible that, by modifying the form to increase the objectivity of the items recorded, the validity of the data would be improved. For example, for the cognitive domain, it may be possible to identify a standardised set of cognitive test items, perhaps from the Bayley-III assessment or other tools, that can be easily administered in a clinical setting. Language function can be ascertained by determining if a child can identify or say words from a list of commonly expressed

words. The list of words can be designed to be independent of language. Clearly, the validity of these approaches will need to be assessed in research studies.

8.4.3 Identification of risk factors for symptoms of ASD in the preterm population

Although a range of perinatal conditions have been linked to autism risk (Gardener 2009, Guinchat 2012), I have not been able to identify specific neonatal factors that contribute to social-communication difficulties. Future studies would benefit from an examination of a more comprehensive set of neonatal and environmental variables in order to identify potential moderators and mediators of risk for ASD. With this understanding, it would then be possible to develop risk scores or risk prediction models that could aid in early diagnosis and initiation of interventional therapies. It would be important to follow-up these children and gain further insight into the developmental trajectories of social-communication in preterm children. This would further our understanding of the significance of these early autistic behaviours as well as the validity of early autism screening in the preterm population using the Q-CHAT and the Bayley-III SE questionnaires. Future studies will also need to focus on the challenges faced in the early assessment of autistic features of children with major functional disabilities and in non-English speaking groups. For example, it would be useful to recognise certain unique behaviours exhibited by toddlers with high risk for autism that are not dependent on physical abilities or language.

8.4.4 Determination of factors that affect the predictive validity of early assessment

In study 3, I was not able to determine the sources of heterogeneity that could account for the differences in predictive validity observed between the studies. This information will shed light on how the sensitivity and specificity of early developmental assessment can be improved, for example the type of assessment tools to be used, the optimal ages at follow-up and the target population to test. As discussed in section 7.4.3, future analyses evaluating the factors that influence the

predictability of early neurodevelopmental assessment will need to be conducted on patient-level data to avoid ecological fallacy.

8.4.5 Linkage with school-age outcome data

Since the outcome ascertained in early childhood is poorly predictive of functioning in later childhood and adolescence, it is important that we continue to record and analyse outcome data from the preterm population. Currently, there is no process or provision in the UK for continuing formal follow-up assessment beyond early childhood for this population. Long-term follow-up programmes require significant manpower and financial investment and the likelihood of high attrition rates will further jeopardise the success of such a programme. It is therefore worthwhile considering other sources of school-age outcome data, for example by linking neonatal to primary care or community child health records using the unique NHS number; or to educational data such as requirements for Special Educational Needs or results on Key Stage assessments that are held in the National Pupil Database. UK national structures provide a unique opportunity for data linkage. Future research could investigate the utility of these data sources and the feasibility of linking neonatal data with later outcome data.

8.5 CONCLUSIONS

I believe that with the ever increasing survival of very preterm infants and the high prevalence of impairment, however subtle, in this population, the ability to capture high quality outcome data is of public health importance. It is crucial that relevant and accurate outcomes are used to evaluate and inform clinical practice. Through my work described in this thesis, I have found that population outcome data recorded during routine follow-up assessments in the NHS compares poorly with data obtained to a research standard. I have discussed the possible explanations and implications of my findings and offered potential avenues for future research to improve the validity of routine data. I

also report that children born very preterm experience significantly greater social-communication problems compared to the general population. This has implications for current clinical practice to address the needs of these children. Finally, I have demonstrated that current practices in relation to determining developmental outcome have limitations in predicting school-age cognitive function. I advocate for longer term outcomes to be sought routinely in order to fully understand the impact of preterm birth on society.

REFERENCES

A report of two working groups convened by the National Perinatal Epidemiology Unit and the former Oxford Regional Health Authority 1994. Disability and perinatal care: a measurement of health status at two years. Oxford: National Perinatal Epidemiology Unit.

Aarnoudse-Moens, C. S., Weisglas-Kuperus, N., van Goudoever, J. B. & Oosterlaan, J. (2009) Meta-analysis of neurobehavioral outcomes in very preterm and/or very low birth weight children. *Pediatrics*; 124: 717-28.

Achenbach, T. M. & L.A., R. (2000). Manual for ASEBA Preschool Forms and Profiles. Burlington, VT: University of Vermont, Research Centre for Children, Youth and Families.

Adams, W. G., Mann, A. M. & Bauchner, H. (2003) Use of an electronic medical record improves the quality of urban pediatric primary care. *Pediatrics*; 111: 626-32.

Akobeng, A. (2006) Understanding diagnostic tests 1: sensitivity, specificity and predictive values. *Acta Paediatr*; 96: 338-341.

Allison, C., Baron-Cohen, S., Wheelwright, S., Charman, T., Richler, J., Pasco, G. & Brayne, C. (2008) The Q-CHAT (Quantitative CHECKlist for Autism in Toddlers): a normally distributed quantitative measure of autistic traits at 18-24 months of age: preliminary report. *J Autism Dev Disord*; 38: 1414-25.

Als, H. (1986) A synactive model of neonatal behavioral organization: Framework for the assessment and support of the neurobehavioral development of the premature infant and his parents in the environment of the neoantal intensive care unit. *Physical and Occupational Therapy in Pediatrics*; 6: 3-55.

American Academy of Pediatrics. (2006) Identifying infants and young children with developmental disorders in the medical home: an algorithm for developmental surveillance and screening. *Pediatrics*; 118: 405-20.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (1999) Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association.

Amess, P., Young, T., Burley, H. & Khan, Y. (2010) Developmental outcome of very preterm babies using an assessment tool deliverable by health visitors. *Eur J Paediatr Neuro*; 14: 219-23.

Anderson, P. J., De Luca, C. R., Hutchinson, E., Roberts, G. & Doyle, L. W. (2010) Underestimation of developmental delay by the new Bayley-III Scale. *Arch Pediatr Adolesc Med*; 164: 352-6.

Anderson, P. J. & Doyle, L. W. (2008) Cognitive and educational deficits in children born extremely preterm. *Semin Perinatol*; 32: 51-8.

Ari-Even Roth, D., Hildesheimer, M., Maayan-Metzger, A., Muchnik, C., Hamburger, A., Mazkeret, R. & Kuint, J. (2006) Low prevalence of hearing impairment among very low birthweight infants as detected by universal neonatal hearing screening. *Arch Dis Child Fetal Neonatal Ed*; 91: F257-62.

Audit Commission. (1993) Children first. A study of hospital services. (Audit Commission NHS report No. 7). London: HMSO.

Autism and Developmental Disabilities Monitoring Network Surveillance Year 2006 Principal Investigators; Centers for Disease Control and Prevention (CDC). (2009) Prevalence of autism spectrum disorders - Autism and Developmental Disabilities Monitoring Network, United States 2006. *MMWR Surveillance Summaries*; 58: 1-20.

Aylward, G. P. (2002) Cognitive and neuropsychological outcomes: more than IQ scores. *Ment Retard Dev Disabil Res Rev*; 8: 234-40.

Aylward, G. P. (2004) Prediction of function from infancy to early childhood: implications for pediatric psychology. *J Pediatr Psychol*; 29: 555-64.

Aylward, G. P. & Aylward, B. S. (2011) The changing yardstick in measurement of cognitive abilities in infancy. *J Dev Behav Pediatr*; 32: 465-8.

Bar-Shalita, T., Vatine, J. J. & Parush, S. (2008) Sensory modulation disorder: a risk factor for participation in daily life activities. *Dev Med Child Neurol*; 50: 932-7.

Baron-Cohen, S., Allen, J. & Gillberg, C. (1992) Can autism be detected at 18 months? The needle, the haystack, and the CHAT. *Br J Psychiatry*; 161: 839-43.

Baron-Cohen, S., Wheelwright, S., Cox, A., Baird, G., Charman, T., Swettenham, J., Drew, A. & Doehring, P. (2000) Early identification of autism by the CHecklist for Autism in Toddlers (CHAT). *J R Soc Med*; 93: 521-5.

Barre, N., Morgan, A., Doyle, L. W. & Anderson, P. J. (2011) Language abilities in children who were very preterm and/or very low birth weight: a meta-analysis. *J Pediatr*; 158: 766-774 e1.

Bart, O., Shayevits, S., Gabis, L. V. & Morag, I. (2011) Prediction of participation and sensory modulation of late preterm infants at 12 months: a prospective study. *Res Dev Disabil*; 32: 2732-8.

Bassan, H., Stolar, O., Geva, R., Eshel, R., Fattal-Valevski, A., Leitner, Y., Waron, M., Jaffa, A. & Harel, S. (2011) Intrauterine growth-restricted neonates born at term or preterm: how different? *Pediatr Neurol*; 44: 122-30.

Bayley, N. (1993) Manual for the Bayley Scales of Infant Development. San Antonio, TX: Psychological Corporation.

Bayley, N. (2006a) Administration manual for the Bayley Scales of Infant and Toddler Development (third edition), San Antonio, TX: Psychological Corporation.

Bayley, N. (2006b) Technical manual for the Bayley Scales of Infant and Toddler Development (third edition), San Antonio, TX: Psychological Corporation.

Bayley, N. (2008) Bayley-III technical report 2: Factors contributing to differences between Bayley-III and BSID-II scores [Online]. Pearson Education. Available: http://images.pearsonassessments.com/images/tmrs/tmrs_rg/BayleyIII_TechRep.pdf?WT.mc_id=T_MRS_Bayley_III_Technical_Report_2 [Accessed 24 August 2014].

Bayley, N. (2010) Bayley Scales of Infant and Toddler Development (third edition) UK and Ireland supplement manual, San Antonio, TX: Psychological Corporation.

Bedford, H., Walton, S. & Ahn, J. (2013) *Measures of child development: a review* [Online]. London: UCL Institute of Child Health. Available: https://www.ucl.ac.uk/cpru/documents/review_of_measures_of_child_development [Accessed 15 June 2015].

Bellman, M. H., Lingam, S. & Aukett, A. 1996. *Schedule of growing skills II: Reference manual*, London: NFER Nelson Publishing Company Ltd.

Bellman, M. H., Rawson, N. B., Wadsworth, J., Ross, E., Cameron, S. & Miller, D. L. (1985) A developmental test based on the STYCAR sequences used in the national childhood encephalopathy study. *Child: Care, Health & Development*; 11: 309-323.

Bhutta, A. T., Cleves, M. A., Casey, P. H., Cradock, M. M. & Anand, K. J. (2002) Cognitive and behavioral outcomes of school-aged children who were born preterm: a meta-analysis. *JAMA*; 288: 728-37.

Bialystok, E. & Martin, M. M. (2004) Attention and inhibition in bilingual children: evidence from the dimensional change card sort task. *Dev Sci*; 7: 325-39.

Bishop, S. L., Richler, J. & Lord, C. (2006) Association between restricted and repetitive behaviors and nonverbal IQ in children with autism spectrum disorders. *Child Neuropsychol*; 12: 247-67.

Bjorck-Akesson, E., Wilder, J., Granlund, M., Pless, M., Simeonsson, R., Adolfsson, M., Almqvist, L., Augustine, L., Klang, N. & Lillvist, A. (2010) The International Classification of Functioning, Disability and Health and the version for children and youth as a tool in child habilitation/early childhood intervention--feasibility and usefulness as a common language and frame of reference for practice. *Disabil Rehabil*; 32 Suppl 1: S125-38.

Bohin, S., Draper, E. S. & Field, D. J. (1999) Health status of a population of infants born before 26 weeks gestation derived from routine data collected between 21 and 27 months post-delivery. *Early Hum Dev*; 55: 9-18.

Bortolus, R., Parazzini, F., Trevisanuto, D., Cipriani, S., Ferrarese, P. & Zanardo, V. (2002) Developmental assessment of preterm and term children at 18 months: reproducibility and validity of a postal questionnaire to parents. *Acta Paediatr*; 91: 1101-7.

Bossuyt, P., Davenport, C., Deeks, J., Hyde, C., Leeflang, M. & Scholten, R. (2013) Chapter 11: Interpreting results and drawing conclusions. In: Deeks, J. J., Bossuyt, P. M. & Gatsonis, C. (Eds.) *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 0.9.*: The Cochrane Collaboration.

Bowen, J. R., Gibson, F. L., Leslie, G. I., Arnold, J. D., Ma, P. J. & Starte, D. R. (1996) Predictive value of the Griffiths assessment in extremely low birthweight infants. *J Paediatr Child Health*; 32: 25-30.

Bowers, K., Wink, L. K., Pottenger, A., McDougle, C. J. & Erickson, C. (2015) Phenotypic differences in individuals with autism spectrum disorder born preterm and at term gestation. *Autism*; 19: 758-63.

Boyd, L. A., Msall, M. E., O'Shea, T. M., Allred, E. N., Hounshell, G. & Leviton, A. (2013) Social-emotional delays at 2 years in extremely low gestational age survivors: correlates of impaired orientation/engagement and emotional regulation. *Early Hum Dev*; 89: 925-30.

Bracewell, M. & Marlow, N. (2002) Patterns of motor disability in very preterm children. *Ment Retard Dev Disabil Res Rev*; 8: 241-8.

Brenner, H. & Gefeller, O. (1997) Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. *Stat Med*; 16: 981-91.

British Association of Perinatal Medicine. (2001) Standards for hospitals providing intensive and high dependency care. 2nd ed. London: British Association of Perinatal Medicine.

British Association of Perinatal Medicine Working Party. (1997) The BAPM neonatal dataset for the annual reporting of data by neonatal intensive care units. London: The British Association of Perinatal Medicine.

British Association of Perinatal Medicine Working Party. (2008) Classification of health status at 2 years as a perinatal outcome. London: The British Association of Perinatal Medicine.

Brocklehurst, P., Farrell, B., King, A., Juszczak, E., Darlow, B., Haque, K., Salt, A., Stenson, B. & Tarnow-Mordi, W. (2011) Treatment of neonatal sepsis with intravenous immune globulin. *N Engl J Med*; 365: 1201-11.

Brogan, E., Cragg, L., Gilmore, C., Marlow, N., Simms, V. & Johnson, S. (2014) Inattention in very preterm children: implications for screening and detection. *Arch Dis Child*; 99: 834-9.

Bruggink, J. L., Van Braeckel, K. N. & Bos, A. F. (2010) The early motor repertoire of children born preterm is associated with intelligence at school age. *Pediatrics*; 125: e1356-63.

Callanan, C., Doyle, L., Rickards, A., Kelly, E., Ford, G. & Davis, N. (2001) Children followed with difficulty: how do they differ? *J Paediatr Child Health*; 37: 152-6.

Campbell, M. K., Halinda, E., Carlyle, M. J., Fox, A. M., Turner, L. A. & Chance, G. W. (1993) Factors predictive of follow-up clinic attendance and developmental outcome in a regional cohort of very low birth weight infants. *Am J Epidemiol*; 138: 704-13.

Canadian Neonatal Follow-up Network [Online]. Available: <http://cnfun.ca>.

Casenhiser, D., Breinbauer, C. & Greenspan, S. (2007) Evaluating Greenspan's social emotional growth scale/ chart as a screening for autism. *The ICDL 11th Annual International Conference: Critical Factors for Optimal Outcomes for Children with Autism and Special Needs*. Tyson's Corner, VA.

Catlett, A. T., Thompson, R. J., Jr., Johndrow, D. A. & Boshkoff, M. R. (1993) Risk status for dropping out of developmental followup for very low birth weight infants. *Public Health Rep*; 108: 589-94.

Charkaluk, M. L., Truffert, P., Marchand-Martin, L., Mur, S., Kaminski, M., Ancel, P. Y., Pierrat, V. & group, E. s. (2011) 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*; 87: 297-302.

Chaudhary, T., Walch, E., Herold, B., Metze, B., Lejeune, A., Burkhardt, F. & Buhner, C. (2013) Predictive and concurrent validity of standardized neurodevelopmental examinations by the Griffiths scales and Bayley scales of infant development II. *Klin Padiatr*; 225: 8-12.

Claas, M. J., de Vries, L. S., Bruinse, H. W., van Haastert, I. C., Uniken Venema, M. M., Peelen, L. M. & Koopman, C. (2011) Neurodevelopmental outcome over time of preterm born children ≤ 750 g at birth. *Early Hum Dev*; 87: 183-91.

Clinical Standards Advisory Group. (1993) Neonatal intensive care. London: HMSO.

Cochran, W. G. (1954) Some methods for strengthening the common Chi² tests. *Biometrics*; 10: 417-451.

Cohen, S. E. (1995) Biosocial factors in early infancy as predictors of competence in adolescents who were born prematurely. *J Dev Behav Pediatr*; 16: 36-41.

Cristobal, R. & Oghalai, J. S. (2008) Hearing loss in children with very low birth weight: current review of epidemiology and pathophysiology. *Arch Dis Child Fetal Neonatal Ed*; 93: F462-8.

Cumberledge, J. (1993) Changing childbirth. Part I - report of the expert maternity group. Winterton report. London: HMSO.

Cummings, S. M., Savitz, L. A. & Konrad, T. R. (2001) Reported response rates to mailed physician questionnaires. *Health Serv Res*; 35: 1347-55.

D'Amore, A., Broster, S., Le Fort, W. & Curley, A. (2010) Two-year outcomes from very low birthweight infants in a geographically defined population across 10 years, 1993-2002: comparing 1993-1997 with 1998-2002. *Arch Dis Child Fetal Neonatal Ed*; 96: F176-85.

Da Costa, D., Bann, C. M., Hansen, N. I., Shankaran, S. & Delaney-Black, V. (2009) Validation of the Functional Status II questionnaire in the assessment of extremely-low-birthweight infants. *Dev Med Child Neurol*; 51: 536-44.

Dawson, C., Perkins, M., Draper, E., Johnson, A. & Field, D. (1997) Are outcome data regarding the survivors of neonatal care available from routine sources? *Arch Dis Child Fetal Neonatal Ed*; 77: F206-10.

De Groote, I., Vanhaesebrouck, P., Bruneel, E., Dom, L., Durein, I., Hasaerts, D., Laroche, S., Oostra, A., Ortibus, E., Roeyers, H. & van Mol, C. (2007) Outcome at 3 years of age in a population-based cohort of extremely preterm infants. *Obstet Gynecol*; 110: 855-64.

de Kieviet, J. F., Piek, J. P., Aarnoudse-Moens, C. S. & Oosterlaan, J. (2009) Motor development in very preterm and very low-birth-weight children from birth to adolescence: a meta-analysis. *JAMA*; 302: 2235-42.

de Waal, C. G., Weisglas-Kuperus, N., van Goudoever, J. B. & Walther, F. J. (2012) Mortality, neonatal morbidity and two year follow-up of extremely preterm infants born in The Netherlands in 2007. *PLoS One*; 7: e41302.

Deeks, J. J. (2001) Systematic reviews of evaluations of diagnostic and screening tests. In: Egger, M., Davey Smith, G. & Altman, D. G. (Eds.) *Systematic reviews in health care: meta-analysis in context*. 2nd ed. London: BMJ.

Deeks, J. J., Bossuyt, P. M. & Gatsonis, C. (2013) *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 0.9*. [Online]. The Cochrane Collaboration. Available: <http://srdta.cochrane.org/>.

Deeks, J. J., Macaskill, P. & Irwig, L. (2005) The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *J Clin Epidemiol*; 58: 882-93.

Department for Communities and Local Government. (2011) *English Indices of Deprivation 2010* [Online]. London. Available: <http://www.communities.gov.uk/documents/statistics/pdf/1871208.pdf>.

Dhaliwal, C., Fleck, B., Wright, E., Graham, C. & McIntosh, N. (2008) Incidence of retinopathy of prematurity in Lothian, Scotland, from 1990 to 2004. *Arch Dis Child Fetal Neonatal Ed*; 93: F422-6.

Dietz, C., Swinkels, S., van Daalen, E., van Engeland, H. & Buitelaar, J. K. (2006) Screening for autistic spectrum disorder in children aged 14-15 months. II: population screening with the Early Screening of Autistic Traits Questionnaire (ESAT). Design and general findings. *J Autism Dev Disord*; 36: 713-22.

Dorling, J. S. & Field, D. J. (2006) Follow up of infants following discharge from the neonatal unit: structure and process. *Early Hum Dev*; 82: 151-6.

Doyle, L. W. & Anderson, P. J. (2010a) Adult outcome of extremely preterm infants. *Pediatrics*; 126: 342-51.

Doyle, L. W., Roberts, G. & Anderson, P. J. (2010b) Outcomes at age 2 years of infants < 28 weeks' gestational age born in Victoria in 2005. *J Pediatr*; 156: 49-53 e1.

Doyle, L. W., Roberts, G. & Anderson, P. J. (2011) Changing long-term outcomes for infants 500-999 g birth weight in Victoria, 1979-2005. *Arch Dis Child Fetal Neonatal Ed*; 96: F443-7.

Dudova, I., Markova, D., Kasparova, M., Zemankova, J., Beranova, S., Urbanek, T. & Hrdlicka, M. (2014) Comparison of three screening tests for autism in preterm children with birth weights less than 1,500 grams. *Neuropsychiatr Dis Treat*; 10: 2201-8.

Dunn, W. (2002) *Infant/Toddler Sensory Profile. User's Manual*, San Antonio, TX: The Psychological Corporation.

Eldevik, S., Hastings, R. P., Hughes, J. C., Jahr, E., Eikeseth, S. & Cross, S. (2009) Meta-analysis of Early Intensive Behavioral Intervention for children with autism. *J Clin Child Adolesc Psychol*; 38: 439-50.

Eldevik, S., Hastings, R. P., Hughes, J. C., Jahr, E., Eikeseth, S. & Cross, S. (2010) Using participant data to extend the evidence base for intensive behavioral intervention for children with autism. *Am J Intellect Dev Disabil*; 115: 381-405.

Elgen, I., Sommerfelt, K. & Markestad, T. (2002) Population based, controlled study of behavioural problems and psychiatric disorders in low birthweight children at 11 years of age. *Arch Dis Child Fetal Neonatal Ed*; 87: F128-32.

Eliasson, A. C., Krumlinde-Sundholm, L., Rosblad, B., Beckung, E., Arner, M., Ohrvall, A. M. & Rosenbaum, P. (2006) The Manual Ability Classification System (MACS) for children with cerebral palsy: scale development and evidence of validity and reliability. *Dev Med Child Neurol*; 48: 549-54.

