ARTICLE OPEN

Check for updates

Use of term reference infants in assessing the developmental outcome of extremely preterm infants: lessons learned in a multicenter study

Charles E. Green[1]✉, Jon E. Tyson[1], Roy J. Heyne[2], Susan R. Hintz[3], Betty R. Vohr[4], Carla M. Bann[5], Abhik Das[6], Edward F. Bell[7], Sana Boral Debsareea[8], Emily Stephens[1], Marie G. Gantz[5], Carolyn M. Petrie Huitema[6], Karen J. Johnson[7], Kristi L. Watterberg[9], Ricardo Mosquera[10], Myriam Peralta-Carcelen[11], Deanne E. Wilson-Costello[12], Tarah T. Colaizy[7], Nathalie L. Maitre[13], Stephanie L. Merhar[14], Ira Adams-Chapman[15,42], Janell Fuller[16], Michelle E. Hartley-McAndrew[17], William F. Malcolm[18], Sarah Winter[19], Andrea F. Duncan[20], Gary J. Myer[21], Stephen D. Kicklighter[22], Myra H. Wyckoff[2], Sara B. DeMauro[23], Anna Maria Hibbs[12], Barbara J. Stoll[1], Waldemar A. Carlo[11], Krisa P. Van Meurs[3], Matthew A. Rysavy[7], Ravi M. Patel[15], Pablo J. Sánchez[13], Abbot R. Laptook[4], C. Michael Cotten[18], Carl T. D'Angio[22], Michele C. Walsh[24] and Eunice Kennedy Shriver National Institute of Child Health and Human Development Neonatal Research Network*

OBJECTIVE: Extremely preterm (EP) impairment rates are likely underestimated using the Bayley III norm-based thresholds scores and may be better assessed relative to concurrent healthy term reference (TR) infants born in the same hospital.
STUDY DESIGN: Blinded, certified examiners in the Neonatal Research Network (NRN) evaluated EP survivors and a sample of healthy TR infants recruited near the 2-year assessment age.
RESULTS: We assessed 1452 EP infants and 183 TR infants. TR-based thresholds showed higher overall EP impairment than Bayley norm-based thresholds (O.R. = 1.86; [95% CI 1.56–2.23], especially for severe impairment (36% vs. 24%; $p \le 0.001$).Difficulty recruiting TR patients at 2 years extended the study by 14 months and affected their demographics.
CONCLUSION: Impairment rates among EP infants appear to be substantially underestimated from Bayley III norms. These rates may be best assessed by comparison with healthy term infants followed with minimal attrition from birth in the same centers.
CLINICALTRIALS.GOV ID: Term Reference (under the Generic Database Study): NCT00063063

Journal of Perinatology (2023) 43:1398–1405; https://doi.org/10.1038/s41372-023-01729-x

INTRODUCTION

Follow-up assessments of extremely preterm (EP) infants are difficult to perform and interpret for multiple reasons. As for other assessments [1], the expectations or biases of unblinded examiners may have an important effect on the findings. This problem can be minimized by including a concurrently assessed reference group of term infants and assuring that the examiners are masked to gestational age, perinatal complications, and findings of any prior follow-up assessments [2–5].

Another issue is the appropriate comparison group of term infants. One approach is to compare EP infants to term infants matched for maternal age, ethnicity, income, education, marital status, insurance status, etc. This approach has been used in efforts to identify the independent effects of perinatal factors on

[1]Department of Pediatrics, McGovern Medical School at The University of Texas Health Science Center at Houston, Houston, TX, USA. [2]Department of Pediatrics, University of Texas Southwestern Medical Center, Dallas, TX, USA. [3]Department of Pediatrics, Division of Neonatal and Developmental Medicine, Stanford University School of Medicine and Lucile Packard Children's Hospital, Palo Alto, CA, USA. [4]Department of Pediatrics, Women & Infants Hospital, Brown University, Providence, RI, USA. [5]Social, Statistical and Environmental Sciences Unit, RTI International, Research Triangle Park, Greensboro, NC, USA. [6]Social, Statistical and Environmental Sciences Unit, RTI International, Rockville, MD, USA. [7]Department of Pediatrics, University of Iowa, Iowa City, IA, USA. [8]Center for Clinical Research and Evidence-Based Medicine, University of Texas Houston McGovern Medical School, Houston, TX, USA. [9]University of New Mexico Health Sciences Center, Albuquerque, NM, USA. [10]Children's Memorial Hermann Hospital, Houston, TX, USA. [11]Division of Neonatology, University of Alabama at Birmingham, Birmingham, AL, USA. [12]Department of Pediatrics, Rainbow Babies & Children's Hospital, Case Western Reserve University, Cleveland, OH, USA. [13]Department of Pediatrics, Nationwide Children's Hospital, The Ohio State University College of Medicine, Columbus, OH, USA. [14]Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, USA. [15]Emory University School of Medicine, Department of Pediatrics, Children's Healthcare of Atlanta, Atlanta, GA, USA. [16]Department of Pediatrics, University of New Mexico, Albuquerque, NM, USA. [17]Department of Pediatrics, University of Buffalo, Buffalo, NY, USA. [18]Department of Pediatrics, Duke University, Durham, NC, USA. [19]Department of Pediatrics, Division of Neonatology, University of Utah School of Medicine, Salt Lake City, UT, USA. [20]Department of Pediatrics, Pennsylvania Hospital, Philadelphia, PA, USA. [21]University of Rochester School of Medicine and Dentistry, Rochester, NY, USA. [22]Department of Pediatrics, Wake Medical Center, Raleigh, NC, USA. [23]Department of Pediatrics, University of Pennsylvania, Philadelphia, PA, USA. [24]Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD, USA. [42]Deceased: Ira Adams-Chapman. *A list of authors and their affiliations appears at the end of the paper. ✉email: Charles.Green@uth.tmc.edu

Received: 27 July 2022 Revised: 17 March 2023 Accepted: 10 July 2023
Published online: 4 August 2023

outcomes. However, matching is logistically difficult, quite likely to be incomplete, and precludes assessment of how adverse socioeconomic factors and their interactions with biological or medical factors compromise the outcomes of EP infants. A better understanding of all these factors is needed to develop improved methods to reduce rates of impairment among EP born children. For these reasons, a comparison to healthy term infants may be preferred in deciding which EP infants should be considered to have a developmental impairment based on the child's capabilities irrespective of the extent to which these impairments result from medical, socioeconomic, or other factors [6].