EPICure [Online]. Available: <http://www.epicure.ac.uk>.

Farooqi, A., Hagglof, B., Sedin, G., Gothefors, L. & Serenius, F. (2007) Mental health and social competencies of 10- to 12-year-old children born at 23 to 25 weeks of gestation in the 1990s: a Swedish national prospective follow-up study. *Pediatrics*; 120: 118-33.

Fedrizzi, E., Inverno, M., Botteon, G., Anderloni, A., Filippini, G. & Farinotti, M. (1993) The cognitive development of children born preterm and affected by spastic diplegia. *Brain Dev*; 15: 428-32.

Fewtrell, M. S., Kennedy, K., Singhal, A., Martin, R. M., Ness, A., Hadders-Algra, M., Koletzko, B. & Lucas, A. (2008) How much loss to follow-up is acceptable in long-term randomised trials and prospective studies? *Arch Dis Child*; 93: 458-61.

Field, D., Draper, E. S., Gompels, M. J., Green, C., Johnson, A., Shortland, D., Blair, M., Manktelow, B., Lamming, C. R. & Law, C. (2001) Measuring later health status of high risk infants: randomised comparison of two simple methods of data collection. *BMJ*; 323: 1276-81.

Flynn, J. (1999) Searching for justice. The discovery of IQ gains over time. *Am Psychol*; 54: 5-20.

Fooks, J. (1999) Four key questions that identify severe disability. *Arch Dis Child*; 80: 67-8.

Fooks, J., Mutch, L., Yudkin, P., Johnson, A. & Elbourne, D. (1997) Comparing two methods of follow up in a multicentre randomised trial. *Arch Dis Child*; 76: 369-76.

Foster-Cohen, S., Edgin, J. O., Champion, P. R. & Woodward, L. J. (2007) Early delayed language development in very preterm infants: evidence from the MacArthur-Bates CDI. *J Child Lang*; 34: 655-75.

Frisone, M. F., Mercuri, E., Laroche, S., Foglia, C., Maalouf, E. F., Haataja, L., Cowan, F. & Dubowitz, L. (2002) Prognostic value of the neurologic optimality score at 9 and 18 months in preterm infants born before 31 weeks' gestation. *J Pediatr*; 140: 57-60.

Gardener, H., Spiegelman, D. & Buka, S. L. (2009) Prenatal risk factors for autism: comprehensive meta-analysis. *Br J Psychiatry*; 195: 7-14.

Gollenberg, A. L., Lynch, C. D., Jackson, L. W., McGuinness, B. M. & Msall, M. E. (2010) Concurrent validity of the parent-completed Ages and Stages Questionnaires, 2nd Ed. with the Bayley Scales of Infant Development II in a low-risk sample. *Child Care Health Dev*; 36: 485-90.

Gordis, L. (2014) Assessing the validity and reliability of diagnostic and screening tests. *Epidemiology*. 5th ed. Philadelphia, PA: Elsevier.

Gray, D., Woodward, L. J., Spencer, C., Inder, T. E. & Austin, N. C. (2006) Health service utilisation of a regional cohort of very preterm infants over the first 2 years of life. *J Paediatr Child Health*; 42: 377-83.

Gray, P. H., Burns, Y. R., Mohay, H. A., O'Callaghan, M. J. & Tudehope, D. I. (1995) Neurodevelopmental outcome of preterm infants with bronchopulmonary dysplasia. *Arch Dis Child Fetal Neonatal Ed*; 73: F128-34.

Gray, P. H., Edwards, D. M., O'Callaghan, M. J. & Gibbons, K. (2015) Screening for autism spectrum disorder in very preterm infants during early childhood. *Early Hum Dev*; 91: 271-6.

Guillen, U., DeMauro, S., Ma, L., Zupancic, J., Roberts, R., Schmidt, B. & Kirpalani, H. (2012) Relationship between attrition and neurodevelopmental impairment rates in extremely preterm infants at 18 to 24 months: a systematic review. *Arch Pediatr Adolesc Med*; 166: 178-84.

Guinchat, V., Thorsen, P., Laurent, C., Cans, C., Bodeau, N. & Cohen, D. (2012) Pre-, peri- and neonatal risk factors for autism. *Acta Obstet Gynecol Scand*; 91: 287-300.

Gutbrod, T., Wolke, D., Soehne, B., Ohrt, B. & Riegel, K. (2000) 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*; 82: F208-14.

Haataja, L., Mercuri, E., Regev, R., Cowan, F., Rutherford, M., Dubowitz, V. & Dubowitz, L. (1999) Optimality score for the neurologic examination of the infant at 12 and 18 months of age. *J Pediatr*; 135: 153-61.

Hack, M., Taylor, H. G., Drotar, D., Schluchter, M., Cartar, L., Wilson-Costello, D., Klein, N., Friedman, H., Mercuri-Minich, N. & Morrow, M. (2005) Poor predictive validity of the Bayley Scales of Infant Development for cognitive function of extremely low birth weight children at school age. *Pediatrics*; 116: 333-41.

Halfon, N., Regalado, M., Sareen, H., Inkelas, M., Reuland, C. H., Glascoe, F. P. & Olson, L. M. (2004) Assessing development in the pediatric office. *Pediatrics*; 113: 1926-33.

Harbord, R. M., Deeks, J. J., Egger, M., Whiting, P. & Sterne, J. A. (2007) A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics*; 8: 239-51.

Hille, E. T., Weisglas-Kuperus, N., van Goudoever, J. B., Jacobusse, G. W., Ens-Dokkum, M. H., de Groot, L., Wit, J. M., Geven, W. B., Kok, J. H., de Kleine, M. J., Kollée, L. A., Mulder, A. L., van Straaten, H. L., de Vries, L. S., van Weissenbruch, M. M., Verloove-Vanhorick, S. P. & the Dutch Collaborative POPS 19 Study Group. (2007) Functional outcomes and participation in young adulthood for very

preterm and very low birth weight infants: the Dutch Project on Preterm and Small for Gestational Age Infants at 19 years of age. *Pediatrics*; 120: e587-95.

Hintz, S. R., Kendrick, D. E., Wilson-Costello, D. E., Das, A., Bell, E. F., Vohr, B. R. & Higgins, R. D. (2011) Early-childhood neurodevelopmental outcomes are not improving for infants born at <25 weeks' gestational age. *Pediatrics*; 127: 62-70.

Holmstrom, G. E., Kallen, K., Hellstrom, A., Jakobsson, P. G., Serenius, F., Stjernqvist, K. & Tornqvist, K. (2014) Ophthalmologic outcome at 30 months' corrected age of a prospective Swedish cohort of children born before 27 weeks of gestation: the extremely preterm infants in Sweden study. *JAMA Ophthalmol*; 132: 182-9.

House of Commons Health Committee Session 1991-2. (1992) Maternity services: second report. London: HMSO.

Howlin, P., Magiati, I. & Charman, T. (2009) Systematic review of early intensive behavioral interventions for children with autism. *Am J Intellect Dev Disabil*; 114: 23-41.

Hutchinson, E. A., De Luca, C. R., Doyle, L. W., Roberts, G., Anderson, P. J. & the Victorian Infant Collaborative Study Group. (2013) School-age outcomes of extremely preterm or extremely low birth weight children. *Pediatrics*; 131: e1053-61.

Indredavik, M. S., Vik, T., Heyerdahl, S., Kulseng, S., Fayers, P. & Brubakk, A. M. (2004) Psychiatric symptoms and disorders in adolescents with low birth weight. *Arch Dis Child Fetal Neonatal Ed*; 89: F445-50.

Jahnsen, R., Aamodt, G. & Rosenbaum, P. (2006) Gross Motor Function Classification System used in adults with cerebral palsy: agreement of self-reported versus professional rating. *Dev Med Child Neurol*; 48: 734-8.

Jansson-Verkasalo, E., Valkama, M., Vainionpaa, L., Paakko, E., Ilkko, E. & Lehtihalmes, M. (2004) Language development in very low birth weight preterm children: a follow-up study. *Folia Phoniatr Logop*; 56: 108-19.

Jeevanantham, D., Dyszuk, E. & Bartlett, D. (2015) The Manual Ability Classification System: A Scoping Review. *Pediatr Phys Ther*; 27: 236-41.

Johnson, A. & King, R. (1999) Can routine information systems be used to monitor serious disability? *Arch Dis Child*; 80: 63-6.

Johnson, C. P. & Myers, S. M. (2007) Identification and evaluation of children with autism spectrum disorders. *Pediatrics*; 120: 1183-215.

Johnson, S., Hollis, C., Hennessy, E., Kochhar, P., Wolke, D. & Marlow, N. (2011a) Screening for autism in preterm children: diagnostic utility of the Social Communication Questionnaire. *Arch Dis Child*; 96: 73-7.

Johnson, S., Hollis, C., Kochhar, P., Hennessy, E., Wolke, D. & Marlow, N. (2010a) Autism spectrum disorders in extremely preterm children. *J Pediatr*; 156: 525-31 e2.

Johnson, S., Hollis, C., Kochhar, P., Hennessy, E., Wolke, D. & Marlow, N. (2010b) Psychiatric disorders in extremely preterm children: longitudinal finding at age 11 years in the EPICure study. *J Am Acad Child Adolesc Psychiatry*; 49: 453-63 e1.

Johnson, S. & Marlow, N. (2006) Developmental screen or developmental testing? *Early Hum Dev*; 82: 173-83.

Johnson, S. & Marlow, N. (2009) Positive screening results on the modified checklist for autism in toddlers: implications for very preterm populations. *J Pediatr*; 154: 478-80.

Johnson, S. & Marlow, N. (2011b) Preterm birth and childhood psychiatric disorders. *Pediatr Res*; 69: 11R-8R.

Johnson, S., Marlow, N., Wolke, D., Davidson, L., Marston, L., O'Hare, A., Peacock, J. & Schulte, J. (2004) Validation of a parent report measure of cognitive development in very preterm infants. *Dev Med Child Neurol*; 46: 389-97.

Johnson, S., Moore, T. & Marlow, N. (2014) Using the Bayley-III to assess neurodevelopmental delay: which cut-off should be used? *Pediatr Res*; 75: 670-4.

Johnson, S., Wolke, D. & Marlow, N. (2008) Developmental assessment of preterm infants at 2 years: validity of parent reports. *Dev Med Child Neurol*; 50: 58-62.

Jones, H. P., Guildea, Z. E., Stewart, J. H. & Cartlidge, P. H. (2002) The Health Status Questionnaire: achieving concordance with published disability criteria. *Arch Dis Child*; 86: 15-20.

Kilbride, H. W., Daily, D. K., Clafin, K., Hall, R. T., Maulik, D. & Grundy, H. O. (1990) Improved survival and neurodevelopmental outcome for infants less than 801 grams birthweight. *Am J Perinatol*; 7: 160-5.

Kim, M. M., O'Connor, K. S., McLean, J., Robson, A. & Chance, G. (1996) Do parents and professionals agree on the developmental status of high-risk infants? *Pediatrics*; 97: 676-81.

Korres, S., Nikolopoulos, T. P., Komkotou, V., Balatsouras, D., Kandiloros, D., Constantinou, D. & Ferekidis, E. (2005) Newborn hearing screening: effectiveness, importance of high-risk factors, and characteristics of infants in the neonatal intensive care unit and well-baby nursery. *Otol Neurotol*; 26: 1186-90.

Korres, S., Nikolopoulos, T. P., Peraki, E. E., Tsiakou, M., Karakitsou, M., Apostolopoulos, N., Economides, J., Balatsouras, D. & Ferekidis, E. (2008) Outcomes and efficacy of newborn hearing screening: strengths and weaknesses (success or failure?). *Laryngoscope*; 118: 1253-6.

Kuban, K. C., Allred, E. N., O'Shea, M., Paneth, N., Pagano, M. & Leviton, A. (2008) An algorithm for identifying and classifying cerebral palsy in young children. *J Pediatr*; 153: 466-72.

Kuban, K. C., O'Shea, T. M., Allred, E. N., Tager-Flusberg, H., Goldstein, D. J. & Leviton, A. (2009) Positive screening on the Modified Checklist for Autism in Toddlers (M-CHAT) in extremely low gestational age newborns. *J Pediatr*; 154: 535-540 e1.

- Kutz, P., Horsch, S., Kuhn, L. & Roll, C. (2009) Single-centre vs. population-based outcome data of extremely preterm infants at the limits of viability. *Acta Paediatr*; 98: 1451-5.
- Kuzniewicz, M. W., Wi, S., Qian, Y., Walsh, E. M., Armstrong, M. A. & Croen, L. A. (2014) Prevalence and neonatal factors associated with autism spectrum disorders in preterm infants. *J Pediatr*; 164: 20-5.
- Landis, J. R. & Koch, G. G. (1977) The measurement of observer agreement for categorical data. *Biometrics*; 33: 159-74.
- Larroque, B., Delobel, M., Arnaud, C. & Marchand, L. (2008) Outcome at 5 and 8 years of children born very preterm. *Arch Pediatr*; 15: 589-91.
- Leavey, A., Zwaigenbaum, L., Heavner, K. & Burstyn, I. (2013) Gestational age at birth and risk of autism spectrum disorders in Alberta, Canada. *J Pediatr*; 162: 361-8.
- Leeflang, M. M., Bossuyt, P. M. & Irwig, L. (2009) Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. *J Clin Epidemiol*; 62: 5-12.
- Leveresen, K. T., Sommerfelt, K., Ronnestad, A., Kaaresen, P. I., Farstad, T., Skranes, J., Stoen, R., Bircow Elgen, I., Rettedal, S., Egil Eide, G., Irgens, L. M. & Markestad, T. (2011) Prediction of neurodevelopmental and sensory outcome at 5 years in Norwegian children born extremely preterm. *Pediatrics*; 127: e630-8.
- Liao, P. J. & Campbell, S. K. (2004) Examination of the item structure of the Alberta infant motor scale. *Pediatr Phys Ther*; 16: 31-8.
- Limperopoulos, C., Bassan, H., Sullivan, N. R., Soul, J. S., Robertson, R. L., Jr., Moore, M., Ringer, S. A., Volpe, J. J. & du Plessis, A. J. (2008) Positive screening for autism in ex-preterm infants: prevalence and risk factors. *Pediatrics*; 121: 758-65.
- Littenberg, B. & Moses, L. E. (1993) Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. *Med Decis Making*; 13: 313-21.
- London Perinatal Group. 2008. *London Perinatal Networks Annual Report* [Online]. Available: <http://www.neonatal.org.uk/documents/4391.pdf>.
- Losh, M., Esserman, D., Anckarsater, H., Sullivan, P. F. & Lichtenstein, P. (2012) Lower birth weight indicates higher risk of autistic traits in discordant twin pairs. *Psychol Med*; 42: 1091-102.
- Lowe, J. R., Erickson, S. J., Schrader, R. & Duncan, A. F. (2012) Comparison of the Bayley II Mental Developmental Index and the Bayley III Cognitive Scale: are we measuring the same thing? *Acta Paediatr*; 101: e55-8.
- Lowe, J. R., Nolen, T. L., Vohr, B., Adams-Chapman, I., Duncan, A. F. & Watterberg, K. (2013) Effect of primary language on developmental testing in children born extremely preterm. *Acta Paediatr*; 102: 896-900.

Luttikhuisen dos Santos, E. S., de Kieviet, J. F., Konigs, M., van Elburg, R. M. & Oosterlaan, J. (2013) 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*; 89: 487-96.

Luyster, R. J., Kuban, K. C., O'Shea, T. M., Paneth, N., Allred, E. N. & Leviton, A. (2011) The Modified Checklist for Autism in Toddlers in extremely low gestational age newborns: individual items associated with motor, cognitive, vision and hearing limitations. *Paediatr Perinat Epidemiol*; 25: 366-76.

Lyon, A. (2007) How should we report neonatal outcomes? *Semin Fetal Neonatal Med*; 12: 332-6.

Marlow, N. (2004) Neurocognitive outcome after very preterm birth. *Arch Dis Child Fetal Neonatal Ed*; 89: F224-8.

Marlow, N., Greenough, A., Peacock, J. L., Marston, L., Limb, E. S., Johnson, A. H. & Calvert, S. A. (2006) Randomised trial of high frequency oscillatory ventilation or conventional ventilation in babies of gestational age 28 weeks or less: respiratory and neurological outcomes at 2 years. *Arch Dis Child Fetal Neonatal Ed*; 91: F320-6.

Marlow, N., Wolke, D., Bracewell, M. A. & Samara, M. (2005) Neurologic and developmental disability at six years of age after extremely preterm birth. *N Engl J Med*; 352: 9-19.

Martin, A. J., Darlow, B. A., Salt, A., Hague, W., Sebastian, L., McNeill, N. & Tarnow-Mordi, W. (2013) Performance of the Parent Report of Children's Abilities-Revised (PARCA-R) versus the Bayley Scales of Infant Development III. *Arch Dis Child*; 98: 955-8.

McGrath, M. M., Sullivan, M. C., Lester, B. M. & Oh, W. (2000) Longitudinal neurologic follow-up in neonatal intensive care unit survivors with various neonatal morbidities. *Pediatrics*; 106: 1397-405.

Mercier, C. E., Dunn, M. S., Ferrelli, K. R., Howard, D. B. & Soll, R. F. (2010) Neurodevelopmental outcome of extremely low birth weight infants from the Vermont Oxford network: 1998-2003. *Neonatology*; 97: 329-38.

Milligan, D. W. (2010) Outcomes of children born very preterm in Europe. *Arch Dis Child Fetal Neonatal Ed*; 95: F234-40.

Modi, N. & Carpenter, T. (1997) Fetal growth and coronary heart disease. *Lancet*; 349: 286-7.

Moher, D., Liberati, A., Tetzlaff, J. & Altman, D. G. (2009) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ*; 339: b2535.

Moore, G. S., Kneitel, A. W., Walker, C. K., Gilbert, W. M. & Xing, G. (2012a) Autism risk in small- and large-for-gestational-age infants. *Am J Obstet Gynecol*; 206: 314 e1-9.

Moore, T., Hennessy, E. M., Myles, J., Johnson, S. J., Draper, E. S., Costeloe, K. L. & Marlow, N. (2012b) Neurological and developmental outcome in extremely preterm children born in England in 1995 and 2006: the EPICure studies. *BMJ*; 345: e7961.

Moore, T., Johnson, S., Haider, S., Hennessy, E. & Marlow, N. (2012c) Relationship between test scores using the second and third editions of the Bayley Scales in extremely preterm children. *J Pediatr*; 160: 553-8.

Moore, T., Johnson, S., Hennessy, E. & Marlow, N. (2012d) Screening for autism in extremely preterm infants: problems in interpretation. *Dev Med Child Neurol*; 54: 514-20.

Morrison, A., Polisena, J., Husereau, D., Moulton, K., Clark, M., Fiander, M., Mierzwinski-Urban, M., Clifford, T., Hutton, B. & Rabb, D. (2012) The effect of English-language restriction on systematic review-based meta-analyses: a systematic review of empirical studies. *Int J Technol Assess Health Care*; 28: 138-44.

Moses, L. E., Shapiro, D. & Littenberg, B. (1993) Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Stat Med*; 12: 1293-316.

Movsas, T. Z. & Paneth, N. (2012) The effect of gestational age on symptom severity in children with autism spectrum disorder. *J Autism Dev Disord*; 42: 2431-9.

Munck, P., Niemi, P., Lapinleimu, H., Lehtonen, L., Haataja, L. & PIPARI Study Group. (2012) Stability of cognitive outcome from 2 to 5 years of age in very low birth weight children. *Pediatrics*; 129: 503-8.

Myers, S. M. & Johnson, C. P. (2007) Management of children with autism spectrum disorders. *Pediatrics*; 120: 1162-82.

Nadig, A. S., Ozonoff, S., Young, G. S., Rozga, A., Sigman, M. & Rogers, S. J. (2007) A prospective study of response to name in infants at risk for autism. *Arch Pediatr Adolesc Med*; 161: 378-83.

National Institute for Health and Care Excellence Topic Expert Group and project team 2010. Specialist neonatal care quality standard. London: National Institute for Health and Care Excellence.

National Neonatal Audit programme [Online]. Available: <http://www.rcpch.ac.uk/improving-child-health/quality-improvement-and-clinical-audit/national-neonatal-audit-programme-nn-3>

National Perinatal Epidemiology Unit and the former Oxford Regional Health Authority. (1994) A report of two working groups convened by the National Perinatal Epidemiology Unit and the former Oxford Regional Health Authority. Disability and perinatal care: a measurement of health status at two years. Oxford: National Perinatal Epidemiology Unit.

NICE. National Institute for Health and Care Excellence: Developmental follow-up of preterm babies guideline [Online]. Available: <https://www.nice.org.uk/guidance/indevelopment/gid-cgwave0752>.

O'Connor, A. R., Stephenson, T., Johnson, A., Tobin, M. J., Moseley, M. J., Ratib, S., Ng, Y. & Fielder, A. R. (2002) Long-term ophthalmic outcome of low birth weight children with and without retinopathy of prematurity. *Pediatrics*; 109: 12-8.

Obuchowski, N. A. & Zhou, X. H. (2002) Prospective studies of diagnostic test accuracy when disease prevalence is low. *Biostatistics*; 3: 477-92.

Oddie, S., Morris, S., Gray, D., Fitz-Simon, N. & NNAP Project Board. (2014) National neonatal audit programme annual report 2013. London: Royal College of Paediatrics and Child Health.

Ohrvall, A. M., Krumlinde-Sundholm, L. & Eliasson, A. C. (2014) The stability of the Manual Ability Classification System over time. *Dev Med Child Neurol*; 56: 185-9.

Oosterling, I. J., Swinkels, S. H., van der Gaag, R. J., Visser, J. C., Dietz, C. & Buitelaar, J. K. (2009) Comparative analysis of three screening instruments for autism spectrum disorder in toddlers at high risk. *J Autism Dev Disord*; 39: 897-909.

Orchinik, L. J., Taylor, H. G., Espy, K. A., Minich, N., Klein, N., Sheffield, T. & Hack, M. (2011) Cognitive outcomes for extremely preterm/extremely low birth weight children in kindergarten. *J Int Neuropsychol Soc*; 17: 1067-79.

Ozonoff, S., Macari, S., Young, G. S., Goldring, S., Thompson, M. & Rogers, S. J. (2008) Atypical object exploration at 12 months of age is associated with autism in a prospective sample. *Autism*; 12: 457-72.

Paez, M. M., Tabors, P. O. & Lopez, L. M. (2007) Dual language and literacy development of Spanish-speaking preschool children. *J Appl Dev Psychol*; 28: 85-102.

Palisano, R., Rosenbaum, P., Walter, S., Russell, D., Wood, E. & Galuppi, B. Gross Motor Function Classification System (GMFCS) [Online]. Hamilton, Ontario: CanChild. Available: http://www.motorgrowth.canchild.ca/en/GMFCS/resources/GMFCS_English.pdf [Accessed 8 July 2014].

Palisano, R., Rosenbaum, P., Walter, S., Russell, D., Wood, E. & Galuppi, B. (1997) Development and reliability of a system to classify gross motor function in children with cerebral palsy. *Dev Med Child Neurol*; 39: 214-23.

Palisano, R. J., Cameron, D., Rosenbaum, P. L., Walter, S. D. & Russell, D. (2006) Stability of the gross motor function classification system. *Dev Med Child Neurol*; 48: 424-8.

Pallás Alonso, C. R., de La Cruz Bértolo, J., Medina López, M. C., Orbea Gallardo, C., Gómez Castillo, E. & Simón De Las Heras, R. (2000) [Cerebral palsy and age of sitting and walking in children weighing less than 1,500 g at birth]. *An Esp Pediatr*; 53: 48-52.

Paneth, N., Qiu, H., Rosenbaum, P., Saigal, S., Bishai, S., Jetton, J., den Ouden, L., Broyles, S., Tyson, J. & Kugler, K. (2003) Reliability of classification of cerebral palsy in low-birthweight children in four countries. *Dev Med Child Neurol*; 45: 628-33.

Pedersen, S. J., Sommerfelt, K. & Markestad, T. (2000) Early motor development of premature infants with birthweight less than 2000 grams. *Acta Paediatr*; 89: 1456-61.

Pietz, J., Peter, J., Graf, R., Rauterberg-Ruland, I., Rupp, A., Sontheimer, D. & Linderkamp, O. (2004) Physical growth and neurodevelopmental outcome of nonhandicapped low-risk children born preterm. *Early Hum Dev*; 79: 131-43.

Pinto-Martin, J. A., Levy, S. E., Feldman, J. F., Lorenz, J. M., Paneth, N. & Whitaker, A. H. (2011) Prevalence of autism spectrum disorder in adolescents born weighing <2000 grams. *Pediatrics*; 128: 883-91.

Piper, M. C., Pinnell, L. E., Darrah, J., Maguire, T. & Byrne, P. J. (1992) Construction and validation of the Alberta Infant Motor Scale (AIMS). *Can J Public Health*; 83 Suppl 2: S46-50.

Potharst, E. S., Houtzager, B. A., van Sonderen, L., Tamminga, P., Kok, J. H., Last, B. F. & van Wassenaer, A. G. (2012) 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. *Dev Med Child Neurol*; 54: 240-6.

Pritchard, M. A., Colditz, P. B. & Beller, E. M. (2005) Parents' evaluation of developmental status in children born with a birthweight of 1250 g or less. *J Paediatr Child Health*; 41: 191-6.

Public Health England. (2013) Public health functions to be exercised by NHS England service specification number 20: newborn hearing screening [Online]. London: Department of Health. Available:

https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/256487/20_nhs_newborn_hearing.pdf [Accessed 28 August 2014].

Rahi, J. S. & Cable, N. (2003) Severe visual impairment and blindness in children in the UK. *Lancet*; 362: 1359-65.

Rattihalli, R. R., Lamming, C. R., Dorling, J., Manktelow, B. N., Bohin, S., Field, D. J. & Draper, E. S. (2011) Neonatal intensive care outcomes and resource utilisation of infants born <26 weeks in the former Trent region: 2001-2003 compared with 1991-1993. *Arch Dis Child Fetal Neonatal Ed*; 96: F329-34.

Reitsma, J. B., Glas, A. S., Rutjes, A. W., Scholten, R. J., Bossuyt, P. M. & Zwinderman, A. H. (2005) Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol*; 58: 982-90.

Reuss, M. L., Paneth, N., Pinto-Martin, J. A., Lorenz, J. M. & Susser, M. (1996) The relation of transient hypothyroxinemia in preterm infants to neurologic development at two years of age. *N Engl J Med*; 334: 821-7.

Roberts, G., Anderson, P. J., Doyle, L. W. & the Victorian Infant Collaborative Study Group. (2010) 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*; 95: 786-90.

Robins, D. L., Fein, D., Barton, M. L. & Green, J. A. (2001) The Modified Checklist for Autism in Toddlers: an initial study investigating the early detection of autism and pervasive developmental disorders. *J Autism Dev Disord*; 31: 131-44.

Rose, S. A., Feldman, J. F. & Jankowski, J. J. (2009) A cognitive approach to the development of early language. *Child Dev*; 80: 134-50.

Royal College of Paediatrics and Child Health, Royal College of Ophthalmologists, British Association of Perinatal Medicine & Bliss 2008. UK Retinopathy of Prematurity Guideline May 2008. London.

Rutter, C. M. & Gatsonis, C. A. (2001) A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat Med*; 20: 2865-84.

Rutter, M., Bailey, A. & Lord, C. (2003) Social communication questionnaire. Los Angeles, CA: Western Psychological Services.

Salt, A., D'Amore, A., Ahluwalia, J., Seward, A., Kaptoge, S., Halliday, S. & Dorling, J. (2006) Outcome at 2 years for very low birthweight infants in a geographical population: risk factors, cost, and impact of congenital anomalies. *Early Hum Dev*; 82: 125-33.

Sand, N., Silverstein, M., Glascoe, F. P., Gupta, V. B., Tonniges, T. P. & O'Connor, K. G. (2005) Pediatricians' reported practices regarding developmental screening: do guidelines work? Do they help? *Pediatrics*; 116: 174-9.

Sansavini, A., Guarini, A., Alessandrini, R., Faldella, G., Giovanelli, G. & Salvioli, G. (2007) Are early grammatical and phonological working memory abilities affected by preterm birth? *J Commun Disord*; 40: 239-56.

Schenger, N. & Gentleman, J. F. (2001) On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician*; 55: 182-6.

Schiariti, V., Matsuba, C., Hoube, J. S. & Synnes, A. R. (2008) Severe retinopathy of prematurity and visual outcomes in British Columbia: a 10-year analysis. *J Perinatol*; 28: 566-72.

Schlapbach, L. J., Adams, M., Proietti, E., Aebischer, M., Grunt, S., Borradori-Tolsa, C., Bickle-Graz, M., Bucher, H. U., Latal, B. & Natalucci, G. (2012) Outcome at two years of age in a Swiss national cohort of extremely preterm infants born between 2000 and 2008. *BMC Pediatr*; 12: 198.

Schonhaut, L., Armijo, I., Schonstedt, M., Alvarez, J. & Cordero, M. (2013) Validity of the ages and stages questionnaires in term and preterm infants. *Pediatrics*; 131: e1468-74.

Schopler, E., Van Bourgondien, M. E., Wellman, G. J. & Love, S. R. (2010) *Childhood Autism Rating Scale*, Torrance, CA: WPS.

SCPE. (2000) Surveillance of cerebral palsy in Europe: a collaboration of cerebral palsy surveys and registers. Surveillance of Cerebral Palsy in Europe (SCPE). *Dev Med Child Neurol*; 42: 816-24.

Seigel, B. (2004) The Pervasive Developmental Disorders Screening Test II (PDDST-II). San Antonio, TX: Harcourt Assessment.

Serenius, F., Källén, K., Blennow, M., Ewald, U., Fellman, V., Holmström, G., Lindberg, E., Lundqvist, P., Maršál, K., Norman, M., Olhager, E., Stigson, L., Stjernqvist, K., Vollmer, B., Strömberg, B. & Group, E. (2013) Neurodevelopmental outcome in extremely preterm infants at 2.5 years after active perinatal care in Sweden. *JAMA*; 309: 1810-20.

Shribman, S. & Billingham, K. 2009. *Healthy Child Programme - pregnancy and the first five years* [Online]. London: Department of Health. Available: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/167998/Health_Child_Programme.pdf [Accessed 7th November 2014].

Skarsgard, E. D. (2006) Networks in Canadian paediatric surgery: Time to get connected. *Paediatr Child Health*; 11: 15-8.

Skellern, C. Y., Rogers, Y. & O'Callaghan, M. J. (2001) A parent-completed developmental questionnaire: follow up of ex-premature infants. *J Paediatr Child Health*; 37: 125-9.

Skranes, J., Vik, T., Nilsen, G., Smevik, O., Andersson, H. W. & Brubakk, A. M. (1998) Can cerebral MRI at age 1 year predict motor and intellectual outcomes in very-low-birthweight children? *Dev Med Child Neurol*; 40: 256-62.

Smith, K. E., Landry, S. H. & Swank, P. R. (2006) The role of early maternal responsiveness in supporting school-aged cognitive development for children who vary in birth status. *Pediatrics*; 117: 1608-17.

Smith, R. D. (1978) The use of developmental screening tests by primary-care pediatricians. *J Pediatr*; 93: 524-7.

Sondaar, M., B.J.M., v. K., de Klein, M. J. K., Briet, J. M., den Ouden, A. L. & van Baar, A. (2008) Do pediatricians recognize cognitive developmental problems in preterm children at age 5 years? *J Dev Phys Disabil*; 20: 21-9.

Song, F., Khan, K. S., Dinnes, J. & Sutton, A. J. (2002) Asymmetric funnel plots and publication bias in meta-analyses of diagnostic accuracy. *Int J Epidemiol*; 31: 88-95.

Stephens, B. E., Bann, C. M., Watson, V. E., Sheinkopf, S. J., Peralta-Carcelen, M., Bodnar, A., Yolton, K., Goldstein, R. F., Dusick, A. M., Wilson-Costello, D. E., Acarregui, M. J., Pappas, A., Adams-Chapman, I., McGowan, E. C., Heyne, R. J., Hintz, S. R., Ehrenkranz, R. A., Fuller, J., Das, A., Higgins, R. D. & Vohr, B. R. (2012) Screening for autism spectrum disorders in extremely preterm infants. *J Dev Behav Pediatr*; 33: 535-41.

Stoelhorst, G. M., Rijken, M., Martens, S. E., van Zwieten, P. H., Feenstra, J., Zwinderman, A. H., Wit, J. M. & Veen, S. (2003) Developmental outcome at 18 and 24 months of age in very preterm children: a cohort study from 1996 to 1997. *Early Hum Dev*; 72: 83-95.

Sullivan, M., Finelli, J., Marvin, A., Garrett-Mayer, E., Bauman, M. & Landa, R. (2007) Response to joint attention in toddlers at risk for autism spectrum disorder: a prospective study. *J Autism Dev Disord*; 37: 37-48.

Swinkels, S. H., Dietz, C., van Daalen, E., Kerkhof, I. H., van Engeland, H. & Buitelaar, J. K. (2006) Screening for autistic spectrum in children aged 14 to 15 months. I: the development of the Early Screening of Autistic Traits Questionnaire (ESAT). *J Autism Dev Disord*; 36: 723-32.

Synnes, A. R., Anson, S., Baum, J. & Usher, L. (2012) Incidence and pattern of hearing impairment in children with ≤ 800 g birthweight in British Columbia, Canada. *Acta Paediatr*; 101: e48-54.

The EIPAGE Study [Online]. Available: <http://www.perinat-france.org>.

The National Audit Office study team 2007. Caring for vulnerable babies: the reorganisation of neonatal services in England. London: The Stationery Office.

The Victorian Infant Collaborative Study [Online]. Available: <http://www.vics-infantstudy.org.au>.

The Victorian Infant Collaborative Study Group. (1997) Improved outcome into the 1990s for infants weighing 500-999 g at birth. The Victorian Infant Collaborative Study Group. *Arch Dis Child Fetal Neonatal Ed*; 77: F91-4.

Tin, W., Fritz, S., Wariyar, U. & Hey, E. (1998) Outcome of very preterm birth: children reviewed with ease at 2 years differ from those followed up with difficulty. *Arch Dis Child Fetal Neonatal Ed*; 79: F83-7.

Tommiska, V., Heinonen, K., Kero, P., Pokela, M. L., Tammela, O., Järvenpää, A. L., Salokorpi, T., Virtanen, M. & Fellman, V. (2003) A national two year follow up study of extremely low birthweight infants born in 1996-1997. *Arch Dis Child Fetal Neonatal Ed*; 88: F29-35.

Tversky, A. & Kahneman, D. (1974) Judgment under Uncertainty: Heuristics and Biases. *Science*; 185: 1124-31.

UK National Screening Committee. (2009) The UK NSC policy on autism screening in children [Online]. Available: <http://www.screening.nhs.uk/autism>.

van Noort-van der Spek, I. L., Franken, M. C. & Weisglas-Kuperus, N. (2012) Language functions in preterm-born children: a systematic review and meta-analysis. *Pediatrics*; 129: 745-54.

Veelken, N., Stollhoff, K. & Claussen, M. (1991) Development of very low birth weight infants: a regional study of 371 survivors. *Eur J Pediatr*; 150: 815-20.

Veen, S., Sassen, M. L., Schreuder, A. M., Ens-Dokkum, M. H., Verloove-Vanhorick, S. P., Brand, R., Grote, J. J. & Ruys, J. H. (1993) Hearing loss in very preterm and very low birthweight infants at the age of 5 years in a nationwide cohort. *Int J Pediatr Otorhinolaryngol*; 26: 11-28.

Vohr, B. (2014) Speech and language outcomes of very preterm infants. *Semin Fetal Neonatal Med*; 19: 78-83.

Vohr, B. R., Msall, M. E., Wilson, D., Wright, L. L., McDonald, S. & Poole, W. K. (2005) Spectrum of gross motor function in extremely low birth weight children with cerebral palsy at 18 months of age. *Pediatrics*; 116: 123-9.

Vohr, B. R., Stephens, B. E., Higgins, R. D., Bann, C. M., Hintz, S. R., Das, A., Newman, J. E., Peralta-Carcelen, M., Yolton, K., Dusick, A. M., Evans, P. W., Goldstein, R. F., Ehrenkranz, R. A., Pappas, A., Adams-Chapman, I., Wilson-Costello, D. E., Bauer, C. R., Bodnar, A., Heyne, R. J., Vaucher, Y. E., Dillard, R. G., Acarregui, M. J., McGowan, E. C., Myers, G. J. & Fuller, J. (2012) Are outcomes of extremely preterm infants improving? Impact of Bayley assessment on outcomes. *J Pediatr*; 161: 222-8.

Vohr, B. R., Wright, L. L., Dusick, A. M., Perritt, R., Poole, W. K., Tyson, J. E., Steichen, J. J., Bauer, C. R., Wilson-Costello, D. E. & Mayes, L. C. (2004) Center differences and outcomes of extremely low birth weight infants. *Pediatrics*; 113: 781-9.

Walch, E., Chaudhary, T., Herold, B. & Obladen, M. (2009) Parental bilingualism is associated with slower cognitive development in very low birth weight infants. *Early Hum Dev*; 85: 449-54.

Watkinson, M., Davis, K. & Neonatal Data Analysis Unit. (2009) National neonatal audit programme annual report 2009. London: Royal College of Paediatrics & Child Health.

Wechsler, D. (2002) Wechsler preschool and primary scale of intelligence for children - third edition. San Antonio, TX: The Psychological Corporation.

Wetherby, A. & Prizant, B. (2002) Communication and Symbolic Behavior Scales Developmental Profile, Baltimore, MD: Paul H Brookes Publishing.

Wetherby, A. M., Brosnan-Maddox, S., Peace, V. & Newton, L. (2008) Validation of the Infant-Toddler Checklist as a broadband screener for autism spectrum disorders from 9 to 24 months of age. *Autism*; 12: 487-511.

Whiting, P., Harbord, R. & Kleijnen, J. (2005) No role for quality scores in systematic reviews of diagnostic accuracy studies. *BMC Med Res Methodol*; 5: 19.

Whiting, P. F., Rutjes, A. W., Westwood, M. E., Mallett, S., Deeks, J. J., Reitsma, J. B., Leeflang, M. M., Sterne, J. A. & Bossuyt, P. M. (2011) QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*; 155: 529-36.

Williams, J. G., Higgins, J. P. & Brayne, C. E. (2006) Systematic review of prevalence studies of autism spectrum disorders. *Arch Dis Child*; 91: 8-15.

Wilson-Ching, M., Molloy, C. S., Anderson, V. A., Burnett, A., Roberts, G., Cheong, J. L., Doyle, L. W. & Anderson, P. J. (2013) Attention difficulties in a contemporary geographic cohort of adolescents born extremely preterm/extremely low birth weight. *J Int Neuropsychol Soc*; 19: 1097-108.

Wilson-Costello, D., Friedman, H., Minich, N., Siner, B., Taylor, G., Schluchter, M. & Hack, M. (2007) Improved neurodevelopmental outcomes for extremely low birth weight infants in 2000-2002. *Pediatrics*; 119: 37-45.

Wolke, D. & Meyer, R. (1999) 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*; 41: 94-109.

Wolke, D., Ratschinski, G., Ohrt, B. & Riegel, K. (1994) The cognitive outcome of very preterm infants may be poorer than often reported: an empirical investigation of how methodological issues make a big difference. *Eur J Pediatr*; 153: 906-15.

Wolke, D., Sohne, B., Ohrt, B. & Riegel, K. (1995) Follow-up of preterm children: important to document dropouts. *Lancet*; 345: 447.

Wood, E. & Rosenbaum, P. (2000a) The gross motor function classification system for cerebral palsy: a study of reliability and stability over time. *Dev Med Child Neurol*; 42: 292-6.

Wood, N. S., Marlow, N., Costeloe, K., Gibson, A. T. & Wilkinson, A. R. (2000b) Neurologic and developmental disability after extremely preterm birth. EPICure Study Group. *N Engl J Med*; 343: 378-84.

World Health Organization. (2007) International Classification of Functioning, Disability, and Health : Children & Youth version. Geneva: World Health Organization.

Zimmerman, I. L., Steiner, V. G. & Pond, R. E. (2002) Preschool language scale - fourth edition. San Antonio, TX: The Psychological Corporation.

Appendix 1: Research outputs

Related publications

Wong HS, Huertas-Ceballos A, Cowan FM, Modi N on behalf of the Medicines for Neonates Investigator Group. Evaluation of early childhood social-communication difficulties in children born preterm using the Quantitative Checklist of Autism in Toddlers. *J Pediatr* 2014; 164(1):26-33. [PMID: 23972644]

Other publications

Wong HS, Santhakumaran S, Statnikov Y, Gray D, Watkinson M, Modi N, the UK Neonatal Collaborative. Retinopathy of prematurity in English neonatal units: a national population-based analysis using NHS operational data. *Arch Dis Child Fetal Neonatal Ed* 2014; 99(3):F196-202 [PMID: 24361602]

Wong HS, Edwards P. Nature or nurture: a systematic review of the effect of socio-economic status on the developmental and cognitive outcomes of children born preterm. *Matern Child Health J* 2013; 17(9):1689-700. [PMID: 23135625]

Presentations

Predictive validity of early developmental assessments in identifying school-age cognitive deficits in children born preterm or very low birthweight: systematic review and meta-analysis

Wong HS, Santhakumaran S, Cowan FM, Modi N on behalf of the Medicines for Neonates Investigator Group

Oral presentation – Fifth Congress of the European Academy of Paediatric Societies, Barcelona (17-21 October 2014)

Sociodemographic and neonatal factors associated with early childhood social-communication difficulties in children born preterm

Wong HS, Huertas-Ceballos A, Cowan FM, Modi N on behalf of the Medicines for Neonates Investigator Group

Poster – Fourth Congress of the European Academy of Paediatric Societies, Istanbul (5-9 October 2012)

Oral presentation – Neonatal Society Summer Meeting, Canterbury (21-22 June 2012)

Evaluation of early childhood social-communication difficulties in children born preterm using the Quantitative Checklist for Autism in Toddlers

Wong HS, Huertas-Ceballos A, Cowan FM, Modi N on behalf of the Medicines for Neonates Investigator Group

Oral presentation (award best presentation by a trainee) – Neonatal Society Spring Meeting, London (28 March 2012)

Comparison of two parent-completed questionnaires for the identification of children at risk for autism spectrum disorder in the preterm population

Wong HS, Huertas-Ceballos A, Cowan FM, Modi N

Poster – European Society for Paediatric Research, Newcastle (14-17 October 2011)

Peer-review activities

I have been invited by the Archives of Disease in Childhood to peer-review two original papers submitted to the journal and have completed both reviews.

Evaluation of Early Childhood Social-Communication Difficulties in Children Born Preterm Using the Quantitative Checklist for Autism in Toddlers

Hilary S. Wong, MRCPCH, MSc¹, Angela Huertas-Ceballos, MSc, FRCPC², Frances M. Cowan, PhD, FRCPC¹, and Neena Modi, MD, FRCPC¹, on behalf of the Medicines for Neonates Investigator Group*

Objectives To characterize early childhood social-communication skills and autistic traits in children born very preterm using the Quantitative Checklist for Autism in Toddlers (Q-CHAT) and explore neonatal and sociodemographic factors associated with Q-CHAT scores.

Study design Parents of children born before 30 weeks gestation and enrolled in a study evaluating routinely collected neurodevelopmental data between the post-menstrual ages of 20 and 28 months were invited to complete the Q-CHAT questionnaire. Children with severe neurosensory disabilities and cerebral palsy were excluded. Participants received neurodevelopmental assessments using the *Bayley Scales of Infant and Toddler Development, 3rd edition* (Bayley-III). Q-CHAT scores of this preterm cohort were compared with published general population scores. The association between Bayley-III cognitive and language scores and neonatal and sociodemographic factors with Q-CHAT scores were examined.