Additional issues include the choice of the developmental test and whether its norms are fully appropriate in designating which EP infants should be considered impaired [2, 7–11]. While the Bayley Scales of Infant and Toddler Development (Bayley III) have been widely used, multiple investigators have reported that the impairment rates are likely to be underestimated in applying its norms [2, 7–14]. Moreover, it is difficult to assure that the examiners in all centers perform the Bayley III assessments in the same way that the assessments were performed when the Bayley III was normed.

For all these reasons, the NICHD Neonatal Research Network (NRN) undertook the study described below to assess EP and a concurrent sample of healthy term reference (TR) infants examined by the same blinded and certified examiners in the same centers at two years corrected age. We hypothesized that the proportion of EP infants with developmental impairment based on standard deviations (SDs) from the mean for the TR sample would be higher than that based on Bayley III norms. If so, we hoped to identify threshold values for Bayley III scores based on our term reference infants that would be more appropriate than those based on the Bayley III norms for categorizing EP infants as impaired in NRN centers.

METHODS

The study was conducted in 15 NRN centers between January 2017 and March 2020. The addition of the TR infants to the follow-up assessments was approved by each center's Institutional Review Board (IRB). Consent was obtained in accordance with each study site's IRB requirements.

Design

To augment the reliability of the assessments and reduce the likelihood that examiner expectations would affect the scores, the TR and EP infants in the study were assessed concurrently by examiners not informed of their gestational age at birth or their prior clinical or developmental findings.

Eligibility and sampling

The eligible EP infants were inborn at NICHD NRN centers and <27 weeks gestation by best obstetric estimate. Infants with at least one Bayley III composite score at the 24-month follow-up visit were included in the analysis.

Eligible TR infants met the following criteria assessed using the medical record: singleton birth at 39 0/7-40 6/7 weeks gestation by best obstetric estimate; birth weight appropriate for gestational age; no resuscitation at birth; absence of congenital anomalies or other abnormalities on physical examination; benign neonatal course with all care given in a low risk nursery and no neonatal problem delaying discharge home; and parent(s) willing and able to come into the clinic. Exclusion criteria included major central nervous system disorder (e.g., cerebral palsy, deafness, blindness or the effects of major insults identified by parent report or medical records [e.g. meningitis or traumatic brain injury before two years]), child protective services custody, parental incarceration, and parental psychosis.

Our goal was to assess one healthy term infant for every fifth EP survivor at 22–26 months corrected for prematurity in the same center to evaluate 180 total TR infants in a one-year study. (See Sample Size and Power.) Recruitment of each healthy TR infant began shortly (e.g. 1–2 months) before the corresponding EP infant's scheduled assessment. If the EP infant was lost to follow-up, the TR infant was still to be assessed.

Center coordinators used medical records to identify and attempt to recruit the first healthy term infant born on or after the expected due date of the index EP infant. The potential value of developmental testing was emphasized in recruiting. The methods of contact (letter, text, phone call) and incentives used to promote participation (e.g. up to $100 plus parking or $50 plus cab fare) varied as allowed by the individual site's IRB. When a parent or guardian declined participation or missed two scheduled clinic visits, the coordinator contacted the next eligible infant's parent or guardian by delivery time and date until one agreed for her child to participate within the testing window.

To assess the representativeness of the TR sample with all term births in the NRN centers we requested the information for all term infants born in NRN hospitals during the study period. To further characterize the sample of TR infants we qualitatively contrasted the estimates on available data from the Bayley-III normative data.

Measurement and comparisons

Certified Bayley III examiners, trained to reliability and re-evaluated annually, provided assessments at each NRN center [15, 16]. The Bayley III was administered to Spanish-speaking children by either a Spanish-speaking evaluator or an English-speaking evaluator with a translator. Means and SD's of Bayley III scores among the TR infants were used to determine new thresholds for each of the Bayley III composites (cognitive, language, and motor) to indicate three levels of impairment: (1) Normal/mild, a score greater than or equal to 1 SD below the mean; (2) Moderate, a score between one and two SD below the mean; and (3) Severe, a score lower than 2 SD below the mean. Application of these new cut points to the EP infants determined the proportion falling into each category [15, 16].

Statistical analysis

Generalized linear multilevel models compared the proportion of infants in each category using the norm-based vs. TR thresholds, accounting for clustering of infants within centers. Levels of impairment were analyzed using an ordinal logistic model and dichotomous variables (moderate/severe vs. normal/mild) were analyzed using a binomial model. Analyses were conducted using SAS version 9.3.

Sample size and power

Assuming a 5% rate of impairment based on Bayley III manual norms, a 15% rate of impairment based on thresholds derived from the reference group [2], and an intraclass correlation of 0.05 due to center membership, a sample of $N = 180$ provided 91% power to detect a 10% absolute difference in impairments > 2 S.D.'s below the mean with alpha = 0.05. Given prior, annual rates of enrollment for EP infants we anticipated that recruiting EP to TR in a 1:5 ratio would result in $N = 180$ within one year.

RESULTS

A total of 1452 EP infants (86% of survivors at 2 years) were evaluated during the time required to accrue and successfully assess 183 TR infants (Fig. 1). This accrual of TR infants took longer than expected (38 versus 24 months with an accrual ratio of 1:8 versus 1:5. Based on querying site coordinators, reasons for slower accrual than expected varied among centers but included difficulty accessing the medical records in some hospitals that were not owned by the university, problems contacting the parents using letters (as required by some IRBs), variable incentives for participation allowed by the IRBs, parental inconvenience, transportation problems, and in one center, contract negotiations between the university and an affiliated hospital.

Demographic comparison of the TR sample and EP infants

Mothers of TR infants were more often White, married and more highly educated. Mothers of EP infants were more often African-American (Table 1).

Demographic comparison of the TR sample, term births at NRN hospitals and the Bayley III normative population

The information that NRN hospitals provided about their term births was incomplete and varied between hospitals, resulting in

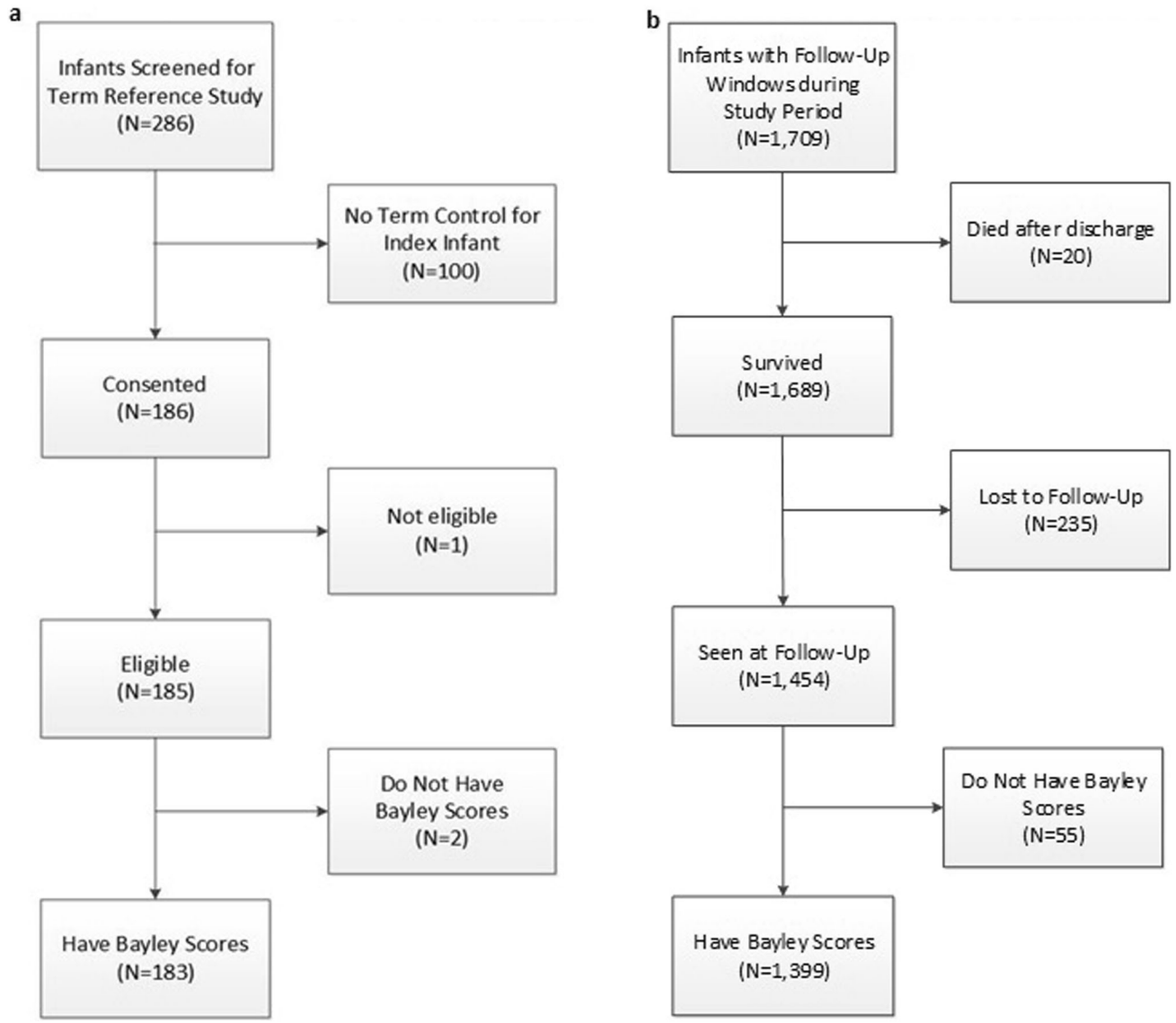

Fig. 1 Sampling diagram. Sample selection for term-reference (**a**) and pre-term (**b**) infants.

uncertainty in how the TR sample differed from all children born at term in these centers. Since the TR group included only healthy infants, modest differences would be expected. However, in the 11 centers where the information was provided (Table 2), there were 30% fewer TR children with Medicaid/public insurance and 24% more with private insurance.

The data for our TR sample were compared with the data provided for the normative Bayley III sample at age two years gathered by the test company (i.e. $n = 100$ children at 24 months). The TR sample differed from the Bayley normative sample with respect to percent who were Hispanic (20 vs 16%), African-American (28% vs 14%) and parents with ≥ 16 years of education (51% vs. 29%). Surprisingly, the Bayley III Technical Manual did not characterize the normative sample in terms of marital or insurance status, did not report the proportion of children approached for inclusion who did not participate, or indicate any measures to blind the evaluators to any unfavorable social, medical, or biologic factors that might influence scores [17].

Bayley III scores for TR and EP infants

The mean composite cognitive, motor, and language scores were 83.9, 83.3, and 80.2, respectively, for the EP infants and 97.5, 98.2, and 97.9, respectively for the TR group (Table 3). As expected, with the deliberate inclusion of children with developmental problems in the Bayley normative sample, the SDs were less for our healthy TR sample for the Cognitive Composite (11.2, 95% CI 10.2–12.5) and the Motor Composite (10.9, 95% CI 9.9–12.2) than for the Bayley normative sample (SD = 15 for all composites). The SD for the Language Composite in the TR sample was 16.0 (95% CI 14.5–17.9), similar to the manual-based SD (15). The composite score SDs for the EP infants ranged from 15.1–17.4.