Results Q-CHAT questionnaires were completed from 141 participants. At a mean post-menstrual age of 24 months, the Q-CHAT scores of the preterm cohort (mean 33.7, SD 8.3) were significantly higher than published general population scores (mean 26.7; SD 7.8), indicating greater social-communication difficulty and autistic behavior. Preterm children received higher scores, particularly in the categories of restricted, repetitive, stereotyped behavior, communication, and sensory abnormalities. Lower Bayley-III language scores and non-white ethnicity were associated with higher Q-CHAT scores.

Conclusions Preterm children display greater social-communication difficulty and autistic behavior than the general population in early childhood as assessed by the Q-CHAT. The implications for longer-term outcome will be important to assess. (*J Pediatr* 2014;164:26-33).

See editorial, p 6 and related article, p 20

The risk for autism spectrum disorders (ASD), which are characterized by impairments in communication, reciprocal socialization, and repetitive behavior,^{1,2} is significantly higher among children born preterm compared with their term-born counterparts.^{3,4} The estimated prevalence of ASD has been reported to be 5% in children born weighing less than 2000 g⁴ and 8% in children born at <26 weeks gestation.³ This represents an approximate 10-fold increase over the 2-9 per 1000 prevalence estimate in the general population.^{5,6} Reported risk factors common among preterm children that are associated with autism include multiple birth, small for gestational age, and birth by cesarean delivery.^{7,8}

Since 2007, the American Academy of Pediatrics has recommended universal ASD-specific screening of all children at 18 to 24 months of age.⁹ However, the pattern of early development of social-communication skills and autistic traits among children born preterm and, hence, the applicability of the recommended screening tools, is unknown. Several authors have studied the use of the Modified Checklist for Autism in Toddlers (M-CHAT) in the high-risk preterm population. The M-CHAT consists of binary items that dichotomize children into 'high risk' or 'low risk' for autism. Using the M-CHAT, high positive screening rates of 25% in very low birth weight (<1500 g) infants¹⁰ and 21%-41% in infants born before 28 weeks gestation^{11,12}

ASD	Autism spectrum disorders
Bayley-III	<i>Bayley Scales of Infant and Toddler Development, 3rd edition</i>
IMD	Index of multiple deprivation
M-CHAT	Modified Checklist for Autism in Toddlers
NNRD	National Neonatal Research Database
Q-CHAT	Quantitative Checklist for Autism in Toddlers

From the ¹Section of Neonatal Medicine, Department of Medicine, Chelsea and Westminster Hospital, Imperial College London and ²Neonatal Service, University College London Hospital NHS Foundation Trust, London, United Kingdom

*A list of members of the Medicines for Neonates Investigator Group is available at www.jpeds.com (Appendix).

Funded by the National Institute for Health Research (NIHR) (Program grant for Applied Research RP-PG-0707-10010), and sponsored by Imperial College London. The views expressed are those of the authors and not necessarily those of the National Health Service, the NIHR, or the Department of Health. The authors declare no conflicts of interest.

Portions of this study were presented as a poster during the 4th Congress of the European Academy of Paediatric Societies, October 5-9, 2012, Istanbul, Turkey.

0022-3476/\$ - see front matter. Copyright © 2014 Mosby Inc. All rights reserved. <http://dx.doi.org/10.1016/j.jpeds.2013.07.013>

were found. Although an increased rate of positive screening for autism might be expected among children born preterm, it has also become apparent that the M-CHAT is poor at differentiating autistic symptoms from neurosensory, cognitive, and motor impairments and that the specificity of screening for ASD in the preterm population is confounded by the high prevalence of these coexisting morbidities.¹¹⁻¹³

The Quantitative Checklist for Autism in Toddlers (Q-CHAT) is a recent revision of the M-CHAT.¹⁴ It is a parent-completed questionnaire consisting of updated items, with each item having a 5-point rating scale instead of a binary scoring system. Assessments of the test properties and clinical validity of the Q-CHAT are ongoing. In a preliminary report, the authors showed that the Q-CHAT scores from an unselected group of 754 toddlers aged between 17 and 26 months (mean age 21.2 months), living in Cambridgeshire, United Kingdom, followed a near-normal distribution and were significantly lower than the scores of children with ASD.¹⁴ The quantitative nature of the Q-CHAT and the near-normal distribution of its scores make it a useful tool as an autistic trait measure. Therefore, we aimed to use the Q-CHAT at 24 months post-menstrual age (age corrected for prematurity) to characterize the social-communication skills and autistic traits in children born before 30 weeks gestation. We compared the Q-CHAT scores of the preterm cohort with the reported scores of the general population, examined correlations with the cognitive and language scores from the *Bayley Scales of Infant and Toddler Development, 3rd edition* (Bayley-III),¹⁵ and investigated associated neonatal and sociodemographic factors.

Methods

Between June 2010 and July 2012, children born before 30 weeks gestation were invited to enroll in a larger study evaluating the reliability of routinely collected neurodevelopmental data at 2 years post-menstrual age in one of 13 participating hospitals (the United Kingdom Clinical Research Network Study Portfolio ID 8626). The 13 study sites included hospitals within the North-East and North-West London Neonatal Networks, St. Thomas' Hospital in South-East London, and Addenbrooke's Hospital in Cambridge, United Kingdom. The study received approval from the Royal Free Hospital National Research Ethics Committee (10/H0720/35). Children from non-English speaking families whose parents required interpretation of the English language were excluded for the following reasons: (1) no provision for English translation was available to ensure that informed consent was obtained; (2) it was not possible to conduct the Bayley-III neurodevelopmental assessment reliably; and (3) the parents would not be able to complete the Q-CHAT questionnaire independently.

In addition to their routine follow-up assessment, all participants had a Bayley-III assessment and a standardized neurologic examination, based on the Hammersmith Infant Neurological Examination,¹⁶ conducted by a single trained

assessor (H.W.) who was blinded to the past medical history of the child. The Bayley-III includes assessments of cognitive and language domains. For each domain, a composite score standardized to a mean 100 and SD 15 was calculated using the participant's post-menstrual age at the time of assessment. The language composite score was converted from combining subscale scores, which were standardized to a mean 10 and SD 3, from the receptive communication and expressive communication subtests.

Q-CHAT

The parents of all participants were sent the Q-CHAT questionnaire prior to the appointment for the neurodevelopmental assessment. If the questionnaire was not completed by the time of the appointment, the parents were given a further copy of the questionnaire and asked to return it by mail. Parents were asked to complete the questionnaire independently at home rather than at the appointment in order to minimize information bias that may result from subjective influence from the researcher.

To reduce the potential confounding effect of coexisting neurosensory and physical impairments on the association between preterm birth and Q-CHAT scores, for the purpose of assessing the applicability of the Q-CHAT for the majority of children born preterm, we excluded children with cerebral palsy and severe neurosensory impairments (defined as a hearing deficit not correctable with hearing aids or a visual deficit not correctable with glasses).

All returned questionnaires were scored according to the methods described by the research team who developed the Q-CHAT.¹⁴ Each of the 25 items was scored using a 5-point Likert scale (0-4 points) with higher scores indicating a higher frequency of autistic behavior. Responses that were ambiguous or incomplete were scored 0, in accordance with the conservative approach adopted by the developers of the Q-CHAT. Questionnaires with more than 6 incomplete responses were excluded. The scores from all items were summed to obtain a total Q-CHAT score within a possible range of 0-100.

To examine different aspects of autistic behavior, we classified the Q-CHAT items into categories that explored social-relatedness (9 items), restricted, repetitive, stereotyped behavior (9 items), communication (4 items), and sensory abnormalities (3 items), based on the nature of the questions.

Neonatal and Sociodemographic Data Collection

The following neonatal and sociodemographic factors were identified a priori for analysis as possible variables associated with Q-CHAT scores: birth gestation, birth weight z-score, sex, single vs multiple pregnancy, ethnicity (white/non-white), maternal age, mode of delivery, length of mechanical ventilation, supplemental oxygen requirement at 36 weeks post-menstrual age, received breast milk during neonatal unit stay, and index of multiple deprivation (IMD). Multiple deprivation relates to the concurrent occurrence of several forms of social and economic disadvantage. The IMD is a summary measure of relative area deprivation, calculated

through a weighted combination of scores in 38 different indicators covering factors such as income, employment, education, health, living environment, and crime for each area in England, using national census data.¹⁷ The IMD for each participant was obtained based on the area code the child was living in at the time of assessment and according to the English Indices of Deprivation 2010.¹⁷ For the purpose of this analysis, IMD was categorized into quintiles based on ranking, with IMD Quintile One presenting the least deprived 20% of areas in England.

Data on the neonatal variables were extracted from the United Kingdom National Neonatal Research Database (NNRD). In England, all neonatal units now employ electronic databases to prospectively collect and document standardized routine patient clinical records. Since 2007, the anonymized clinical records of all infants admitted to 165 participating neonatal units have been used to create the NNRD at the Neonatal Data Analysis Unit. The NNRD is a national resource to facilitate neonatal research and support clinical services in the United Kingdom. We extracted data for each participant by linkage with the NNRD through the participants' unique identifiers.

Statistical Analyses

Data were double-entered and verified to ensure accuracy. All analyses were performed using the Stata statistical package v. 11.0 (StataCorp, College Station, Texas). Group differences in characteristics between respondents and non-respondents were compared using χ^2 tests for categorical variables and Student *t* tests or Mann-Whitney U tests for continuous variables. We used the Student *t* test to compare the overall and sex-specific Q-CHAT scores from the study preterm population with the published scores from the general population (general population overall mean 26.7, SD 7.8; mean for boys 27.5, SD 7.8; mean for girls 25.8, SD 7.7).¹⁴ Differences in the distributions of item-specific scores between the study cohort and the general population in each category of autistic behavior were examined by χ^2 tests. To overcome the χ^2 test restriction for low expected numbers, we combined the proportions in adjacent score categories to ensure that all expected values were larger than five.¹⁸

As cognitive and communication abilities are known to be associated with ASD, we explored the correlation between the Q-CHAT scores and the Bayley-III cognitive and language scores using linear regression to determine if any observed difference in Q-CHAT scores between the preterm and the general populations were explained by delayed cognitive and language development in the preterm population. Post-hoc analysis of the correlation between subcategorical Q-CHAT scores (total score from items within each category of autistic behavior) and Bayley-III cognitive and language scores was carried out with Bonferroni correction for multiple testing.

Linear regression models were created to determine the association of predictive variables with Q-CHAT scores. To account for correlated outcomes within multiple birth

sets, we used cluster bootstrap to estimate standard errors. Variables identified to be significant at 5% significance level in univariable models were included in forward stepwise multivariable regression analyses to determine the independent effect of each factor on Q-CHAT scores. We also conducted post-hoc analyses to explore possible interactions between ethnicity, Bayley-III language scores, and IMD.

Results

Q-CHAT questionnaires were sent to all 211 participants in the larger study. Ten children were found to have major functional impairments (9 with cerebral palsy; 1 with severe hearing impairment) and were ineligible for this study. One hundred fifty questionnaires, including 8 from children who were ineligible, were returned. One questionnaire with 7 missing responses was treated as a non-respondent and excluded, leaving 141 participants (70.1% of eligible participants) for the analyses.

Table 1 (available at www.jpeds.com) compares the neonatal and sociodemographic characteristics between respondents and non-respondents. Non-respondents were more likely to be parents of girls (66.7%, $P = .02$). Nonetheless, both boys and girls were equally represented in the respondent group. All children had received breast milk during their neonatal unit stay. Lower proportions of non-white participants and participants living in the most deprived IMD quintile responded to the questionnaire although the differences in proportions between respondents and non-respondents were not statistically significant. Nevertheless, our sample consisted of an overrepresentation of children living in more deprived areas. A χ^2 analysis conducted to compare the IMD between our study cohort and that of all infants born before 30 weeks gestation and admitted to our participating hospitals between 2008 and 2010, using information in the NNRD, showed no difference in the 2 samples. Hence, the socioeconomic status of our study sample was reflective of the target preterm population from which it was drawn. The mean post-menstrual age of the respondents was 24.7 (SD 2.6, range 18.5-35.6) months at the time of completion of the questionnaire.

Q-CHAT Scores of Preterm Population

The Q-CHAT scores of the preterm population (mean 33.7, SD 8.3, range 15-55; **Figure**) were normally distributed and significantly higher (less favorable) than the published general population scores (mean difference 7.0 (95% CI 5.6-8.3); $P < .001$). The mean Q-CHAT scores were 33.8 (SD 7.8, range 15-55) for preterm boys and 33.5 (SD 8.8, range 15-54) for preterm girls. Compared with the general population, sex-specific scores in both preterm boys and girls were significantly higher (mean difference 6.3 (95% CI 4.5-8.1) for boys, 7.7 (95% CI 5.6-9.8) for girls; $P < .001$ for both sexes). In contrast to the higher scores described in boys in the general population,¹⁴ no sex difference in Q-CHAT scores were observed in our preterm population

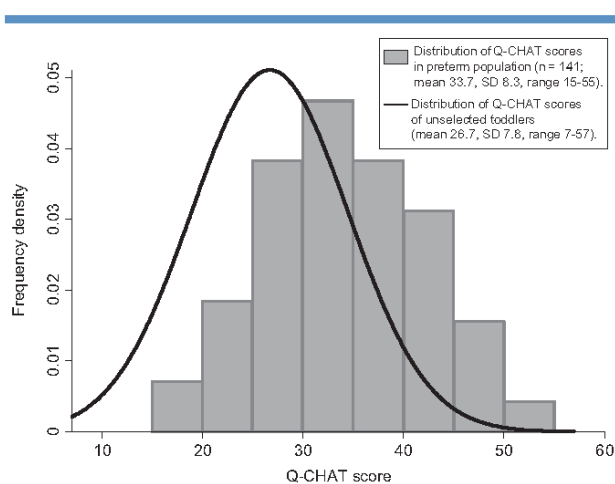


Figure. Histogram of Q-CHAT scores of preterm population with superimposed distribution of published Q-CHAT scores of unselected toddlers.

($P = .85$). Overall, 23 (16.3%) participants (10 boys and 13 girls) had Q-CHAT scores higher than 2 SD above the general population mean (ie, higher than 42.3). When we examined this using the sex-specific scores of the general population, 8 (11.0%) boys and 13 (19.1%) girls obtained Q-CHAT scores higher than 2 SD deviations above the sex-specific means of the general population (ie, higher than 43.1 for boys; 41.2 for girls). There was no correlation between Q-CHAT scores and the corrected ($P = .21$) and uncorrected ($P = .36$) ages at assessments. There was no difference in the mean Q-CHAT scores between respondents whose parents had completed the questionnaire before and those who had completed the questionnaire after their observation of the Bayley-III assessment ($P = .84$).

The distribution of item-specific scores by category of autistic behavior is displayed in [Table II](#). The distribution of scores between the preterm study cohort and the general population differed significantly in 17 items. In all these items, there were greater proportions of preterm children receiving higher scores, indicating greater social-communication difficulties and autistic behavior characteristics. The differences were most predominant in the categories of restricted, repetitive, stereotyped behavior (7 of 9 items differ significantly), communication (3 of 4 items), and sensory abnormalities (all 3 items). Only 4 out of the 9 items exploring social-relatedness were scored differently in the preterm population.

Association of Q-CHAT Scores with Cognitive and Language Abilities at 2 Years

With univariable analyses, Q-CHAT scores were significantly associated with both Bayley-III cognitive and language composite scores. When both variables were input into a multivariable regression model, the effect of cognitive function on Q-CHAT scores was no longer significant ($P = .24$). Bayley-III language composite scores indepen-

dently predicted Q-CHAT scores in a linear fashion (regression coefficient: -0.25 ; 95% CI -0.40 to -0.10 ; correlation coefficient -0.51 ; $P = .001$) and accounted for 24.3% of the variance in Q-CHAT scores. The observed relationship between language and Q-CHAT scores was entirely due to expressive communication ability. The regression coefficient between Bayley-III expressive communication subscale scores and Q-CHAT scores was -1.35 (95% CI -1.96 to -0.74 ; correlation coefficient -0.43 ; $P < .001$). There was no association between receptive communication ability and Q-CHAT scores ($P = .22$). Apart from the expected association between Bayley-III language scores and subcategorical Q-CHAT scores from items exploring communication skills, no significant association was found between cognitive and language scores and other subcategorical Q-CHAT scores (data not shown). In a separate analysis, Bayley-III motor composite scores were associated with Q-CHAT scores in univariable analysis but did not have an independent effect when included in a multivariable model with language scores ($P = .95$).

Neonatal and Sociodemographic Predictors of Q-CHAT Scores

Successful linkage with the NNRD was achieved for 139 of 141 respondents and complete data of all predictor variables were obtained for 126 (89.4%) children. All variables examined and their test statistics are listed in [Table III](#). Non-white ethnicity and living in deprived areas were found to be associated with higher Q-CHAT scores following univariable analysis. Although non-white children were more likely to live in areas of higher deprivation (test for trend $P < .001$, data not shown), there was no interaction between ethnicity and IMD ($P = .72$) in the association with Q-CHAT scores. As we observed lower Bayley-III language scores among non-white children (mean difference 7.31 [95% CI 3.07-11.5; $P < .001$]) and children living in more deprived areas (mean decrease of 1.89 [95% CI 0.24-3.55; $P = .03$] points per IMD quintile increase in deprivation), we considered language ability to be a potential confounder in the relationship between ethnicity, IMD, and Q-CHAT scores. There was no interaction between Bayley-III language scores and IMD quintiles ($P = .88$) or ethnicity ($P = .51$). The final multivariable regression model included all variables found to be statistically significant during univariable analysis (Bayley-III language composite score, ethnicity, and IMD) ([Table IV](#)).

Discussion

In this study, we demonstrated that at 24 months post-menstrual age, children born before 30 weeks gestation were rated by their parents as having greater social-communication difficulties and autistic traits compared with the general population. To our knowledge, this is the first study characterizing the distribution and spread of autistic traits of preterm children in early childhood. Utilizing properties of

Table II. Item-specific distribution of Q-CHAT scores

Q-CHAT item	Score (% of responses)					Difference in distribution compared with general population (<i>P</i>)
	0	1	2	3	4	
Items exploring social-relatedness						
1. Look when name is called*	49.6	44.7	5.0	0.0	0.7	.13
2. Eye contact*	52.5	44.0	2.8	0.7	0.0	.004
3. Protoimperative pointing*	68.8	25.5	3.5	0.7	1.4	.58
4. Protodeclarative pointing*	61.7	24.1	11.3	0.7	2.1	.56
5. Pretend play*	64.5	22.0	7.8	2.1	3.5	.02
6. Follow a gaze*	53.6	37.1	7.1	0.7	1.4	.45
7. Offer comfort	34.8	32.6	24.6	4.3	3.6	.04
8. Use simple gestures*	76.6	18.4	3.5	0.7	0.7	.28
9. Check reaction	33.3	32.6	22.0	9.9	2.1	<.001
Restricted, repetitive, stereotyped behavior						
10. Line objects up†	7.1	10.6	40.4	24.1	17.7	<.001
11. Interest maintained by spinning object†	16.8	43.1	26.3	6.6	7.3	<.001
12. Adapt to change in routine*	23.4	58.2	17.0	1.4	0.0	<.001
13. Do the same thing over and over again	8.5	7.8	13.5	23.4	46.8	<.001
14. Echolalia	2.8	3.5	19.1	25.5	48.9	.06
15. Unusual finger movement†	53.2	10.1	12.9	13.7	10.1	<.001
16. Maintenance of interest†	23.9	24.6	30.4	13.0	8.0	<.001
17. Twiddle objects repetitively†	33.1	14.4	17.3	24.5	10.8	<.001
18. Stare at nothing with no purpose*	59.6	22.7	9.2	5.0	3.5	.15
Communication abnormalities						
19. Understand child's speech	18.4	36.9	31.2	12.1	1.4	<.001
20. Number of words	17.9	17.9	41.4	19.3	3.6	.69
21. Typicality of first words*	56.2	34.3	4.4	1.5	3.6	.02
22. Use of hand as tool	7.8	5.0	10.6	36.2	40.4	<.001
Sensory abnormalities						
23. Sniff or lick unusual objects	15.9	14.5	20.3	29.0	20.3	<.001
24. Walk on tiptoe†	13.5	22.0	46.8	9.9	7.8	<.001
25. Oversensitive to noise†	23.4	36.9	21.3	9.2	9.2	<.001

* χ^2 test was performed by combining proportions with scores 2, 3, and 4.
 † χ^2 test was performed by combining proportions with scores 3 and 4.

the Q-CHAT as quantitative measures of autistic features, we showed that Q-CHAT scores in the preterm population, although following a normal distribution with similar variability as the general population, were shifted to the right, yielding higher, less favorable scores. These findings corroborate the report by Johnson et al of a similar right-shift in frequency distribution of ASD symptoms in preterm children compared with term-born classmate controls as measured by the Social Communication Questionnaire at age 11 years.³ Our result, thus, suggests that in the preterm population, suboptimal development of social-communication skills exists from early childhood. The 7-point right shift in mean Q-CHAT score of the preterm population corresponds to nearly a 1 SD difference. Because ASD exists as a continuum with autism representing the extreme end of the spectrum, our results

also support the likelihood that a large proportion of preterm children experience clinically significant social-communication difficulties below the diagnostic threshold for ASD from a young age, when early intervention may be possible.

Our study sample was comprised of children without major functional disability. Previous studies have reported significantly higher odds of positive autism screening on the M-CHAT in children with motor, visual, hearing, and cognitive impairments.^{11,12} Exclusion of children with these impairments reduced the positive screening rates at age 2 years from 21%-10% in the extremely low gestational age newborns (ELGAN) study¹¹ and from 41%-16.5% in the extremely premature babies (EPICure) study.¹² Many 'critical' items on the M-CHAT such as "Does your child

Table III. Univariable associations of neonatal and sociodemographic factors with Q-CHAT scores

Variable	n	Coefficient (Q-CHAT score)	95% CI	z-statistic	<i>P</i>
Gestation (per completed wk)	141	-0.77	-1.61-0.06	-1.82	.07
Birthweight z-score (per point increase)	126	0.07	-1.41-1.55	0.09	.09
Male sex	141	0.27	-2.54-3.07	0.19	.85
Singleton pregnancy	141	3.80	-0.42-8.01	1.76	.08
White ethnicity	141	-7.55	-10.2 to -4.86	-5.51	.001
Maternal age (per y)	141	-0.17	-0.40-0.07	-1.38	.17
Cesarean delivery	132	-2.22	-5.38-0.93	-1.38	.17
Length of mechanical ventilation (per d)	139	0.10	-0.01-0.20	1.84	.07
Supplemental oxygen requirement at 36 wk post-menstrual age	141	1.30	-2.06-4.67	0.76	.45
IMD quintile (per quintile increase in deprivation)	141	2.07	1.04-3.11	3.94	<.001

Table IV. Final multivariable model of factors associated with Q-CHAT scores

Variable	Coefficient (Q-CHAT score)	95% CI	z-statistic	P
Bayley-III language composite score (per point)	-0.23	-0.33 to -1.39	-4.82	<.001
White ethnicity	-5.30	-7.92 to -2.67	-3.96	<.001
IMD quintile (per quintile increase in deprivation)	0.96	-2.00-0.08	1.81	.07

Note: n = 136. $r^2 = 0.38$.

ever bring objects over to you to show you something?" and "Does your child respond to his/her name when you call?" depended on intact neurosensory and motor functions and failed more frequently by children with overt motor and other disabilities.¹³ The Q-CHAT also contains similar questions, and we would expect children with such disabilities to receive higher Q-CHAT scores on these questions. Therefore, it is likely that the distribution of Q-CHAT scores in this very preterm population would be even higher if children with cerebral palsy and severe neurosensory disabilities were included.

In the general population, there was a 1.7 point sex difference in mean scores, with the scores for boys being significantly higher than the scores for girls. We did not find a sex difference in our population. This may be due to insufficient statistical power, given that a sample size of 24 000 children would be required to detect the 0.3 point sex difference in Q-CHAT scores as significantly different. Nevertheless, it has been suggested that the autism risk phenotype seen in preterm children resembles more closely that of children with syndromic or medically-explained autism where the sex ratio is closer to 1:1 than those with idiopathic autism,¹¹ supporting the hypothesis that the autism risk phenotype in preterm children, rather than being a primary deficit, represents part of a 'preterm phenotype' with a different etiology.

In addition to comparing the overall measure, we conducted analyses on each questionnaire item and found that preterm children experience difficulties across all aspects of autistic behavior but particularly in the categories of restricted, repetitive, stereotyped behavior, communication, and sensory abnormalities. The presence of reduced language ability among children born preterm is well-described.^{19,20} Dysfunction in sensory modulation in preterm children, characterized by either hyposensitivity or hypersensitivity to sensory input, is a problem anecdotally recognized by parents and clinicians. There is, however, a paucity of studies in this area. It is hypothesized that illness-related inflammatory processes, which affect the development of brain cells, structures, and circuits, combined with the intense multisensory environment of the neonatal intensive care unit and life-sustaining interventions, result in alternations of the infant's sensory system and regulation.^{21,22} Sensory modulation dysfunction is thought to be negatively associated with emotional development and can affect social-interactive capabilities.²³

There is some evidence that restricted and repetitive behaviors are associated with cognitive status.^{24,25} EPICure study investigators also concluded that cognitive deficits

in their extremely preterm cohort accounted for the excess of repetitive and stereotyped behavior compared with term controls.²⁶ Although we did not demonstrate a correlation between cognitive scores and subcategorical Q-CHAT scores in the restricted and repetitive behavior domain, as the mean cognitive score of our preterm population was lower than would be expected in the general population, the potential association between cognition and restricted and repetitive behavior could in part explain the higher Q-CHAT scores obtained by the participants in this category.

We observed fewer differences between preterm children and the general population in response to items exploring social-relatedness. Previous work on early autism screening highlighted the absence of pretend play and joint-attention as strong predictors for later diagnosis of autism.²⁷ We speculate that Q-CHAT items exploring social-relatedness may provide a higher degree of specificity for differentiating early autistic features from concurrent developmental delay in children without severe physical and neurosensory impairment compared with items in the other categories. Although parents reported a lower frequency of pretend play among the preterm children, development in joint-attention (elucidated by questions on protodeclarative pointing and following a gaze) were similar to the general population. Focusing on elucidating social-relatedness for autism screening in the preterm population may reduce the 'false-positive' screening rate associated with currently available screening tools.

This study highlights the inter-relationship between ethnicity, area deprivation, language skills, and Q-CHAT scores. Despite the obvious limitations of this approach, as ethnicity is correlated with area deprivation, it is often used as a proxy measure for socioeconomic status. The relationships between Q-CHAT scores, ethnicity, and socioeconomic status in the general population are unknown. Our findings suggest the possibility of an environmental impact of socioeconomic disadvantage on early social-communication development. It could also represent ethnic and cultural differences, language abilities, and other socioeconomic influences on parental reporting on the Q-CHAT. We lack information about individual socioeconomic status, and the ecologic fallacy in using IMD must be noted. The reliability of the Q-CHAT questionnaire across diverse ethnic and socioeconomic groups will need to be further determined.

The methodological strengths of this study included prospective collection of both neonatal and 2-year data, which precluded recall bias, and the stringent administration

of the Bayley-III developmental assessment by a single assessor blinded to the medical history of the participants. In addition, our recruitment of participants, based on a gestational age rather than a birth weight criterion that was not limited to extreme prematurity, allows broader generalizability of the results.

We also recognize several limitations. The parents of eligible participants self-selected to attend the routine clinical follow-up appointment and agreed to participate in our larger study to investigate the quality of routine outcome data collection. The Q-CHAT response rate of 70%, although high for a questionnaire, may have introduced additional selection bias. We were also unable to assess children from non-English speaking families, thus limiting the broader applicability of our findings to other populations, particularly in those with higher proportions of non-English speakers. The lack of a contemporaneous control group meant that Q-CHAT scores could only be compared with the general population estimates. The proportion of preterm children in the unselected population from which the published estimates were based is unknown. In addition, there were higher proportions of children in the preterm study population living in more deprived areas. The preterm children were at a slightly older age when the Q-CHAT questionnaires were completed by their parents compared with the unselected general population. Nevertheless, no correlation between Q-CHAT score and age was found in both the general population and in our study cohort. Around one-quarter of the parents had completed the Q-CHAT after their child had received the Bayley-III assessment. Although the Q-CHAT scores of these children did not differ from the scores of those children whose parents had completed questionnaire before the Bayley-III assessment, knowledge of the results from the developmental assessment might have influenced parents' responses on the questionnaire and resulted in reporting bias.

Both the Q-CHAT and the M-CHAT were designed to screen for ASD in the general population. The predictive validity of these screening tools when applied to the preterm population had not been investigated. It is likely that preterm children and those with ASD share common early developmental problems, such as specific language impairment and attention deficits. Hence, results from these early screening tools need to be interpreted with caution. Although a range of perinatal conditions have been linked to autism risk,²⁸ our analysis was restricted by the range of data available through the NNRD. We have not been able to identify specific neonatal factors that contribute to social-communication difficulties that could aid in early risk identification for targeted intervention. In addition, the challenges faced in the early assessment of autistic features of children with major functional disabilities and in non-English speaking groups will require consideration. Further clarification of social-communication developmental trajectories in preterm children will enable us to understand the significance of these early autistic features, the validity of early autism screening in the preterm population and,

indeed, give a better insight into the largely unknown etiology of ASD. ■

The authors thank the staff from the participating hospitals (Addenbrooke's, Chelsea and Westminster, Ealing, Hillingdon, Homerton, Newham, North Middlesex, Northwick Park, Queen's [Romford], Royal London, St. Thomas', West Middlesex, Whipps Cross) for assistance with recruitment and in capturing electronic data, the Neonatal Data Analysis Unit, Imperial College London team (Eugene Statnikov and Daniel Gray [data analysts], Shalini Santhakumaran [statistician], and Richard Colquhoun [manager]) for data management and administrative support.

Submitted for publication Nov 12, 2012; last revision received Apr 26, 2013; accepted Jul 9, 2013.

References

1. American Psychiatric Association. Task Force for the Handbook of Psychiatric Measures. Diagnostic and Statistical Manual of Mental Disorders: DSM-IV-TR. 4th ed., text revision. Washington, DC: American Psychiatric Association; 2000.
2. World Health Organization. International Statistical Classification of Diseases and Related Health Problems. 10th revision. 2nd ed. Geneva: World Health Organization; 2004.
3. Johnson S, Hollis C, Kochhar P, Hennessy E, Wolke D, Marlow N. Autism spectrum disorders in extremely preterm children. *J Pediatr* 2010;156:525-31.
4. Pinto-Martin JA, Levy SE, Feldman JF, Lorenz JM, Paneth N, Whitaker AH. Prevalence of autism spectrum disorder in adolescents born weighing <2000 grams. *Pediatrics* 2011;128:883-91.
5. Williams JG, Higgins JP, Brayne CE. Systematic review of prevalence studies of autism spectrum disorders. *Arch Dis Child* 2006;91:8-15.
6. Autism and Developmental Disabilities Monitoring Network Surveillance Year 2006 Principal Investigators, Centers for Disease Control and Prevention (CDC). Prevalence of autism spectrum disorders—Autism and Developmental Disabilities Monitoring Network, United States 2006. *MMWR Surveillance Summaries* 2009;58:1-20.
7. Gardener H, Spiegelman D, Buka SL. Perinatal and neonatal risk factors for autism: a comprehensive meta-analysis. *Pediatrics* 2011;128:344-55.
8. Guinchat V, Thorsen P, Laurent C, Cans C, Bodeau N, Cohen D. Pre-, peri-, and neonatal risk factors for autism. *Acta Obstet Gynecol Scand* 2012;91:287-300.
9. American Academy of Pediatrics. Identifying infants and young children with developmental disorders in the medical home: an algorithm for developmental surveillance and screening. *Pediatrics* 2006;118:405-20.
10. Limperopoulos C, Bassan H, Sullivan NR, Soul JS, Robertson RL Jr, Moore M, et al. Positive screening for autism in ex-preterm infants: prevalence and risk factors. *Pediatrics* 2008;121:758-65.
11. Kuban KC, O'Shea TM, Allred EN, Tager-Flusberg H, Goldstein DJ, Leviton A. Positive screening on the Modified Checklist for Autism in Toddlers (M-CHAT) in extremely low gestational age newborns. *J Pediatr* 2009;154:535-40.
12. Moore T, Johnson S, Hennessy E, Marlow N. Screening for autism in extremely preterm infants: problems in interpretation. *Dev Med Child Neurol* 2012;54:514-20.
13. Luyster RJ, Kuban KC, O'Shea TM, Paneth N, Allred EN, Leviton A. The Modified Checklist for Autism in Toddlers in extremely low gestational age newborns: individual items associated with motor, cognitive, vision and hearing limitations. *Paediatr Perinat Epidemiol* 2011;25:366-76.
14. Allison C, Baron-Cohen S, Wheelwright S, Charman T, Richler J, Pasco G, et al. The Q-CHAT (Quantitative Checklist for Autism in Toddlers): a normally distributed quantitative measure of autistic traits at 18-24 months of age: preliminary report. *J Autism Dev Disord* 2008;38:1414-25.
15. Bayley N. Technical Manual for the Bayley Scales of Infant and Toddler Development. 3rd ed. San Antonio, TX: Psychological Corporation; 2006.

16. Haataja L, Mercuri E, Regev R, Cowan F, Rutherford M, Dubowitz V, et al. Optimality score for the neurologic examination of the infant at 12 and 18 months of age. *J Pediatr* 1999;135:153-61.
17. Department for Communities and Local Government. English Indices of Deprivation 2010. London; 2011. Available at: <https://www.gov.uk/government/publications/english-indices-of-deprivation-2010>. [uploaded March 24, 2011; cited February 16, 2013].
18. Cochran WG. Some methods for strengthening the common χ^2 tests. *Biometrics* 1954;10:417-51.
19. van Noort-van der Spek IL, Franken MC, Weisglas-Kuperus N. Language functions in preterm-born children: a systematic review and meta-analysis. *Pediatrics* 2012;129:745-54.
20. Barre N, Morgan A, Doyle LW, Anderson PJ. Language abilities in children who were very preterm and/or very low birth weight: a meta-analysis. *J Pediatr* 2011;158:766-74.
21. Als H. A syntactic model of neonatal behavioral organization: Framework for the assessment and support of the neurobehavioral development of the premature infant and his parents in the environment of the neonatal intensive care unit. *Phys Occupational Ther Pediatr* 1986; 6:3-55.
22. Bar-Shalita T, Vatine JJ, Parush S. Sensory modulation disorder: a risk factor for participation in daily life activities. *Dev Med Child Neurol* 2008;50:932-7.
23. Bart O, Shayevits S, Gabis LV, Morag I. Prediction of participation and sensory modulation of late preterm infants at 12 months: a prospective study. *Res Dev Disabil* 2011;32:2732-8.
24. Bishop SL, Richler J, Lord C. Association between restricted and repetitive behaviors and nonverbal IQ in children with autism spectrum disorders. *Child Neuropsychol* 2006;12:247-67.
25. Ozonoff S, Macari S, Young GS, Goldring S, Thompson M, Rogers SJ. Atypical object exploration at 12 months of age is associated with autism in a prospective sample. *Autism* 2008;12:457-72.
26. Johnson S, Marlow N. Positive screening results on the modified checklist for autism in toddlers: implications for very preterm populations. *J Pediatr* 2009;154:478-80.
27. Baron-Cohen S, Allen J, Gillberg C. Can autism be detected at 18 months? The needle, the haystack, and the CHAT. *Br J Psychiatr* 1992;161:839-43.
28. Angelidou A, Asadi S, Alysandratos KD, Karagkouni A, Kourembanas S, Theoharides TC. Perinatal stress, brain inflammation, and risk of autism—review and proposal. *BMC Pediatr* 2012;12:89.

50 Years Ago in *THE JOURNAL OF PEDIATRICS*

Central Nervous System Complications of Children with Acute Leukemia: An Evaluation of Treatment Methods

Evans AE, D'Angio GJ, Mitus A. *J Pediatr* 1964;64:94-6

Fifty years ago in *The Journal*, Evans et al reported an historical cohort study of 53 children symptomatic from intracranial leukemia, treated by either lumbar puncture alone, cranial irradiation, or intrathecal methotrexate. Today the results are not surprising; 98% of the irradiated patients became symptom-free for 2.8 months, and 88% of the methotrexate-treated patients were symptom-free 3.7 months ($P = .07$). Only 48% of the children undergoing simple lumbar puncture became asymptomatic and even then for only about 2 weeks. Whether this study's small sample size would have survived today's peer review process to merit publication is debatable.

Nonetheless, this paper presaged a landmark shift in the management and cure of childhood acute lymphoblastic leukemia (ALL). Indeed, the authors mention in their last paragraph the "prophylactic" use of monthly intrathecal methotrexate in 12 patients whose central nervous system leukemia had already been eradicated and their symptom-free survival of 7 or more months. That success was unheralded! But, by the end of the 1960s, the "Total Therapy" studies at St. Jude Children's Research Hospital using prophylactic craniospinal irradiation and intrathecal methotrexate reduced the rate of leukemia relapse into the central nervous system from more than 50% to approximately 10%, and afforded cures in greater than 50% of children in an era when ALL took the lives of 80% or more.¹

The next 50 years of advancing cures in ALL and other childhood cancers will be more challenging. Now that many childhood cancers have 5-year survival rates in excess of 80%, clinical trials will be cumbersome, as much larger sizes will be necessary to demonstrate increasingly smaller increments of success. Accumulating the larger samples sizes required will be difficult because of cost as well as the increasing division of childhood cancers into smaller and smaller diseases based upon molecular subtypes. Surely we will conceive new clinical trial designs to tackle these logistical issues. However, we should never give short shrift to initial observations such as that of Evans et al, even when sample size is small, if the findings are compelling and the impact potentially immense.

Paul Graham Fisher, MD

Departments of Neurology, Pediatrics, and Human Biology

Lucile Packard Children's Hospital

Stanford University

Palo Alto, California

<http://dx.doi.org/10.1016/j.jpeds.2013.08.003>

Reference

1. Simone J, Aur RJA, Hustu HO, Pinkel D. "Total therapy" studies of acute lymphocytic leukemia in children: current results and prospects for cure. *Cancer* 1972;30:1488-94.

Table I. Neonatal and sociodemographic characteristics of Q-CHAT respondents and non-respondents

Characteristics	Respondents (n = 141)	Non-respondents (n = 60)	P
Gestation, median (range), wk	27 (23-29)	27 (23-29)	.15
Birth weight, median (range), g	957.5 (490-1720)	920 (560-1400)	.07
Birth weight z-score, mean (SD)	-0.10 (0.98)	-0.21 (0.95)	.50
Male sex, n (%)	73 (51.8)	20 (33.3)	.02
Singleton pregnancy, n (%)	109 (77.3)	48 (80.0)	.67
Ethnicity, n (%)			
White	66 (46.8)	21 (35.0)	.12
Non-white	75 (53.2)	39 (65.0)	
Maternal age, median (range)	32 (14-46)	31 (18-58)	.90
Cesarean section delivery, n (%)	71 (50.4)	32 (53.3)	.34
Length of mechanical ventilation, median (range), d	1 (0-54)	1 (0-61)	.13
Supplemental oxygen requirement at 36 wks post-menstrual age, n (%)	38 (27.0)	23 (38.3)	.11
Received breast milk during neonatal stay, n (%)	141 (100.0)	60 (100.0)	-
IMD quintile, n (%)			
One (least deprived)	14 (9.9)	6 (10.0)	
Two	15 (10.6)	5 (8.3)	
Three	22 (15.6)	5 (8.3)	.12
Four	42 (29.8)	12 (20.0)	
Five (most deprived)	48 (34.0)	32 (53.3)	
Bayley-III cognitive composite score, mean (SD)	94.6 (13.0)	91.6 (11.7)	.12
Cognitive impairment*, n (%)			
None	137 (97.2)	59 (98.3)	
Mild-moderate	4 (2.8)	1 (1.7)	.62
Severe	0 (0)	0 (0)	
Bayley-III language composite score, mean (SD)	87.7 (13.0)	89.3 (12.1)	.41
Language skills impairment*, n (%)			
None	124 (87.9)	56 (93.3)	
Mild-moderate	16 (11.4)	2 (3.3)	.08
Severe	1 (0.7)	2 (3.3)	

*Impairments were classified based on Bayley-III cognitive and language scores as follow: none = scores ≥ 70 , mild-moderate = scores ≥ 55 and < 70 , severe = scores < 55 .

Appendix

Members of the Medicines for Neonates Investigator Group include: Neena Modi, MD, FRCPCH, Imperial College London, United Kingdom; Peter Brocklehurst, MBChB, FRCOG, Institute for Women's Health, University College London, United Kingdom; Jane Abbott, Bliss, United Kingdom; Kate Costeloe, MBBChir, FRCPCH, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, United Kingdom; Elizabeth S Draper, PhD, University of Leicester, United Kingdom; Azeem Majeed, MBBCh, FRCGP, Imperial College London, United Kingdom; Jacquie Kemp, MSc, RGN, London Specialised Commissioning Group, United Kingdom; Deborah Ashby, PhD, School of Public Health, Imperial College London, United Kingdom; Alys Young, PhD, Manchester Academic Health Sciences Centre, University of Manchester, United Kingdom; and Stavros Petrou, PhD, Warwick Medical School, University of Warwick, United Kingdom.

Appendix 2: The NPEU/Oxford classification of disability

DOMAIN Key questions	CRITERIA FOR DISABILITY (Criteria for severe disability in bold)
<p><u>Malformation</u> Does the child have a malformation?</p>	<p>Any anomaly detected at birth or apparent within the first two postnatal years, which is likely to result in death, disfigurement or disability, and which is likely to require medical or surgical treatment</p> <p>Any malformation which despite physical assistance impairs the performance of daily activities</p>
<p><u>Neuromotor function</u> Does the child have any difficulty walking?</p> <p>Does the child have any difficulty sitting?</p> <p>Does the child have any difficulty with hand use?</p> <p>Does the child have any difficulty with head control?</p>	<p>Non-fluent gait Abnormal gait reducing mobility Unable to walk without assistance</p> <p>Sits unsupported but unstable Sits supported Unable to sit</p> <p>Some difficulty feeding with one hand Some difficulty feeding with both hands Unable to use hands to feed self</p> <p>Unstable but no support required Unable to control head movement without support No head control</p>
<p><u>Auditory function</u> Does the child have any difficulty hearing</p>	<p>Hearing impaired, not aided Hearing impaired, corrects with aids Hearing impaired, uncorrected even with aids</p>
<p><u>Communication</u> Is there any difficulty with communication</p>	<p>Unable to comprehend word/sign out of familiar context Unable to comprehend word/sign in cued situation</p> <p>Uses single words only, vocabulary > 10 words Vocabulary < 10 words Unable to produce > 5 recognisable sounds No vocalisation</p>
<p><u>Visual function</u> Does the child have any difficulty with vision?</p>	<p>Normal vision with correction Not fully correctable Blind or sees light only</p>
<p><u>Cognitive function</u> Does the child have any learning difficulty? (function assessed on standardised test)</p>	<p>2 to 3 standard deviations below mean (between 6 - 11 months behind at 2 years) More than 3 standard deviations below mean (approximately 12 months behind at 2 years or more)</p>

Seizures

Does the child have seizures? No treatment required
No seizures on treatment
Seizures less than 1/month despite treatment
Seizures more than 1/month despite treatment

Other physical disability

Does the child have any other disability

Respiratory:
Limited exercise tolerance, no drug treatment
Limited exercise tolerance, on drug treatment
Requires continual oxygen therapy
Requires mechanical ventilation

Gastro-intestinal function:
Requires special diet
Has stoma
Requires tube feeding
Requires parenteral nutrition

Renal function:
Renal impairment known, no treatment
Renal impairment, on drug or dietary treatment only
Requires dialysis

Growth
Height or weight 2 to 3 standard deviations below mean for age
Height or weight more than 3 standard deviations below mean for age

Appendix 3: Definitions of disability used by studies reporting neonatal outcomes

Definitions of disability in different domains						
Developmental	Communication	Vision	Hearing	Motor	Other/comments	
EPICure (UK) (Wood 200b)						
Cohort: All infants born at less than 26 completed weeks of gestation in UK and Ireland between March and December 1995.						
At 30 months:						
<u>Severe disability:</u> BSID-II MDI<55						
	Communicating by systematized method only or not communicating by speech or other method	Blind or perceives light only	Impaired and uncorrected even with hearing aid*	Unable to sit Unable to walk without assistance Unable to use hands to feed self Unable to control head movement without support/ no head control		
<u>Other disability:</u> Mild: BSID-II MDI 70-84 Moderate: BSID-II MDI 55-69						
	Delay in speech and other systematized method of communicating	Normal with correction Useful vision but not fully correctable	Impaired but hearing aid not required Impaired, corrected with hearing aid	Nonfluent, abnormal gait, reduced mobility Sits unsupported but unstable, sits supported Some difficulty feeding with both hands Unstable head control but no support required	Non-febrile seizure	

EpiCure 2 (UK) (Moore 2012b)

Cohort: All infants born between 22 and 26 completed weeks of gestation in England during 2006.