The ranges for all three Bayley composite scores based on norm-based thresholds were ≥ 85, 70–84 and 55–69 respectively for all three Bayley III composite scores. Using term-reference data resulted in ranges for normal/mild, moderate and severe thresholds of ≥86.21, 75–86.20 and 63.73–74.97 for the Cognitive Composite, ≥87.31, 76.38–87.30 and 65.45–76.37 for the Motor Composite, and ≥81.91, 65.88–81.90 and 49.85–65.87 for the Language Composite. The Bayley III score thresholds for severe impairment (< 2 SDs below the mean) for Cognitive and Motor Composites were thus were 5-6 points higher than for the Bayley normative sample. However, the Language Composite threshold was approximately 3 points lower.

Table 1. Sociodemographic and medical characteristics of term and preterm infants.

Characteristic	Term ($N = 183$) *N* (%)	Preterm ($N = 1452$) *N* (%)
Maternal		
Maternal age…mean (SD)	30.13 (5.9)	28.72 (6.1)
Race		
African American	51 (28)	580 (40)
White	115 (63)	742 (51)
American Indian/Alaskan Native	0 (0)	11 (1)
Asian, Native Hawaiian, or Other Pacific Islander	5 (3)	44 (3)
More than One Race	4 (2)	30 (2)
Unknown	8 (4)	45 (3)
Ethnicity		
Hispanic or Latino	37 (20)	239 (16)
Not Hispanic or Latino	145 (79)	1199 (83)
Unknown	1 (1)	14 (1)
Gravidity…mean (SD)	2.56 (1.6)	2.96 (2.2)
Parity…mean (SD)	1.91 (1.1)	2.20 (1.4)
Marital status		
Married	108 (59)	647 (45)
Not married	74 (40)	801 (55)
Unknown	1 (1)	4 (0)
Education		
8th grade or less	2 (1)	41 (3)
9th to 12th grade	8 (4)	105 (7)
High school diploma	31 (17)	274 (19)
Trade or technical school	6 (3)	165 (11)
Partial college/associate degree	39 (21)	387 (27)
College degree	41 (22)	193 (13)
Graduate degree	53 (29)	93 (6)
Unknown	3 (2)	194 (13)
Neonatal		
Gestational age…mean (SD)	39.25 (0.6)	24.88 (1.1)
Birth weight…mean (SD)	3406.8 (359.8)	753.2 (165.1)
Sex		
Male	92 (50)	711 (49)
Female	91 (50)	741 (51)
Follow-Up		
GMFCS[a] level		
Normal/Level 0	167 (91)	961 (66)
Possible Level 1	0 (0)	16 (1)
Level 1	14 (8)	330 (23)
Level 2	0 (0)	77 (5)
Level 3	0 (0)	23 (2)
Level 4	0 (0)	16 (1)
Level 5	0 (0)	23 (2)
Unknown	2 (1)	6 (0)
Moderate/severe CP[a]		
Yes	0 (0)	110 (8)
No	181 (99)	1341 (92)
Unknown	2 (1)	1 (0)
Vision impairment (Bilateral blind with no/some functional vision)		
Yes	0 (0)	16 (1)
No	182 (99)	1436 (99)
Unknown	1 (1)	0 (0)
Hearing impairment (Any impairment with or without amplification)		
Yes	1 (1)	39 (3)
No	181 (99)	1375 (95)
Unknown	1 (1)	38 (3)

[a]*GMFCS* Gross Motor Function Classification System.
CP Cerebral Palsy.

Comparison of impairment rates

Term-reference-based impairment thresholds resulted in higher overall rates of moderate/severe impairment (i.e. impairment on *any* one of the Cognitive, Motor or Language Composites Scores) (Table 4 bottom). The same was true for impairment identified using just the Cognitive and Motor Composites. Given the larger, term-reference-estimated SD for the Language Composite, the norm-based thresholds resulted in higher rates of moderate/severe language impairment (Table 4). As evident in Table 4, the differences between the Manual and Term Reference based rates of moderate and severe impairment were largely to the difference in severe impairment. A second set of post-hoc analyses adjusting for maternal education, language spoken at home and age at assessment did not substantially alter these results.

DISCUSSION

We assessed Bayley III scores at two years adjusted age for EP infants and TR infants born in the same NRN centers and examined by the same assessors who had been trained to reliability [18] and were blinded to gestational age at birth, perinatal events, and prior follow-up findings. The mean composite cognitive, motor, and language scores were 83.9, 83.3, and 80.2, respectively, for the EP infants and 97.5, 98.2, and 97.9, respectively, for the TR group.

The mean Bayley III composite scores for our TR group were lower than for term control infants in some other studies [2, 3, 12] despite the high proportion of well-educated TR mothers. This finding may be due to greater socioeconomic disadvantages; our sample contained a higher proportion of Hispanic, African American, and Medicaid-insured children than the term controls in most other studies.

More EP infants had moderate or severe cognitive and motor impairments (composite scores more than 1 or 2 SDs below the mean, respectively) using the scores for TR sample (SD = 10.9–11.3) than the Bayley III normative sample (SD = 15.0). These differences are likely due in part to the different referent populations assessed. To avoid under-identification of impaired EP children, children with major congenital anomalies, perinatal problems, or postnatal insults likely to affect development [2, 3, 6] were systematically excluded from our TR sample. A different approach was used for the Bayley III normative sample, in which 10% of the children had such problems as Down's syndrome, cerebral palsy and language impairments [17]. While a reference population that includes the full spectrum of child development is desirable for some purposes [6], this approach would likely understate the proportion of impaired preterm infants when threshold scores 1 or 2 SDs below the mean for the Bayley III normative population are used to designate impairments. Accordingly, Sharp and DeMauro [7], among others, suggest that different and higher threshold Bayley III scores are needed.