At 3 years:

<u>Severe impairment:</u>	Developmental quotient less than 3 SD below mean for age	Communicating by systematized method only or not communicating by speech or other method	Blindness	Profound sensorineural hearing loss not improved by aids	Non-ambulant cerebral palsy (GMFCS levels 3-5)
<u>Moderate impairment:</u>	Developmental quotient 2-3 SD below mean for age	Delay in speech and other systematized method of communicating	Functionally impaired vision	Hearing loss improved by aids	Ambulant cerebral palsy (GMFCS level 2)
<u>Mild impairment:</u>	Developmental quotient 1-2 SD below mean for age		Squints or refractive errors	Hearing loss not sufficient to require aids	Cerebral palsy with GMFCS level 1

Former Trent region cohort study (UK) (Rattihalli 2011)

Cohort: Infants born <26 weeks of gestation in 2001-2003 in the former Trent region.

At 2 years, 2 criteria were used to define disability:

(i) NPEU/Oxford criteria:

<u>Severe disability:</u>	Cognitive score < -3 SD below mean; About 12 months behind at 2 years	Inability to comprehend word/sign in cued situation, Produce less than 5 recognisable sounds or no vocalisation	Blind or seeing light only	Hearing impairment, uncorrected even with aids	Inability to sit Inability to use hands to feed self Inability to control head movement without support No head	Malformation which despite physical assistance impairs the performance of daily activities; Seizures;
---------------------------	---	---	----------------------------	--	--	--

			movement	Requiring mechanical ventilation or continuous oxygen therapy; requiring parenteral nutrition or tube feeding; requiring dialysis; or height or weight < -3 SD below mean for age	
<hr/>					
(ii) Audit Commission criteria:					
<u>Severe disability:</u>	Developmental score < -2SD on Griffiths scale	Blind or seeing light only or corrected binocular visual acuity of <6/24	Hearing loss uncorrected even with aids or loss more than 60dB	Cerebral palsy	Epilepsy (needing regular medication); breathlessness requiring intervention in form of oxygen, ventilation or tracheostomy; hydrocephalus requiring shunt; presence of stoma; loss of limb

EIPAGE (France) (Larroque 2008)

Cohort: All infants born between 22 and 32 completed weeks of gestation from nine regions in France in 1997.

At 5 years:

<u>Severe disability:</u> KABC MPC <55	Assessment with the Rossano test: visual deficiency <3/10 for both eyes	Hearing loss >70dB for one or both ears, or use of hearing aid	Non-ambulatory cerebral palsy
<u>Moderate disability:</u> KABC MPC 55-69			Cerebral palsy, able to walk with aid
<u>Minor disability:</u> KABC MPC 70-84	Visual deficit <3/10 for one eye		Cerebral palsy, able to walk without aid

EPIBEL (Belgium) (De Groot 2007)

Cohort: All infants born at or below 26 weeks of gestation in a geographically defined region of Belgium from 1999-2000.

At 3 years:

<u>Severe disability:</u> BSID-II MDI and PDI of 70 or greater were considered as normal. Therefore, if MDI or PDI were >70 and no impairment was	No speech or systematized methods of communication	No useful vision	No useful hearing	No head control Unable to sit No independent walking Unable to dress and feed self
<u>Mild-moderate disability:</u> detected in other domains, the outcome is considered to be 'no overall disability'	Speech and other formal methods of communication	Impairment but some or little useful vision (with aids if worn)	Impairment but useful hearing (with aids if worn)	Head and neck unstable but without support or only for very short periods Sitting Non-febrile seizure

					unsupported but unstable Nonfluent or abnormal, reduced mobility Some difficulty but able to dress self, Unable to dress self but able to feed self
EXPRESS (Sweden) (Serenius 2013)					
Cohort: All infants born before 28 weeks gestation in Sweden between 2004-2007.					
At 30 months:					
<u>Severe disability:</u>	Bayley-III cognitive score < -3SD from mean	Bayley-III language score < -3SD from mean	Bilateral blindness (unable to fixate and follow a light with either eye)	Deafness (hearing loss uncorrectable with aids)	Bayley-III motor score < mean-3SD Severe cerebral palsy (unable to walk even with aid)
<u>Moderate disability:</u>	Bayley-III cognitive score between -2 and -3SD from mean	Bayley-III language score between -2 and -3SD from mean	Registered at low-vision centres without blindness	Hearing loss corrected with aids	Bayley-III motor score between -2 and -3SD from mean Moderate cerebral palsy (able to walk with aid)
<u>Mild disability:</u>	Bayley-III cognitive score between -1 and -2SD from mean	Bayley-III language score between -1 and -2SD from mean			Bayley-III motor score between -1 and -2SD from mean

						Mild cerebral palsy (able to walk independently)
Helsinki follow-up study (Finland) (Tommiska 2003)						
Cohort: All ELBW (birth weight <1000g) infants born in Finland in 1996-1997.						
At 18 months:						
<u>Severe impairment:</u>	Not considered in the classification of disability		Blindness	Hearing impairment requiring hearing aid	Cerebral palsy	Seizures or combination of ≥ 3 milder impairments in vision, hearing, motor or speech assessment
<u>Mild impairment:</u>	Not considered in the classification of disability	No definition of mild speech impairment given	No definition of mild visual impairment given	No definition of mild hearing impairment given	No definition of mild motor impairment given	1-2 impairments in vision, hearing, motor or speech assessment but does not meet criteria for severe impairment.
Swiss national cohort (Schlapbach 2012)						
Cohort: All infants born between 24+0 and 27+6 weeks gestation during 2000-2008.						
At 2 years:						
<u>Severe disability:</u>	BSID-II MDI<55		Blindness or only perception of light or light reflecting objects	No useful hearing even with aids	Cerebral palsy with GMFCS levels 3-5, or	BSID-II PDI<55

<u>Moderate disability:</u> BSID-II MDI 55-69	Moderately reduced vision but better than severe visual disability, or unilateral blindness with good vision in contralateral eye	Hearing loss corrected with aids	Cerebral palsy with GMFCS level 2, or BSID-II PDI 55-69
The Netherlands national cohort (de Waal 2012) Cohort: All infants born at 23 to 27 weeks of gestation in The Netherlands in 2007. At 2 years:			
<u>Moderate-severe disability:</u> BSID-II MDI<70	Blind, only perceiving light or sight worse than 6/18 when corrected	No useful hearing even with aids or more than 40dB hearing loss with aids	Cerebral palsy with GMFCS levels 2-5; BSID-II PDI<70
<u>Mild disability:</u> BSID-II MDI 70-84			Cerebral palsy with GMFCS level 1; BSID-II PDI 70-84
Norwegian national cohort (Leveresen 2011) Cohort: All infants born between 22 and 27 weeks or with birth weight between 500g and 999g in 1999-2000. At 5 years:			
<u>Severe disability:</u> WPPSI-R full-scale IQ<55	Legal blindness	Complete deafness	Cerebral palsy with GMFCS levels 4-5
<u>Moderate disability:</u> WPPSI-R full-scale IQ 55-69	Severe visual impairment	Need of hearing aid	Cerebral palsy with GMFCS levels 2-3
<u>Mild disability:</u> WPPSI-R full-scale IQ 70-84	Squint or refractive error	Mild hearing loss	Movement Assessment Battery for

				Children score >95 th percentile
Victorian Infant Collaborative Study 2005 cohort* (Australia) (Doyle 2010b)				
Cohort: All infants born between 22-27 completed weeks of gestation in the state of Victoria, Australia in 2005.				
At 2 years:				
<u>Severe disability:</u>	Bayley-III cognitive score < -3 SD of control mean	Bayley-III language score < -3 SD of control mean	Blindness (visual acuity <20/200 in the better eye)	Cerebral palsy with GMFCS levels 4-5
<u>Moderate disability:</u>	Bayley-III cognitive score -2 to -3 SD of control mean	Bayley-III language score -2 to -3 SD of control mean	Deafness (hearing loss requiring amplification or worse)	Cerebral palsy with GMFCS levels 2-3
<u>Mild disability:</u>	Bayley-III cognitive score -1 to -2 of control mean	Bayley-III language score -1 to -2 of control mean		Cerebral palsy with GMFCS level 1
National Institute of Child Health and Human Development Neonatal Research Network Follow-up Study of High Risk Infants 2002-2004 cohort* (USA)				
(Hintz 2011)				
Cohort: Infants born at <25 weeks gestation with birth weight 401-1000g at and NICHD Neonatal Research Network site in 2002-2004.				
At 18-22 months:				
<u>Profound disability:</u>	BSID-II MDI <50			Cerebral palsy with GMFCS levels 4-5
<u>Moderate-severe disability:</u>	BSID-II MDI 50-69		Bilateral blindness (absence of functional vision in	Deafness (permanent hearing loss that required GMFCS levels 2-3

	either eye)	amplification in both ears)	
Vermont Oxford Network Extremely Low Birth Weight Follow-up 1998-2003 cohort* (International) (Mercier 2010)			
Cohort: Infants with birth weight 401-1000g born between 1 July 1998 and 31 December 2003 who received follow-up at 33 participating North American Vermont Oxford network centres.			
At 18-24 months:			
<u>Severe disability:</u>	BSID-II MDI<70	Bilateral blindness	Hearing impairment requiring amplification
			BSID-II MDI<70; Cerebral palsy; Inability to walk 10 steps with support

*These studies have either ongoing recruitment or repeated recruitments at different periods. The latest cohort or publication reporting early neurodevelopmental outcomes were chosen to be included in this table for comparison.

Study Number

**Reliability of two-year neurodevelopmental
assessment in preterm infants**

Assessment Record

SECTION A: GENERAL INFORMATION

Study Number:

Gender: Male Female

Gestational age at birth:

	Years	Months	Days
Date of assessment:	<input type="text"/>	<input type="text"/>	<input type="text"/>
EDD:	<input type="text"/>	<input type="text"/>	<input type="text"/>
Adjusted age:	<input type="text"/>	<input type="text"/>	<input type="text"/>
(in months and days)	<input type="text"/>	<input type="text"/>	<input type="text"/>

Consultant Paediatrician:

Permission to forward results to Paediatrician:

General Practitioner:

Permission to forward results to GP:

What is (are) the main language(s) spoken at home?

Child accompanied by: Mother Father Others
please specify: _____

Site: _____

SECTION B: NEUROSENSORY INFORMATION

VISUAL OR EYE PROBLEM

	Left eye	Right eye
Is there a visual or eye defect of any type present?	Yes <input type="checkbox"/>	Yes <input type="checkbox"/>
	No <input type="checkbox"/>	No <input type="checkbox"/>

Does the child wear glasses?	Yes <input type="checkbox"/>	No <input type="checkbox"/>
------------------------------	------------------------------	-----------------------------

Usual vision (with glasses with worn)

Normal or near normal	<input type="checkbox"/>
Impaired but appears to have useful vision	<input type="checkbox"/>
Sees light or gross movement only	<input type="checkbox"/>
No useful vision (blind)	<input type="checkbox"/>

Is there a squint present?	Left <input type="checkbox"/>	Right <input type="checkbox"/>	No <input type="checkbox"/>
Are there abnormal eye movement present	Left <input type="checkbox"/>	Right <input type="checkbox"/>	No <input type="checkbox"/>
Has the child had any ophthalmic assessment or intervention?	Left <input type="checkbox"/>	Right <input type="checkbox"/>	No <input type="checkbox"/>

If yes, please describe

Details of ophthalmic specialist
(if applicable):

Other comments:
(including any parental concerns regarding the child's vision)

HEARING PROBLEM

Is there a hearing impairment of any type present? **Left ear** **Right ear**
Yes Yes
No No

Does the child normally wear aids? Left Right No

Usual hearing (with aids if worn)

Normal or near normal
Hearing loss corrected with aids
Some hearing but loss not corrected by aids
No useful hearing even with aids

Has the child had any hearing assessment or intervention? Left Right No

If yes, please describe

Details of hearing specialist
(if applicable):

Other comments:
*(including any parental
concerns regarding the child's
hearing)*

SECTION C: NEUROLOGICAL EXAMINATION

SCORES



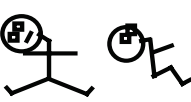








Cranial Nerves (Max 15)	Posture (Max 18)	Movements (Max 6)	Tone (Max 24)	Reflexes (Max 15)	Total (Max 78)
			R	R	R
			L	L	L

Overall comments:

CRANIAL NERVE FUNCTION

	Score 3	Score 2	Score 1	Score 0	Score
Facial appearance (at rest and when crying or stimulated)	smiles or reacts to stimuli by closing eyes and grimacing		closes eyes but not tightly poor facial expression	expressionless, does not react to stimuli	
Eye appearance	normal conjugated eye movements		Intermittent deviation of eyes or abnormal movements	continuous deviation of eyes or abnormal movements	
Auditory response test the response to rattle or bell	reacts to stimuli on both sides		doubtful reaction to stimuli or asymmetrical	does not react to stimuli	
Visual response test the ability to follow a red ball or moving object	follows the object for a complete arc		follows the object for an incomplete arc or asymmetry	does not follow the object	
Sucking/swallowing watch the infant suck on breast or bottle	good suck and swallowing,		poor suck and/or swallowing	no sucking reflex no swallowing	












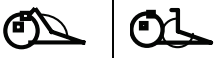

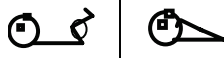

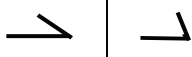






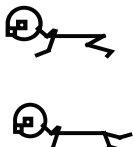
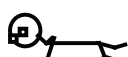


POSTURE

	Score 3	Score 2	Score 1	Score 0	Score
Head in sitting	 straight; in midline		 slightly to side <i>or</i> backward <i>or</i> forward	 markedly to side <i>or</i> backward <i>or</i> forward	
Trunk in sitting	 straight		 slightly curved <i>or</i> bent to side	   very rounded rocking back bent sideways	
Arms at rest	in neutral position, central straight <i>or</i> slightly bent		slight internal rotation <i>or</i> external rotation intermittent dystonic posture	marked internal rotation <i>or</i> external rotation <i>or</i> dystonic posture hemiplegic posture	
Hands	hands open		intermittent adducted thumb <i>or</i> fisting	persistent adducted thumb <i>or</i> fisting	
Legs in sitting in supine and in standing	 able to sit with straight back and legs straight <i>or</i> slightly bent (long sitting) legs in neutral position; straight <i>or</i> slightly bent	slight internal rotation <i>or</i> external rotation	 sit with straight back but knees bent at 15-20 ° internal rotation <i>or</i> external rotation at hips	 unable to sit straight unless knees markedly bent (no long sitting) marked internal rotation <i>or</i> external rotation <i>or</i> fixed extension <i>or</i> flexion <i>or</i> contractures at hips and knees	
Feet in supine and in standing	central; in neutral position toes straight midway between flexion and extension		slight internal rotation <i>or</i> external rotation intermittent tendency to stand on tiptoes <i>or</i> toes up <i>or</i> curling under	marked internal rotation <i>or</i> external rotation at the ankle persistent tendency to stand on tiptoes <i>or</i> toes up <i>or</i> curling under	





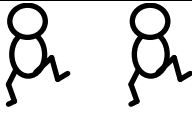
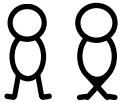






MOVEMENTS

	Score 3	Score 2	Score 1	Score 0	Score
Quantity Watch infant lying in supine	Normal		Excessive <i>or</i> sluggish	Minimal <i>or</i> none	
Quality	Free, alternating, and smooth		Jerky, Slight tremor	<ul style="list-style-type: none"> • Cramped & synchronous, • Extensor spasms, • Athetoid; • Ataxic, • Very tremulous, • Myoclonic spasm • Dystonic movement 	

TONE

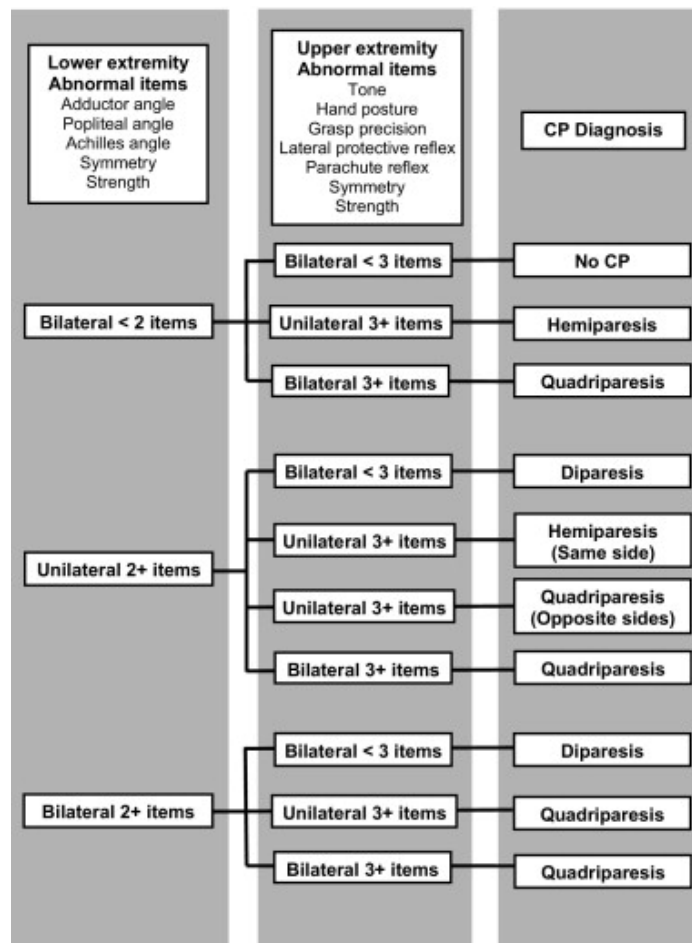
	Score 3	Score 2	Score 1	Score 0	Score
Scarf sign Take the infant's hand and pull the arm across the chest until there is resistance. Note the position of the elbow.	Range:  R L R L		 R L	 or  R L R L	R
					L
Passive shoulder elevation Lift arm next to the infant's head. Note resistance at shoulder and elbow.	resistance but overcomeable  R L	resistance difficult to overcome R L	no resistance  R L	resistance, not overcomeable  R L	R
					L
Pronation/ supination Steady upper arm while pronating and supinating forearm, note resistance	full pronation and supination, no resistance R L		full pronation and supination but resistance to be overcome R L	full pronation and supination not possible, marked resistance R L	R
					L
Adductors With the infant's legs extended, open them as far as possible. The angle formed by the legs is noted.	Range: 150°-80°  R L R L	150°-160°  R L	>170°  R L	<80°  R L	R
					L
Popliteal angle Legs are flexed at the hip simultaneously on to the side of the abdomen, then extended at the knee until there is resistance. Note angle between lower and upper leg.	Range: 100°-150°  R L R L	150°-160°  R L	90° or >170°  R L R L	<80°  R L	R
					L
Ankle dorsiflexion With knee extended, dorsiflex ankle. Note the angle between foot and leg.	Range: 30°- 85°  R L R L	20°-30°  R L	<20° or 90°  R L R L	>90°  R L	R
					L
Pulled to sit Pull infant to sit by wrists					
Ventral suspension Hold infant in ventral suspension; note position of back, limbs and head	 				

REFLEXES AND REACTIONS

	Score 3	Score 2	Score 1	Score 0	Score
Tendon reflexes	easily elicitable	mild brisk	brisk	clonus or absent	
Arm protection Pull the infant by one arm from the supine position and note the reaction of the opposite side.	 arm & hand extended R L		 arm semi-flexed R L	 arm fully flexed R L	R
					L
Vertical suspension hold infant under axilla make sure legs do not touch any surface	 kicks symmetrically		 kicks one leg more, or poor kicking	 no kicking even if stimulated, or scissoring	
Lateral tilting (describe side up). Infant held vertically tilt quickly to horizontal. Note spine, limbs and head	 R L	 R L	 R L	 R L	R
					L
Forward parachute Infant held vertically and suddenly tilted forward . Note reaction of the arms.		asymmetrical partial			

SECTION D: ASSESSMENT OF CEREBRAL PALSY

CEREBRAL PALSY ALGORITHM



NO CEREBRAL PALSY

Classification of cerebral palsy

- | | | |
|-------------------|--|--------------------------|
| Spastic bilateral | 2 limb involvement/ Diparesis | <input type="checkbox"/> |
| | 3 limb involvement/ Asymmetric quadriparesis | <input type="checkbox"/> |
| | 4 limb involvement/ Quadriparesis | <input type="checkbox"/> |
| Hemiplegia | Right-sided | <input type="checkbox"/> |
| | Left-sided | <input type="checkbox"/> |
| Other | Dyskinetic | <input type="checkbox"/> |

Comments:

GROSS MOTOR FUNCTION CLASSIFICATION SCALE (GMFCS)

Level of gross motor ability

< 24 months corrected age	
Level 1	Infants move in and out of sitting and floor sit with both hands free to manipulate objects. Infants crawl on hands and knees, pull to stand and take steps holding on to furniture. Infants walk between 18 months and 2 years of age without the need for any assistive mobility device.
Level 2	Infants maintain floor sitting but may need to use their hands for support to maintain balance. Infants creep on their stomach or crawl on hands and knees. Infants may pull to stand and take steps holding on to furniture.
Level 3	Infants maintain floor sitting when the low back is supported. Infants roll and creep forward on their stomachs.
Level 4	Infants have head control but trunk support is required for floor sitting. Infants can roll to supine and may roll to prone.
Level 5	Physical impairments limit voluntary control of movement. Infants are unable to maintain antigravity head and trunk postures in prone and sitting. Infants require adult assistance to roll.
≥ 24 months corrected age	
Level 1	Children floor sit with both hands free to manipulate objects. Movements in and out of floor sitting and standing are performed without adult assistance. Children walk as the preferred method of mobility without the need for any assistive mobility device.
Level 2	Children floor sit but may have difficulty with balance when both hands are free to manipulate objects. Movements in and out of sitting are performed without adult assistance. Children pull to stand on stable surface. Children crawl on hands and knees with a reciprocal pattern, cruise holding onto furniture and walk using an assistive mobility device as preferred methods of mobility.
Level 3	Children maintain floor sitting often by “W-sitting” and may require adult assistance to assume sitting. Children creep on their stomach or crawl on hands and knees (often without reciprocal leg movements) as their primary methods of self mobility. Children may pull to stand on a stable surface and cruise short distances. Children may walk short distances indoors using an assistive mobility device and adult assistance for steering and turning.
Level 4	Children floor sit when placed, but are unable to maintain alignment and balance without use of their hands for support. Children frequently require adaptive equipment for sitting and standing. Self mobility for short distances is achieved through rolling, creeping on stomach, or crawling on hands and knees without reciprocal leg movement.
Level 5	Physical impairments restrict voluntary control of movement and the ability to maintain antigravity head and trunk postures. All areas of motor function are limited. Functional limitations in sitting and standing are not fully compensated for through the use of adaptive equipment and assistive technology. Children have no means of independent mobility and are transported.

MANUAL ABILITIES CLASSIFICATION SYSTEM (MACS)

Level of manual ability

Level 1	Handles objects easily and successfully: At most limitations in the ease of performing manual tasks requiring speed and accuracy; however, any limitations in manual activities do not restrict independence in any daily activities.
Level 2	Handles most objects but with somewhat reduced quality and/or speed of achievement: certain activities may be avoided or be achieved with some difficulty; alternative ways of performance might be used but manual abilities do not usually restrict independence in daily activities.
Level 3	Handles objects with difficulty, needs help to prepare and/or modify activities: the performance is slow and achieved with limited success regarding quality and quantity; activities are performed independently if they have been set up or anticipated.
Level 4	Handles a limited selection of easily managed objects in adapted situations: performs part of activities with effort and limited success; requires continuous support and/or adapted equipment for even partial achievement of activity.
Level 5	Does not handle objects and has severely limited ability to perform even simple actions: requires total assistance.

SECTION E: BEHAVIOUR OBSERVATION RECORD

EXAMINER RATING

	Observed most of the time	Observed some of the time	Never or rarely observed
1. Positive affect (smiles and laughs)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2. Shows enthusiasm or excitement	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3. Explores objects in the environment	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4. Readily takes part in activities	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5. Cooperates with requests	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
6. Alertness (quiet and attentive, not drowsy)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7. Appropriate muscle tone (not overly stiff, floppy or with tremor)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
8. Adapts easily to changes in stimulation or routines	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
9. Works without being overly active or fidgety	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
10. Distracted easily, interfering with performance on items	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
11. Overly sensitive to touch or textures	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
12. Approaches new tasks with apprehension	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
13. Negative affect (cries, frown, whines or complains)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

SECTION F: SOCIO-DEMOGRAPHIC INFORMATION

Maternal age at child's birth: _____

Maternal support status

Married and living with husband	
Single and living with partner	
Single, with partner living separately	
Single parent	

Other: _____

Number of children in household: _____

Maternal age last in full time education: _____

Parents' highest qualification from school/college

	Maternal	Paternal
None of the below		
Vocational qualification, NVQ, or CSE		
O Level, GCSE, or Scottish Standards		
BTEC, A Levels or Scottish Highers		
Diploma or HND		
University degree		
Postgraduate university degree		

Other: _____

Parents' occupation

	Maternal	Paternal
Professional		
Self-employed, manager, administrative, public service or equivalent		
Service worker or equivalent		
Manual worker or equivalent		
Unemployed		

Other: _____

CHILD CARE ARRANGEMENTS

Age of child when mother returned to work: _____ or N/A

Alternative childcare:

	Regularly	Sometimes
Partner/ husband/wife	<input type="checkbox"/>	<input type="checkbox"/>
Grandparent(s)	<input type="checkbox"/>	<input type="checkbox"/>
Other relatives	<input type="checkbox"/>	<input type="checkbox"/>
Friends	<input type="checkbox"/>	<input type="checkbox"/>
Nursery	<input type="checkbox"/>	<input type="checkbox"/>
Child minder	<input type="checkbox"/>	<input type="checkbox"/>

Other: _____

Does the child regularly attend:

	Yes	No
Playgroup	<input type="checkbox"/>	<input type="checkbox"/>

If yes, number of half day sessions per week: _____

Nursery	<input type="checkbox"/>	<input type="checkbox"/>
----------------	--------------------------	--------------------------

If yes, number of half day sessions per week: _____

Study Number

Reliability of two-year neurodevelopmental assessment in preterm infants

Parental Questionnaire Social & Communication Abilities at 2 years corrected age

An important part of your child's evaluation is to learn how he or she interacts with you. Because you understand your child so well, you are the best person to provide this information. We will use this information to help us understand how premature children learn social skills.

Please complete all the questions in this booklet as accurately as possible.

Please bring the completed booklet to the research appointment and give it to the doctor.

PART ONE

Based on Bayley-III Social-Emotional Scale STANLEY I. GREENSPAN, M.D.

For each question, circle the number in the column that best describes how often you observe the behaviour in your child. Circle **only one** number for each question.

	Behaviour Frequency					
	Can't tell	None of the time	Some of the time	Half of the time	Most of the time	All of the time
1. Takes a calm and enjoyable interest in most sounds.	0	1	2	3	4	5
2. You can easily get your child's attention without having to be very dramatic.	0	1	2	3	4	5
3. Takes a calm and enjoyable interest in most sights, including colourful or bright things.	0	1	2	3	4	5
4. You can easily get your child to look at things without them being very bright or colourful.	0	1	2	3	4	5
5. Calmly enjoys touching or being touched by different things.	0	1	2	3	4	5
6. You can easily get your child to respond to your touch without having to touch your child firmly to get his or her attention.	0	1	2	3	4	5
7. Likes to be swung around, danced with while in your arms, or quickly lifted up in the air.	0	1	2	3	4	5
8. You can easily get your child's attention by approaching him or her, or moving him or her around slowly.	0	1	2	3	4	5
9. You can help your child to calm down.	0	1	2	3	4	5
10. Looks at interesting sights, such as your face or a toy.	0	1	2	3	4	5
11. Looks at or turns toward interesting sounds.	0	1	2	3	4	5
12. Seems happy or pleased when he or she sees a favourite person (e.g. looks or smiles, makes sounds, or moves arms in a way that expresses joy or delight).	0	1	2	3	4	5
13. Responds to people talking or playing with him or her by making sounds or faces (e.g. happy sounds or a curious or annoyed look).	0	1	2	3	4	5
14. Reaches for or points at things, or makes distinct sounds to show you what he or she wants (e.g. reaches out to be picked up or points at a toy).	0	1	2	3	4	5
15. Exchanges two or more smiles, other looks, sounds, or actions (e.g. reaching, giving or taking) with a favourite person.	0	1	2	3	4	5
16. Shows you that he or she understands your actions or gestures by making an appropriate gesture in return (e.g. makes a funny face back at you, looks at something you point to, stops doing something when you shake your head and use a firm voice to say "No!" or smiles and does more of something when you nod with a big smile and say "Yes!").	0	1	2	3	4	5
17. Uses many consecutive actions in a back-and-forth way to show you what he or she wants or to have fun with you (e.g. smiles, reaches out for a hug, and, when you hug, takes your hat, puts it on his or her head, and smiles proudly OR takes your hand, leads you to the refrigerator, tugs on the handle, and, after you open it, points to something he or she likes, such as food, a bottle of juice, or milk).	0	1	2	3	4	5

	Behaviour Frequency					
	Can't tell	None of the time	Some of the time	Half of the time	Most of the time	All of the time
18. Copies or imitates many of your sounds, words, or actions while playing with you (e.g. if you make funny faces and sounds, he or she copies them).	0	1	2	3	4	5
19. Searches for something he or she wants by looking or getting you to look for it.	0	1	2	3	4	5
20. Shows you what he or she wants or needs by using a few actions in a row (e.g. leads you by the hand to open a door and then touches or bangs on the door).	0	1	2	3	4	5
21. Uses words or tries to use words when people talk with or play with him or her.	0	1	2	3	4	5
22. Copies or imitates familiar make-believe play (e.g. feeds or hugs a doll).	0	1	2	3	4	5
23. Tells you what he or she wants with one or a few words (e.g. "juice", "open" or "kiss").	0	1	2	3	4	5
24. Shows you he or she understands your simple verbal wish (e.g. "Please show me your toy").	0	1	2	3	4	5
25. Plays make-believe (e.g. feeds a doll, plays house, or pretends to be a TV or movie character) with you or others.	0	1	2	3	4	5
26. Uses words or pictures to tell you what he or she is interested in (e.g. "See truck!").	0	1	2	3	4	5
27. Uses words with one or more peers.	0	1	2	3	4	5
28. Uses words or pictures to show what he or she likes or dislikes (e.g. "Want that" or "No want")	0	1	2	3	4	5

Official Use

Total

Raw Score

PART TWO

Based on the Quantitative Checklist for Autism in Toddlers

Please answer the following questions about your child by ticking the appropriate box.

1. Does your child look at you when you call his/her name?

- Always
- Usually
- Sometimes
- Rarely
- Never

2. How easy is it for you to get eye contact with your child?

- Always
- Usually
- Sometimes
- Rarely
- Never

3. When your child is playing alone, does s/he line objects up?

- Always
- Usually
- Sometimes
- Rarely
- Never

4. Can other people easily understand your child's speech?

- Always
- Usually
- Sometimes
- Rarely
- Never

5. Does your child point to indicate that s/he wants something (e.g. a toy that is out of reach)?

- Always
- Usually
- Sometimes
- Rarely
- Never

6. Does your child point to share interest with you (e.g. pointing at an interesting sight)?

- Always
- Usually
- Sometimes
- Rarely
- Never

7. How long can you child's interest be maintained by a spinning object (e.g. washing machine, electric fan, toy car wheels)?
- Always
 - Usually
 - Sometimes
 - Rarely
 - Never
8. How many words can your child say?
- Always
 - Usually
 - Sometimes
 - Rarely
 - Never
9. Does your child pretend (e.g. care for dolls, talk on a toy phone)?
- Always
 - Usually
 - Sometimes
 - Rarely
 - Never
10. Does your child follow where you're looking?
- Always
 - Usually
 - Sometimes
 - Rarely
 - Never
11. How often does your child sniff or lick unusual objects?
- Always
 - Usually
 - Sometimes
 - Rarely
 - Never
12. Does your child place your hand on an object when s/he wants you to use it (e.g. on a door handle when s/he wants you to open the door, on a toy when s/he wants you to activate it)?
- Always
 - Usually
 - Sometimes
 - Rarely
 - Never

13. Does your child walk on tiptoe?

- Always
- Usually
- Sometimes
- Rarely
- Never

14. How easy is it for your child to adapt when his/her routine changes or when things are out of their usual place?

- Always
- Usually
- Sometimes
- Rarely
- Never

15. If you or someone else in the family is visibly upset, does your child show signs of wanting to comfort them (e.g. stroking their hair, hugging them)?

- Always
- Usually
- Sometimes
- Rarely
- Never

16. Does your child do the same thing over and over again (e.g. running the tap, turning the light switch on and off, opening and closing doors)?

- Always
- Usually
- Sometimes
- Rarely
- Never

17. Would you describe your child's first words as:

- Always
- Usually
- Sometimes
- Rarely
- Never

18. Does your child echo things s/he hears (e.g. things that you say, lines from songs or movies, sounds)?

- Always
- Usually
- Sometimes
- Rarely
- Never

19. Does your child use simple gestures (e.g. wave goodbye)?

- Always
- Usually
- Sometimes
- Rarely
- Never

20. Does your child make unusual finger movements near his/her eyes?

- Always
- Usually
- Sometimes
- Rarely
- Never

21. Does your child spontaneously look at your face to check your reaction when faced with something unfamiliar?

- Always
- Usually
- Sometimes
- Rarely
- Never

22. How long can your child's interest be maintained by just one or two objects?

- Always
- Usually
- Sometimes
- Rarely
- Never

23. Does your child twiddle objects repetitively (e.g. pieces of string)?

- Always
- Usually
- Sometimes
- Rarely
- Never

24. Does your child seem oversensitive to noise?

- Always
- Usually
- Sometimes
- Rarely
- Never

25. Does your child stare at nothing with no apparent purpose?

- Always
- Usually
- Sometimes
- Rarely
- Never

Official Use

Total	
--------------	--

Finally, please complete the last 2 questions:

(i) This form was completed by

Name: _____

Date: _____

(ii) What is your relationship with the child?

Mother

Father

Other

*If "Other", please specify (e.g. Grandmother) _____

Thank you for completing this questionnaire

Your help is greatly appreciated

Appendix 6: Systematic electronic literature search strategy

Search on MEDLINE using PubMed interface

Date of electronic search: 13/04/12

	Search terms	No of articles retrieved
1	("Infant, Premature"[MeSH]) preterm infant [tiab] OR preterm neonate [tiab] OR preterm children	37032
2	[tiab] premature infant [tiab] OR premature neonate [tiab] OR premature	3077
3	children [tiab] ("Infant, Very Low Birth Weight"[MeSH]) OR "very low birth weight"	4423
4	[tiab] OR "very low birthweight" [tiab]	9188
5	extremely low birth weight [tiab] OR "extremely low birthweight" [tiab]	1628
6	#1 OR #2 OR #3 OR #4 OR #5	46447
7	cogniti* [tiab]	171344
8	neurodevelopment* [tiab]	10989
9	"Developmental Disabilities"[Mesh] OR disability [tiab]	83996
10	mental retardation	88516
11	intelligence [tiab] OR IQ [tiab]	28057
12	#7 OR #8 OR #9 OR #10 OR #11	351189
13	#6 AND #12	3600
14	Limit to English-language	3257
15	Limit to publication date 01/01/1990 - 31/03/2012	2844

Appendix 7: Study 1 sensitivity analyses

Table S1 The sensitivities and specificities of the NHS assessment in identifying children with *any* impairment against the research assessment, using all singleton births and only one randomly selected child from each multiple birth set

Domain of development*	Method of classification of impairment for research assessment	Identification of any impairment by NHS assessment against the research assessment	
		Sensitivity, % (95% CI)	Specificity, % (95% CI)
Cognitive	Bayley-III scores	65.5 (49.7 - 81.3)	84.0 (75.5 - 92.5)
Receptive communication	Bayley-III scores	25.0 (5.7 - 44.3)	97.0 (94.2 - 99.8)
	NPEU/Oxford	42.1 (9.2 - 75.1)	96.7 (94.3 - 99.1)
Expressive communication	Bayley-III scores	60.8 (48.4 - 73.2)	90.9 (83.8 - 98.0)
	NPEU/Oxford	48.6 (33.9 - 63.4)	93.9 (88.1 - 99.6)
Combined language	Bayley-III scores	54.4 (44.2 - 64.6)	90.9 (84.2 - 97.6)
	NPEU/Oxford	49.3 (34.8 - 63.8)	93.9 (88.1 - 99.6)
Fine motor	Bayley-III scores	27.3 (0.0 - 64.7)	99.4 (98.1 - 100.0)
	NPEU/Oxford	Not estimable	Not estimable
Gross motor	Bayley-III scores	73.3 (46.5 - 100.0)	99.4 (97.9 - 100.0)
	NPEU/Oxford	66.7 (42.0 - 91.4)	98.7 (96.8 - 100.0)
Combined motor	Bayley-III scores	63.2 (34.3 - 92.1)	99.3 (98.0 - 100.0)
	NPEU/Oxford	68.8 (45.6 - 91.9)	98.7 (96.8 - 100.0)
Overall	Bayley-III scores	60.7 (50.0 - 71.3)	86.3 (79.3 - 93.3)

*Combined language impairment was judged as the worst category of outcome from receptive communication and expressive communication; combined motor impairment was judged as the worst category of outcome from fine motor and gross motor. Overall impairment was based on the worst category of outcome from the cognitive, language and motor domains.

Table S2 The sensitivities and specificities of the NHS assessment in identifying children with severe impairment against the research assessment, using all singleton births and only one randomly selected child from each multiple birth set

Domain of development*	Method of classification of impairment for research assessment	Identification of severe impairment by NHS assessment against the research assessment	
		Sensitivity, % (95% CI)	Specificity, % (95% CI)
Cognitive	Bayley-III scores	28.6 (5.0 - 52.2)	98.8 (97.1 - 100.0)
Receptive communication	Bayley-III scores	30.0 (0.0 - 68.2)	98.7 (97.0 - 100.0)
	NPEU/Oxford	Not estimable	Not estimable
Expressive communication	Bayley-III scores	58.3 (36.6 - 80.0)	96.3 (93.2 - 99.3)
	NPEU/Oxford	36.0 (13.7 - 58.3)	97.3 (94.8 - 99.8)
Combined language	Bayley-III scores	45.0 (17.2 - 72.8)	96.6 (93.2 - 100.0)
	NPEU/Oxford	40.9 (18.4 - 63.4)	96.6 (94.0 - 99.2)
Fine motor	Bayley-III scores	50.0 (30.2 - 100.0)	99.4 (98.2 - 100.0)
	NPEU/Oxford	Not estimable	Not estimable
Gross motor	Bayley-III scores	Not estimable	Not estimable
	NPEU/Oxford	Not estimable	Not estimable
Combined motor	Bayley-III scores	Not estimable	Not estimable
	NPEU/Oxford	Not estimable	Not estimable
Overall	Bayley-III scores	52.2 (24.2 - 80.1)	96.4 (92.9 - 99.9)

*Combined language impairment was judged as the worst category of outcome from receptive communication and expressive communication; combined motor impairment was judged as the worst category of outcome from fine motor and gross motor. Overall impairment was based on the worst category of outcome from the cognitive, language and motor domains.

Appendix 8: Studies included in systematic review (study 3)

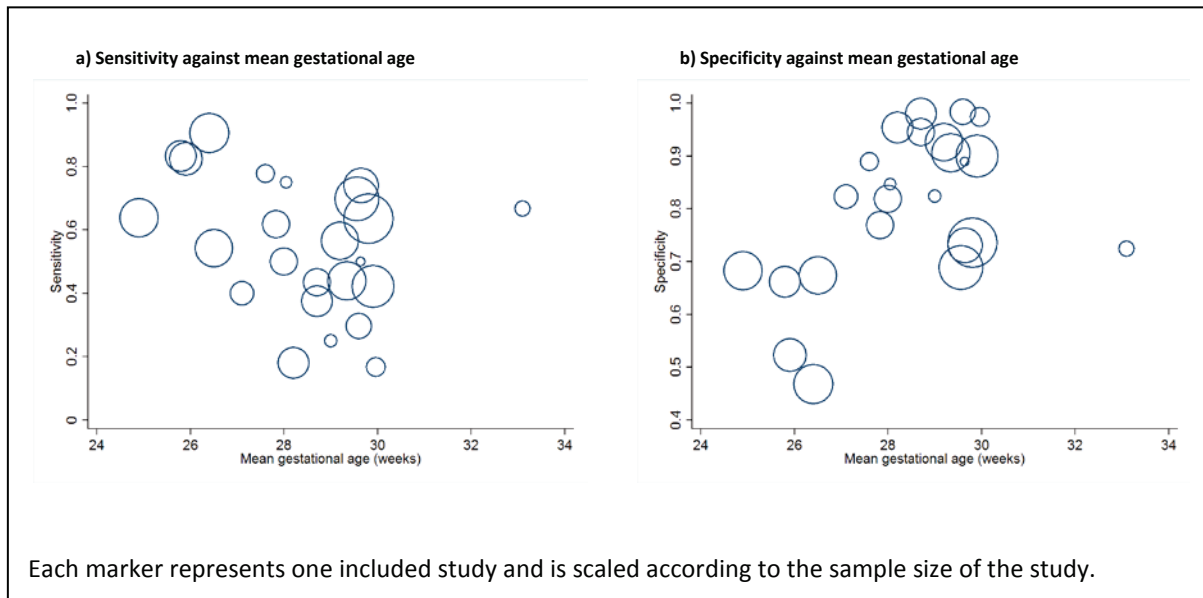
Author and year of publication used to denote study	Reference of published articles
Bassan, 2011	Bassan H, Stolar O, Geva R, Eshel R, Fattal-Valevski A, Leitner Y, Waron M, Jaffa A, Harel S. Intrauterine growth-restricted neonates born at term or preterm: how different? <i>Pediatr Neurol</i> 2011;44:122
Bowen, 1996	Bowen JR, Gibson FL, Leslie GI, Arnold JD, Ma PJ, Starte DR. Predictive value of the Griffiths assessment in extremely low birthweight infants. <i>J Paediatr Child Health</i> 1996;32:25
Bruggink, 2010	Bruggink JLM, Van Braeckel KN, Bos AF. The early motor repertoire of children born preterm is associated with intelligence at school age. <i>Pediatrics</i> 2010;125:1356-63
Charkaluk, 2011	Charkaluk ML, Truffert P, Marchand-Martin L, Mur S, Kaminski M, Ancel PY, Pierrat V for Epipage study group. Very preterm children free of disability or delay at age 2: predictors of schooling at age 8: a population-based longitudinal study. <i>Ear Hum Dev</i> 2011;87:297
Claas, 2011	Claas MJ, de Vries LS, Bruinse HW, van Haastert IC, Uniken Venema MMA, Peelen LM, Koopman C. Neurodevelopmental outcome over time of preterm born children <750g at birth. <i>Ear Hum Dev</i> 2011;87:183.
Cohen, 1995	Cohen SE. Biosocial factors in early infancy as predictors of competence in adolescents who were born prematurely. <i>J Dev Behav Pediatr</i> 1995;16:36
Fedrizzi, 1993	Fedrizzi E, Inverno M, Botteon G, Anderloni A, Filippini G, Farinotti M. The cognitive development of children born preterm and affected by spastic diplegia. <i>Brain Dev</i> 1993;15:428
Gray, 1995	(i) Gray PH, Burns YR, Mohay HA, O'Callaghan MJ, Tudehope DI. Neurodevelopmental outcome of preterm infants with bronchopulmonary dysplasia. <i>Arch Dis Child Fetal Neonatal Ed</i> 1995; 73:F128 (ii) Gray PH, O'Callaghan MJ, Rogers YM. Psychoeducational outcome at school age of preterm infants with bronchopulmonary dysplasia. <i>J Paediatr Child Health</i> . 2004 Mar;40(3):114-20
Gray, 2006	(i) 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. <i>J Paediatr Child Health</i> . 2006 Jun;42(6):377-83 (ii) Pritchard VE, Clark CA, Liberty K, Champion PR, Wilson K, Woodward LJ. Early school-based learning difficulties in children born very preterm. <i>Early Hum Dev</i> . 2009 Apr;85(4):215-24
Hack, 2005	Hack M, Taylor HG, Drotar D, Schluter M, Cartar L, Wilson-Costello W, Klein N, Friedman H, Mercuri-Minich N, Morrow M. Poor predictive validity of the Bayley Scales of Infant Development for cognitive function of ELBW children at school age. <i>Pediatrics</i> 2005;116:333