As hypothesized, the overall proportion of EP infants with a cognitive, motor, or language impairment based on a composite score at least 1 SD below the mean for our TR group was higher than that based on Bayley III normative population (68 vs. 57%, $p < 0.01$)). The difference was particularly marked for severe impairment (one or more composite scores at least 2 SDs below the mean; 36 vs. 24%, $p \leq 0.001$). An unexpected finding was that the proportion of EP infants with composite language scores lower than 1 SD below the mean based on our TR sample was not higher than for the Bayley normative sample. This finding reflects a relatively high SD (16.0) for the TR language scores which may well be due to a high proportion of Hispanics and marked heterogeneity in parental education among the TR parents and a greater influence of education on language than on cognition or motor scores.

Table 2. Demographic characteristics for all term births at participating sites[a].

Variable (Number of centers)	Survey ($N = 101,422$)	Study ($N = 170$)
	n (%)	*n* (%)
Birth Weight (13)[b]	***n* = 100,277**	***n* = 155**
<1500 g	14 (0)	0 (0)
1501-2500 g	3005 (3)	0 (0)
2501-4000 g	89872 (90)	149 (95)
> 4000 g	7305 (7)	6 (4)
Sex (14)	***n* = 101,422**	***n* = 170**
Male	50631 (50)	84 (49)
Female	50778 (50)	86 (51)
Ambiguous	18 (0)	0 (0)
Race[c] (11)	***n* = 80,922**	***n* = 137**
White	39761 (49)	78 (57)
African American	28241 (35)	44 (32)
American Indian/ Alaskan Native	620 (1)	0 (0)
Asian	5051 (6)	5 (4)
Native Hawaiian or other Pacific Islander	316 (0)	0 (0)
More than one race	920 (1)	3 (2)
Unknown	5955 (7)	7 (5)
Ethnicity[c] (14)	***n* = 101,422**	***n* = 170**
Hispanic	26785 (26)	39 (23)
Non-Hispanic	66961 (66)	130 (76)
Unknown	7676 (8)	1 (1)
Insurance[d] (11)	***n* = 77,166**	***n* = 106**
Medicaid/public insurance	46358 (60)	32 (30)
Self-pay/uninsured	1969 (3)	9 (8)
Private	27461 (36)	64 (60)
Unknown	443 (1)	0 (0)
Other	735 (1)	1 (1)

[a]All analyses exclude one due to lack of data on demographics by gestational age.
[b]Birthweight analyses also exclude two due to missing data.
[c]Race analyses also exclude five centers due to missing data.
[d]Insurance analyses also exclude eight centers due to missing data.

Study limitations

The approach in most follow-up studies to assessing EP infants and designating their impairment rates involves some uncertainty about the reliability and inadvertent bias of the examiners as well as the appropriateness of the Bayley normative sample. While our study facilitated blinded Bayley III assessment of EP and TR infants by the same carefully trained and certified examiners, our sample of healthy TR infants was not sufficiently representative of healthy term infants in NRN centers to establish clear impairment thresholds for outcomes in the NRN. Our findings for insurance coverage and parental education indicate that attempts to recruit such infants two years or more after birth are difficult and prone to selection bias. Caregivers who had concerns about their child's development may have been more likely to participate, a problem that would cause us to underestimate the degree to which impairment rates were underestimated using Bayley norms. Future efforts to recruit a representative sample of healthy term infants may be more successful if these infants are enrolled in the neonatal period with special measures to maintain rapport with the parents and achieve high follow-up rates through the age of assessment [19].

The need to minimize bias in assessing EP infants may be achieved more simply by including a convenience sample of term reference controls and blinding the evaluators to gestational age, medical history, and any prior follow-up assessments. However, it is unclear whether the Bayley IV Scales address the need emphasized by Sharp and DeMauro [7] among others to establish higher impairment thresholds for the Bayley III Scales. While the Bayley-IV has superceded the Bayley-III the current results are still informative. The Bayley-IV Technical Manual states, "Because most of the Bayley-4 is a revision of the previous edition, most of the validity evidence reported in the research related to the Bayley-III is still relevant….". (p. 37) [20].

Accurate identification and monitoring of impairment rates in EP infants is critical for multiple reasons, including provision of appropriate services and parental counselling for individual infants, planning their long term education and rehabilitation, testing perinatal interventions in proper clinical trials, and evaluating care and outcomes within and across different perinatal centers over time. The impairment rates identified among EP and other high-risk infants have been almost always

Table 3. Bayley III scores among term reference (TR) and extremely preterm (EP) infants.

	Term Infants			Pre-Term Infants		
	N	Mean	SD	*N*	Mean	SD
Cognitive Composite Score	182	97.5	11.2	1446	83.9	15.1
Language Composite Score	180	97.9	16.0	1409	80.2	17.4
Expressive Language Scaled Score	180	9.5	3.0	1391	6.7	3.0
Receptive Language Scaled Score	180	9.8	3.0	1405	6.6	3.1
Motor Composite Score	178	98.2	10.9	1403	83.3	16.2
Fine Motor Scaled Score	180	10.3	2.0	1412	7.9	3.0
Gross Motor Scaled Score	179	9.1	2.2	1385	6.7	2.7

Table 4. Proportion of EP infants designated moderately/severely impaired using norm-based versus term reference based threshold scores for impairment.

Variable	Manual norm-based impairment threshold	Term reference based impairment threshold	Term reference vs. norm-based	
	N (%)	*N* (%)	OR (95% CI)[a]	*p*-value
Moderate/Severe Impairment				
Cognitive				
Normal/Mild	825 (57)	642 (44)	2.01 (1.69,2.39)	<0.001
Moderate/Severe	621 (43)	804 (56)		
Language				
Normal/Mild	625 (44)	716 (51)	0.68 (0.57,0.82)	<0.001
Moderate/Severe	784 (56)	693 (49)		
Motor				
Normal/Mild	776 (55)	662 (47)	1.60 (1.34,1.92)	<0.001
Moderate/Severe	627 (45)	741 (53)		
Level of Impairment (3 categories)				
Cognitive				
Normal/mild	825 (57)	642 (44)	2.50 (2.10,2.96)	<0.001
Moderate	395 (27)	478 (33)		
Severe	226 (16)	326 (23)		
Language Composite				
Normal/mild	625 (44)	716 (51)	0.60 (0.50,0.71)	<0.001
Moderate	412 (29)	379 (27)		
Severe	372 (26)	314 (22)		
Motor Composite				
Normal/mild	776 (55)	662 (47)	2.48 (2.08,2.96)	<0.001
Moderate	373 (27)	321 (23)		
Severe	254 (18)	420 (30)		
Overall Developmental Impairment (i.e. Cognitive, Motor or Langauage Composite)				
Moderate/severe impairment				
Normal/mild	603 (43)	457 (32)	1.86 (1.56,2.23)	<0.001
Moderate/severe/	806 (57)	965 (68)		
Level of impairment				
Normal/mild	603 (44)	457 (33)	2.72 (2.29,3.22)	<0.001
Moderate/	456 (33)	436 (31)		
Severe	326 (24)	497 (36)		