Kilbride, 1990	(i) Kilbride HW, Daily DK, Claflin K, Hall RT, Maulik D, Grundy HO. Improved survival and neurodevelopmental outcome for infants less than 801 grams birthweight. <i>Am J Perinatol</i> 1990;7:160 (ii) Kilbride HW, Daily DK. Survival and subsequent outcome to five years of age for infants with birthweight less than 801 grams born from 1983 to 1989. <i>J Perinatol</i> 1998;18:102-6
Marlow, 2005	(i) Marlow N, Wolke D, Bracewell MA, Methanna S for the EPICure Study Group. Neurologic and developmental disability at six years of age after extremely preterm birth. <i>NEJM</i> 2005;352:9 (ii) Johnson S, Fawke J, Hennessy E, Rowell V, Thomas S, Wolke D, Marlow N. Neurodevelopmental disability through 11 years of age in children born before 26 weeks of gestation. <i>Pediatrics</i> 2009;124:e249-e257
McGrath, 2000	McGrath MM, Sullivan MC, Lester BM, Oh W. Longitudinal neurologic follow-up in neonatal intensive care unit survivors with various neonatal morbidities. <i>Pediatrics</i> 2000;106:1397
Munck, 2012	Munck P, Niemi P, Lapinleimu H, Lehtonen L, Haataja L and the PIPARI Study Group. Stability of cognitive outcome from 2 to 5 years of age in very low birth weight children. <i>Pediatrics</i> 2012;129:503
Orchinik, 2011	Orchinik LJ, Taylor HG, Espy KA, Minich N, Klein N, Sheffield T, Hack M. Cognitive Outcomes for extremely preterm/extremely low birth weight children in kindergarten. <i>J Int Neuropsychol Soc</i> 2011;17:1067-79
Potharst, 2011	Potharst ES, Houtzager BA, van Sonderen L, Tamminga P, Kok JK, van Wassenaer AG. 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. <i>Dev Med Child Neuro</i> 2011;54:240
Reuss, 1996	(i) 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. <i>N Engl J Med</i> 1996;334:821 (ii) 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 performances at ages 2, 6 and 9 years. <i>Dev Med Child Neurol</i> 1999;41(12):826-33 (iii) Pinto-Martin J, Whitaker A, Feldman J, Cnaan A, Zhao H, Bloch JR, McCulloch D, Paneth N. Special education services and school performance in a regional cohort of low-birthweight infants at age nine. <i>Paediatr Perinat Epidemiol.</i> 2004 Mar;18(2):120-9 (iv) Lorenz JM, Whitaker AH, Feldman JF, Yudkin PL, Shen S, Blond A, Pinto-Martin JA, Paneth N. 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. <i>J Dev Behav Pediatr</i> 2009; 30(6):535-43
Roberts, 2010	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. <i>Arch Dis Child</i> 2010;95:786
Skranes, 1998	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? <i>Dev Med Child Neurol</i> 1998;40:256
Smith, 2006	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. <i>Pediatrics</i> 2006;117:1608

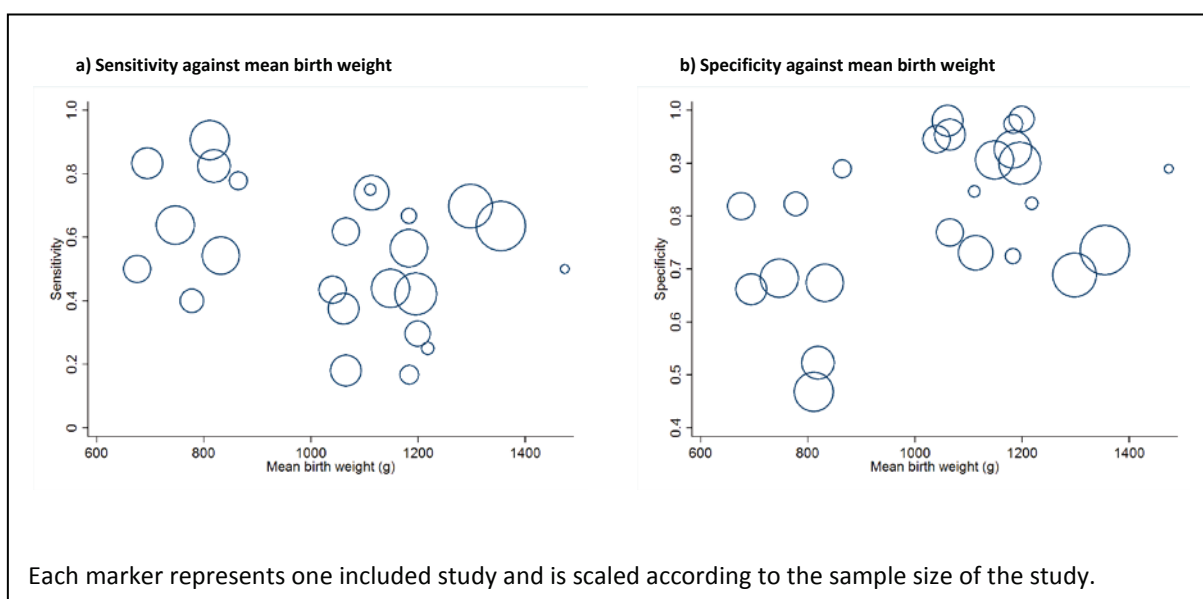
Tommiska, 2003	<p>(i) Tommiska V, Heinonen K, Kero P, Pokela ML, Tammela O, Järvenpää AL, Salokorpi T, Virtanen M, Fellman V. A national two year follow up study of extremely low birthweight infants born in 1996-1997. Arch Dis Child Fetal Neonatal Ed. 2003 Jan;88(1):F29-35</p> <p>(ii) Mikkola K, Ritari N, Tommiska V, Salokorpi T, Lehtonen L, Tammela O, Pääkkönen L, Olsen P, Korkman M, Fellman V. Neurodevelopmental outcome at 5 years of age of a national cohort of extremely low birth weight infants who were born in 1996-1997. Pediatrics. 2005 Dec;116(6):1391-400</p>
Veelken, 1991	<p>(i) Veelken N, Stollhoff K, Calussen M. Development of very low birth weight infants: a regional study of 371 survivors. Eur J Pediatr 1991; 150:815</p> <p>(ii) Dammann O, Drescher J, Veelken N. Maternal fever at birth and non-verbal intelligence at age 9 years in preterm infants. Dev Med Child Neurol. 2003 Mar;45(3):148-51</p>
Vermeulen, 2001	<p>(i) Vermeulen GM, Bruinse HW, de Vries LS. Perinatal risk factors for adverse neurodevelopmental outcome after spontaneous preterm birth. Eur J Obstet Gynecol Reprod Biol 2001;99(2):207-12</p> <p>(ii) Rademaker KJ, Uiterwaal CS, Groenendaal F, Venema MM, van Bel F, Beek FJ, van Haastert IC, Grobbee DE, de Vries LS. Neonatal hydrocortisone treatment: neurodevelopmental outcome and MRI at school age in preterm-born children. J Pediatr 2007; 150(4):351-7</p>
Wolke, 1999	<p>(i) 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 2000;82:F208</p> <p>(ii) 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. 1999 Feb;41(2):94-109</p> <p>(iii) Saigal S, den Ouden L, Wolke D, Hoult L, Paneth N, Streiner DL, Whitake A, Pinto-Martin J. School-age outcomes in children who were extremely low birth weight from four international population-base cohorts. Pediatrics 2003;112:948-50</p> <p>(iv) Wolke D, Schulz J & Meyer R. Longterm developmental outcome of very prematurely born infants. Bavarian longitudinal study. Monatsschrift Kinderheilkunde 2001;149:S53-S61</p>

Appendix 9: Scatterplots showing the relationship of study-level variables with diagnostic validity (study 3)

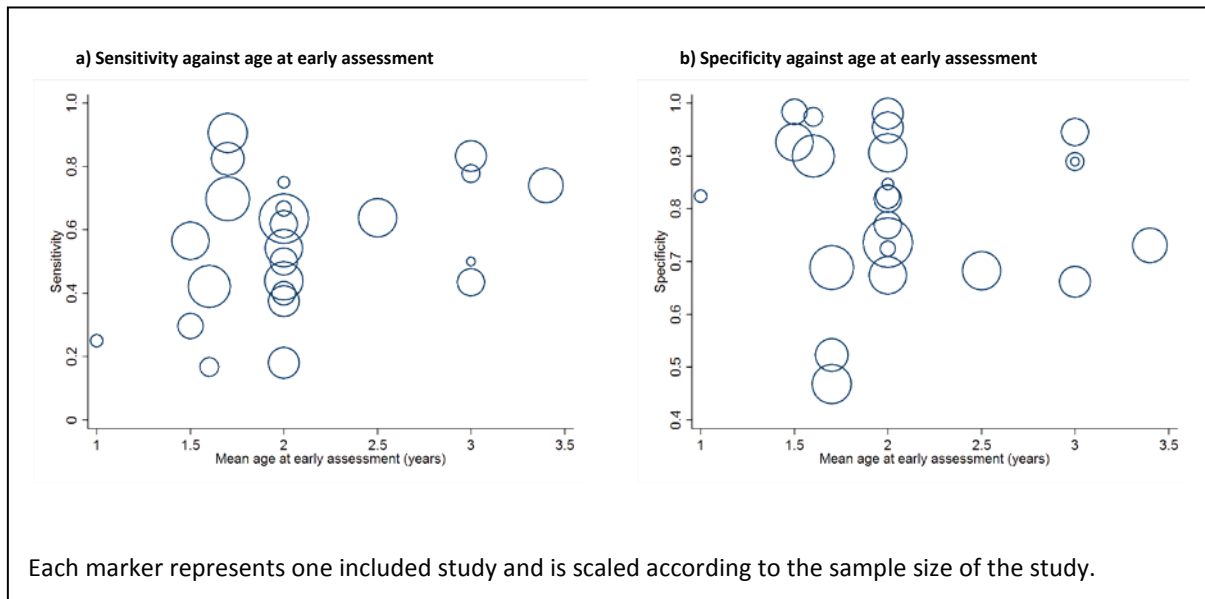
Figures S1a and S1b Scatterplot of the (a) sensitivity and (b) specificity reported in each study against the mean gestational age of the study population



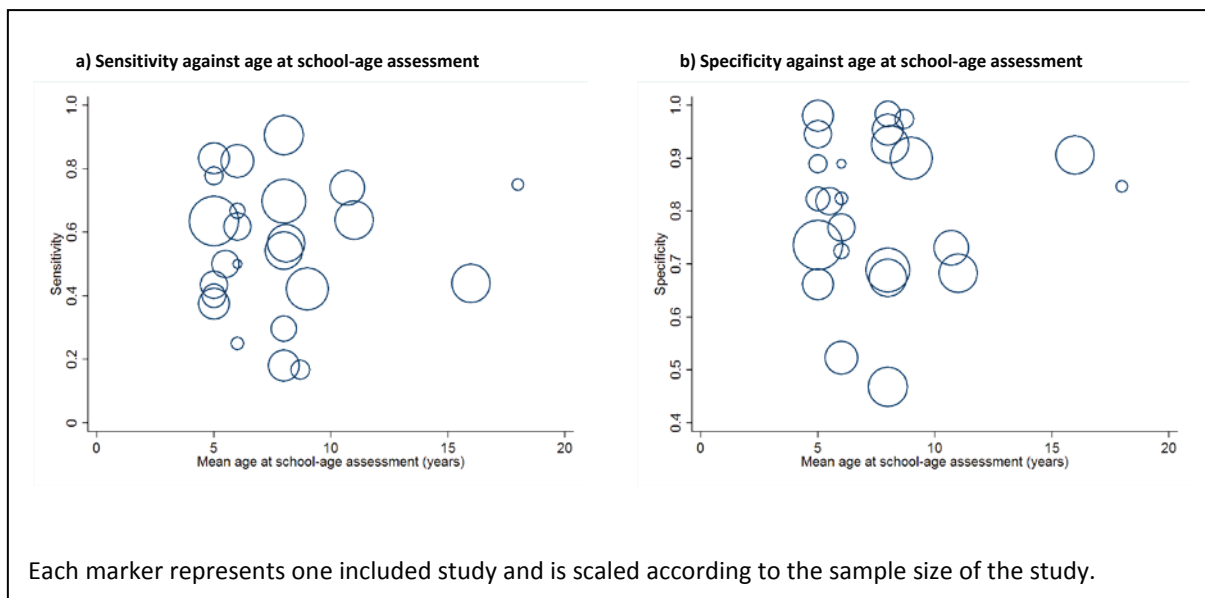
Figures S2a and S2b Scatterplot of the (a) sensitivity and (b) specificity reported in each study against the mean birth weight of the study population



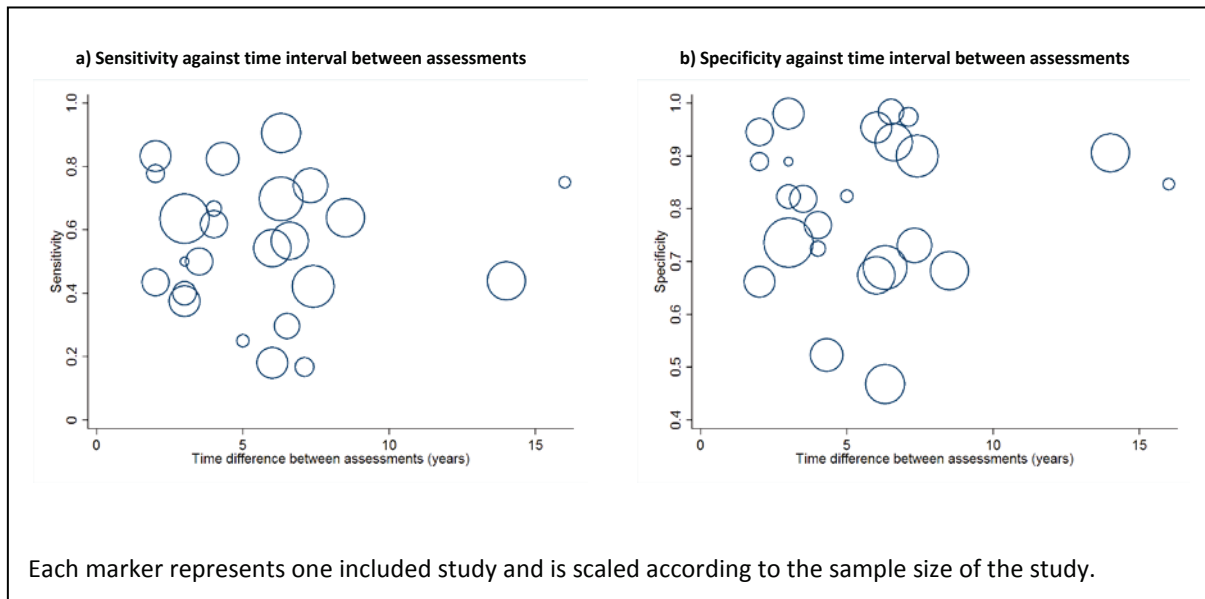
Figures S3a and S3b Scatterplot of the (a) sensitivity and (b) specificity reported in each study against the mean age at early developmental assessment of the study population



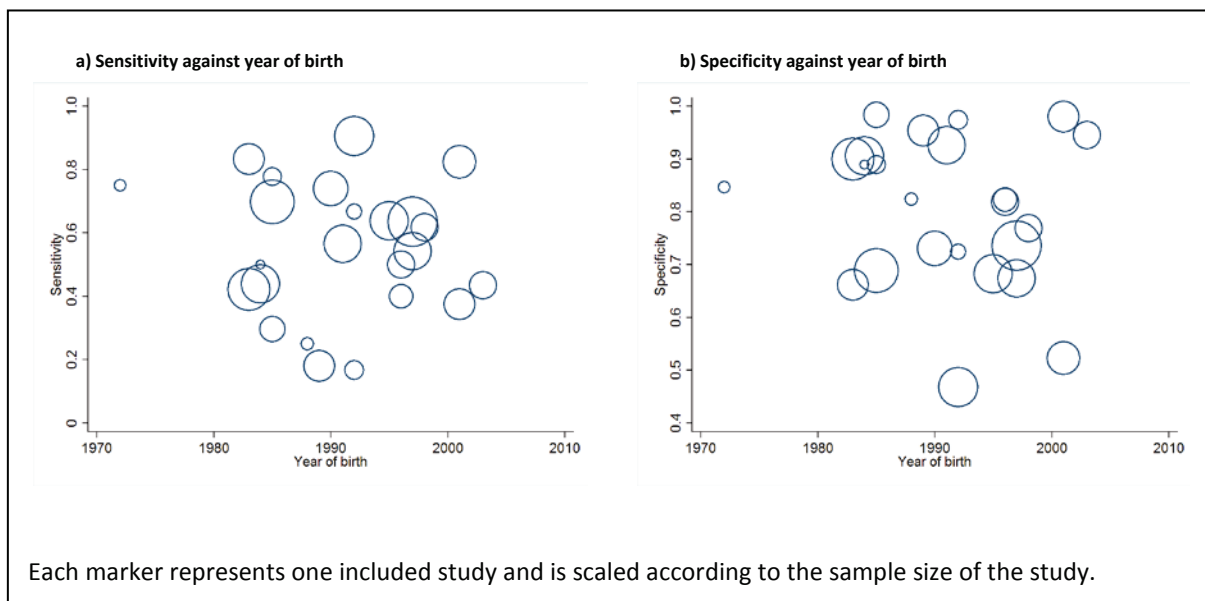
Figures S4a and S4b Scatterplot of the (a) sensitivity and (b) specificity reported in each study against the mean age at school-age cognitive assessment of the study population



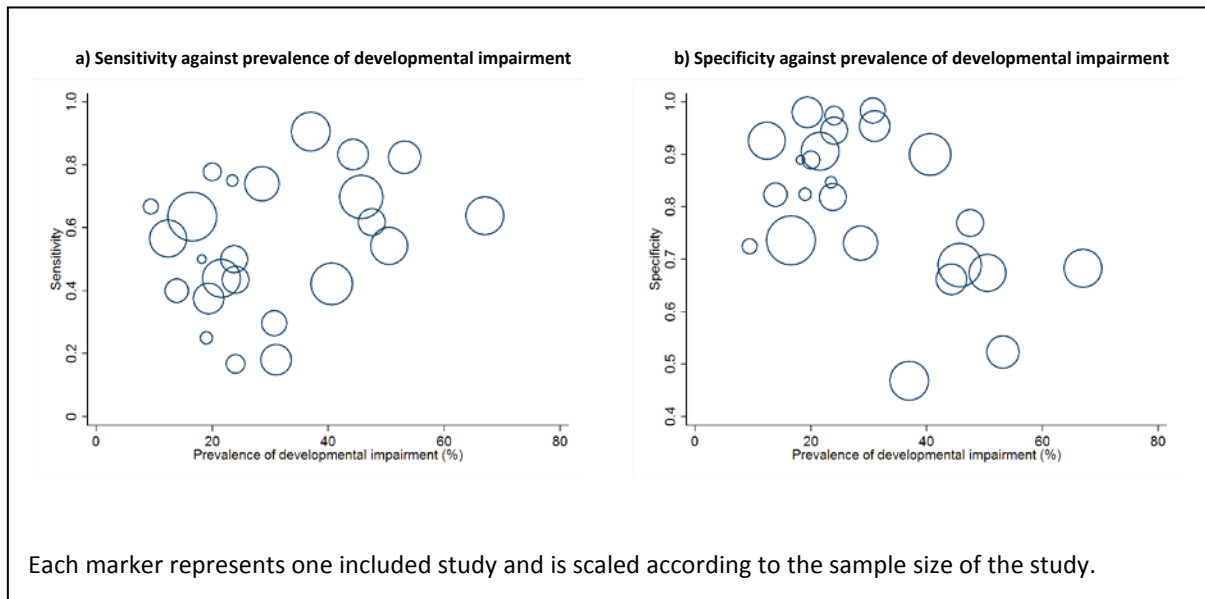
Figures S5a and S5b Scatterplot of the (a) sensitivity and (b) specificity reported in each study against the mean time difference between assessments of the study population



Figures S6a and S6b Scatterplot of the (a) sensitivity and (b) specificity reported in each study against the year of birth of the study population



Figures S7a and S7b Scatterplot of the (a) sensitivity and (b) specificity reported in each study against the prevalence of developmental impairment in the study population



Figures S8a and S7b Scatterplot of the (a) sensitivity and (b) specificity reported in each study against the prevalence of *severe* development impairment in the study population