[a]Odds ratios are based on models accounting for clustering of participants by site. Higher odds ratios indicate greater impairment when using the term reference cut points and lower odds ratios indicate less impairment when using the term reference cut points.

assessed by examiners well aware of the infants' risk factors and prior assessments. Yet,the need for blinded assessors and concurrently assessed control patients should not be assumed to be less important to assure unbiased assessments in follow-up clinics than in other settings. High priority should be given in neonatal follow-up programs to developing effective methods to meet this need and to define appropriate impairment thresholds for EP infants.

DATA AVAILABILITY

Data collected at participating sites of the NICHD Neonatal Research Network were transmitted to RTI International, the data coordinating center (DCC) for the network, which stored, managed and analyzed the data included in this study. On behalf of the NRN, RTI International had full access to all the data in the study and take responsibility for the integrity of the data and accuracy of the data analysis. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Data reported in this paper may be requested through a data use agreement. Further details are available at https://neonatal.rti.org/index.cfm?fuseaction=DataRequest.Home.

REFERENCES

1. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA. 1995;273:408–12. https://doi.org/10.1001/jama.273.5.408.
2. Anderson PJ, Luca CRde, Hutchinson E, Roberts G, Doyle LW. Underestimation of developmental delay by the new Bayley-III Scale. Arch Pediatr Adolesc Med. 2010;164:352–6. https://doi.org/10.1001/archpediatrics.2010.20.
3. Skiöld B, Vollmer B, Bohm B, Hallberg B, Horsch S, Mosskin M. et al. Neonatal magnetic resonance imaging and outcome at age 30 months in extremely preterm infants. J Pediatr. 2012;160:559–66.e1. https://doi.org/10.1016/j.jpeds.2011.09.053.
4. Msall ME. Measuring outcomes after extreme prematurity with the Bayley-III Scales of infant and toddler development: a cautionary tale from Australia. Arch Pediatr Adolesc Med. 2010;164:391–3. https://doi.org/10.1001/archpediatrics.2010.25.
5. Aylward GP, Aylward BS. The changing yardstick in measurement of cognitive abilities in infancy. J Dev Behav Pediatr. 2011;32:465–8. https://doi.org/10.1097/DBP.0b013e3182202eb3.

6. Peña ED, Spaulding TJ, Plante E. The composition of normative groups and diagnostic decision making: shooting ourselves in the foot. Am J Speech-Lang Pathol. 2006;15:247–54. https://doi.org/10.1044/1058-0360(2006/023).
7. Sharp M, DeMauro SB. Counterbalanced Comparison of the BSID-II and Bayley-III at Eighteen to Twenty-two Months Corrected Age. J Dev Behav Pediatr. 2017;38:322–9. https://doi.org/10.1097/DBP.0000000000000441.
8. Vohr BR, Stephens BE, Higgins RD, Bann CM, Hintz SR, Das A. et al. Are outcomes of extremely preterm infants improving? Impact of Bayley assessment on outcomes. J Pediatr. 2012;161:222–8.e3. https://doi.org/10.1016/j.jpeds.2012.01.057.
9. Jary S, Whitelaw A, Walløe L, Thoresen M. Comparison of Bayley-2 and Bayley-3 scores at 18 months in term infants following neonatal encephalopathy and therapeutic hypothermia. Dev Med Child Neurol. 2013;55:1053–9. https://doi.org/10.1111/dmcn.12208.
10. Moore T, Johnson S, Haider S, Hennessy E, Marlow N. Relationship between test scores using the second and third editions of the Bayley Scales in extremely preterm children. J Pediatr. 2012;160:553–8. https://doi.org/10.1016/j.jpeds.2011.09.047.
11. Reuner G, Fields AC, Wittke A, Löpprich M, Pietz J. Comparison of the developmental tests Bayley-III and Bayley-II in 7-month-old infants born preterm. Eur J Pediatr. 2013;172:393–400. https://doi.org/10.1007/s00431-012-1902-6.
12. Serenius F, Kallen K, Blennow M, Ewald U, Fellman V, Holmstron G. et al. Neurodevelopmental outcome in extremely preterm infants at 2.5 years after active perinatal care in Sweden. JAMA. 2013;309:1810–20. https://doi.org/10.1001/jama.2013.3786.
13. Spencer-Smith MM, Spittle AJ, Lee KJ, Doyle LW, Anderson PJ. Bayley-III Cognitive and Language Scales in Preterm Children. Pediatrics. 2015;135:e1258–65. https://doi.org/10.1542/peds.2014-3039.
14. Spittle AJ, Spencer-Smith MM, Cheong JLY, Eeles AL, Lee KJ, Anderson PJ. et al. General movements in very preterm children and neurodevelopment at 2 and 4 years. Pediatrics. 2013;132:e452–8. https://doi.org/10.1542/peds.2013-0177.
15. Newman JE, Bann CM, Vohr BR, Dusick AM, Higgins RD. Improving the Neonatal Research Network annual certification for neurologic examination of the 18-22 month child. J Pediatr. 2012;161:1041–6. https://doi.org/10.1016/j.jpeds.2012.05.048.
16. Adams-Chapman I, Heyne RJ, DeMauro SB, Duncan AF, Hintz SR, Pappas A, et al. Neurodevelopmental Impairment Among Extremely Preterm Infants in the Neonatal Research Network. Pediatrics 2018;141. https://doi.org/10.1542/peds.2017-3091.
17. Bayley N. Bayley Scales of Infant and Toddler Development. Third Edition. Technical Manual (PsychCorp. Pearson Clinical Assessment, Bloomington, MN, 2006).
18. Vohr BR, Wright LL, Dusick AM, Perritt R, Poole WK, Tyson JE. et al. Center differences and outcomes of extremely low birth weight infants. Pediatrics. 2004;113:781–9. https://doi.org/10.1542/peds.113.4.781.
19. Bode MM, D`Eugenio DB, Mettelman BB, Gross SJ. Predictive validity of the Bayley, Third Edition at 2 years for intelligence quotient at 4 years in preterm infants. J Dev Behav Pediatr. 2014;35:570–5. https://doi.org/10.1097/DBP.0000000000000110.
20. Bayley N, Aylward GP. Bayley 4 Scales of Infant and Toddler Development. Fourth Edition. (NCS Pearson, Inc., Bloomington, MN, 2019).

ACKNOWLEDGEMENTS

The National Institutes of Health and the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD) provided grant support for the Neonatal Research Network for the Term Reference Study. NICHD staff provided input into the study design, conduct, analysis, and manuscript drafting; NCATS cooperative agreements provided infrastructure support to the NRN. While NICHD staff had input into the study design, conduct, analysis, and manuscript drafting, the comments and views of the authors do not necessarily represent the views of NICHD, the National Institutes of Health, the Department of Health and Human Services, or the U.S. Government.

AUTHOR CONTRIBUTIONS

JET and CEG fully participated in the design, planning, management of the current study as well as in drafting the manuscript.

FUNDING

The National Institutes of Health and the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD) (U10 HD21373, UG1 HD21364, UG1 HD21385, UG1 HD27851, UG1 HD27853, UG1 HD27856, UG1 HD27880,UG1 HD27904, UG1 HD34216, UG1 HD36790, UG1 HD40492, UG1 HD40689, UG1 HD53089, UG1 HD53109, UG1 HD68244, UG1 HD68270, UG1 HD68278, UG1 HD68263, UG1 HD68284; UG1 HD87226, UG1 HD87229) and the National Center for Advancing Translational Sciences (NCATS) (UL1 TR6, UL1 TR41, UL1 TR42, UL1 TR77, UL1 TR93, UL1 TR105, UL1 TR442, UL1 TR454, UL1 TR1117, provided grant support for the Neonatal Research Network, including for the Follow-up Study.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to Charles E. Green.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

EUNICE KENNEDY SHRIVER NATIONAL INSTITUTE OF CHILD HEALTH AND HUMAN DEVELOPMENT NEONATAL RESEARCH NETWORK

Richard A. Polin[25], Martin Keszler[26], Angelita M. Hensman[26], Elisa Vieira[26], Lucille St. Pierre[26], Robert T. Burke[26], Barbara Alksninis[26], Andrea Knoll[26], Mary L. Keszler[26], Teresa M. Leach[26], Elisabeth C. McGowan[26], Victoria E. Watson[26], Nancy S. Newman[27], Bonnie S. Siner[27], Elizabeth Roth[27], Angelia Williams[27], Brenda B. Poindexter[28], Kurt Schibler[28], K. Tanya E. Cahill[28], Cathy Grisby[28], Kristin Kirker[28], Sara Stacey[28], Sandra Wuertz[28], Ronald N. Goldberg[29], Ricki F. Goldstein[29], Patricia L. Ashley[29], Deesha Mago-Shah[29], Joanne Finkle[29], Kimberley A. Fisher[29], Kathryn E. Gustafson[29], Caitlin Stone[29], Matthew M. Laughon[29], Janice Bernhardt[29], Janice Wereszczak[29], Jennifer Talbert[29], Alexandra Bentley[29], Laura Edwards[29], Ginger Rhodes-Ryan[29], Donna White[29], David P. Carlton[30], Yvonne Loggins[30], Diane Bottcher[30], Sheena L. Carter[30], Salathiel Kendrick-Allwood[30], Maureen Mulligan LaRossa[30], Judith Laursen[30], Colleen Mackie[30], Amy Sanders[30], Gloria Smikle[30], Lynn Wineski[30], Andrew A. Bremer[31], Rosemary D. Higgins[31], Stephanie Wilson Archer[31], Amir M. Khan[32], Kathleen A. Kennedy[32], Andrea F. Duncan[32], Elizabeth Allain[32], Julie Arldt-McAlister[32], Fatima Boricha[32], Allison G. Dempsey[32], Carmen Garcia[32], Donna J. Hall[32], Janice John[32], M. Layne Lillie[32], Karen Martin[32], Georgia E. McDavid[32], Shannon L. McKee[32], Michelle Poe[32], Kimberly Rennie[32], Tina Reddy[32], Shawna Rodgers[32], Daniel K. Sperry[32], Sharon L. Wright[32], Leif D. Nelin[33], Jonathan L. Slaughter[33], Sudarshan R. Jadcherla[33], Christopher Timan[33], Patricia Luzader[33], Melanie Stein[33], Rox Ann Sullivan[33], Helen Carey[33], Stephanie Burkhardt[33], Mary Ann Nelin[33], Erna Clark[33], Kristi Small[33],

Jacqueline McCool[33], Lindsay Pietruszewski[33], Jessica Purnell[33], Kyrstin Warnimont[33], Laura Marzec[33], Bethany Miller[33], Demi R. Beckford[33], Hallie Baugher[33], Katelyn Levengood[33], Nancy Batterson[33], Jill Tonneman[33], Krystal Hay[33], Brittany DeSantis[33], Dennis Wallace[34], Jeanette O' Donnell Auman[34], Margaret Crawford[34], Jenna Gabrio[34], Jamie E. Newman[34], Lindsay Parlberg[34], Kristin M. Zaterka-Baxter[34], David K. Stevenson[35], M. Bethany Ball[35], Valerie Chock[35], Dona Bahmani[35], Barbara Bentley[35], Maria Elena DeAnda[35], Anne M. DeBattista[35], Beth Earhart[35], Lynne C. Huffman[35], Casey E. Krueger[35], Ryan E. Lucash[35], Heather Taylor[35], Hali E. Weiss[35], Namasivayam Ambalavanan[36], Kirstin J. Bailey[36], Fred J. Biasini[36], Stephanie A. Chopko[36], Monica V. Collins[36], Shirley S. Cosby[36], Kristy A. Domnanovich[36], Chantel J. Jno-Finn[36], Morissa Ladinsky[36], Mary Beth Moses[36], Tara E. McNair[36], Vivien A. Phillips[36], Julie Preskitt[36], Richard V. Rector[36], Kimberlly Stringer[36], Sally Whitley[36], Sheree York Chapman[36], Heidi M. Harmon[37], Karen J. Johnson[37], Mendi L. Schmelzel[37], Jacky R. Walker[37], Claire A. Goeke[37], Sarah E. Faruqui[37], Diane L. Eastman[37], Michelle L. Baack[37], Laurie A. Hogden[37], Megan M. Henning[37], Chelsey Elenkiwich[37], Megan Broadbent[37], Sarah Van Muyden[37], Robin K. Ohls[9], Conra Backstrom Lacy[9], Sandra Sundquist Beauman[9], Mary Ruffner Hanson[9], Jean R. Lowe[9], Elizabeth Kuan[9], Eric C. Eichenwald[38], Barbara Schmidt[38], Haresh Kirpalani[38], Soraya Abbasi[38], Aasma S. Chaudhary[38], Toni Mancini[38], Jonathan Snyder[38], Kristina Ziolkowski[38], Ronnie Guillet[39], Satyan Lakshminrusimha[39], Anne Marie Reynolds[39], Rosemary L. Jensen[39], Joan Merzbach[39], William Zorn[39], Osman Farooq[39], Gary J. Myers[39], Mary Rowan[39], Diane Prinzing[39], Melissa Bowman[39], Ann Marie Scorsone[39], Kyle Binion[39], Constance Orme[39], Premini Sabaratnam[39], Alison Kent[39], Rachel Jones[39], Elizabeth Boylin[39], Emily Li[39], Jennifer Kachelmeyer[39], Kimberly G. McKee[39], Kelly R. Coleman[39], Brenna Cavanaugh[39], Luc P. Brion[40], Diana M. Vasil[40], Sally S. Adams[40], Frances Eubanks[40], E. Rebecca McDougald[40], Lara Pavageau[40], Pollieanna Sepulveda[40], Alicia Guzman[40], Elizabeth Heyne[40], Lizette E. Lee[40], Azucena Vera[40], Jillian Waterbury[40], Cathy Twell Boatman[40], Bradley A. Yoder[41], Mariana Baserga[41], Roger G. Faix[41], Stephen D. Minton[41], Mark J. Sheffield[41], Carrie A. Rau[41], Shawna Baker[41], Susan Christensen[41], Sean D. Cunningham[41], Jennifer O. Elmont[41], Becky Hall[41], Erika R. Jensen[41], Manndi C. Loertscher[41], Trisha Marchant[41], Kandace M. McGrath[41], Hena G. Mickelsen[41], Galina Morshedzadeh[41], D. Melody Parry[41], Kelly Stout[41], Ashley L. Stuart[41] and Kimberlee Weaver-Lewis[41]

[25]Division of Neonatology, College of Physicians and Surgeons, Columbia University, New York, NY, USA. [26]Alpert Medical School of Brown University and Women & Infants Hospital of Rhode Island, Providence, RI, USA. [27]Case Western Reserve University, Rainbow Babies & Children's Hospital, Cleveland, OH, USA. [28]Cincinnati Children's Hospital Medical Center, University Hospital, and Good Samaritan Hospital, Cincinnati, OH, USA. [29]Duke University School of Medicine, University Hospital, University of North Carolina, Duke Regional Hospital, and WakeMed Health & Hospitals, Durham, NC, USA. [30]Emory University, Children's Healthcare of Atlanta, Grady Memorial Hospital, and Emory University Hospital Midtown, Atlanta, GA, USA. [31]Eunice Kennedy Shriver National Institute of Child Health and Human Development, Bethesda, MD, USA. [32]McGovern Medical School at The University of Texas Health Science Center at Houston, Children's Memorial Hermann Hospital, and Memorial Hermann Southwest Hospital, Houston, TX, USA. [33]Nationwide Children's Hospital, The Abigail Wexner Research Institute at Nationwide Children's Hospital, Center for Perinatal Research, The Ohio State University College of Medicine, The Ohio State University Wexner Medical Center, and Riverside Methodist Hospital, Columbus, OH, USA. [34]RTI International, Rockville, MD, USA. [35]Stanford University, El Camino Hospital, and Lucile Packard Children's Hospital, Stanford, CA, USA. [36]University of Alabama at Birmingham Health System and Children's Hospital of Alabama, Tuscaloosa, AL, USA. [37]University of Iowa and Sanford Health, Iowa City, IA, USA. [38]University of Pennsylvania, Hospital of the University of Pennsylvania, Pennsylvania Hospital, Children's Hospital of Philadelphia, and Virtua Voorhees Hospital, Philadelphia, PA, USA. [39]University of Rochester Medical Center, Golisano Children's Hospital, and the University of Buffalo Women's and Children's Hospital of Buffalo, Rochester, NY, USA. [40]University of Texas Southwestern Medical Center, Parkland Health & Hospital System, and Children's Medical Center Dallas, Dallas, TX, USA. [41]University of Utah Medical Center, Intermountain Medical Center, McKay-Dee Hospital, Utah Valley Hospital, and Primary Children's Medical Center, Salt Lake City, UT, USA.