Accuracy of Bayley Scores as Outcome Measures in Trials of Neonatal Therapies

Long-term follow-up of extremely preterm infants is essential because the proportion of infants who survive is increasing while neurodevelopmental impairment rates remain high. Psychometric tests, such as the Bayley Scales of Infant and Toddler Development, Third Edition (Bayley-III), are commonly used to measure early childhood outcomes in obstetric and neonatal randomized trials. However, the composite scores on psychometric tests can be inaccurate because of administrative and technical errors. To determine the magnitude of this problem, we performed a post hoc review of all errors detected by central source document verification of Bayley-III assessments.

Methods | Extremely preterm COT participants in 25 centers in 6 countries were randomly assigned to 2 oxygen saturation target ranges. The primary outcome at a corrected age of 18 to 21 months was death or survival with neurodevelopmental disability. Cognitive or language composite scores of less than 85 on the Bayley-III scales were included in the primary outcome. The research ethics boards of all clinical centers approved the protocol, and written informed consent was obtained from a parent or guardian of every study infant.

Original Review Process During COT. Experienced examiners administered the Bayley-III test and submitted copies of the source documents to the coordinating center. Between October 15, 2008, and August 15, 2012, trained staff (J.D. and L.C.) reviewed the source documents for accuracy, completeness, and agreement with the electronic database entries. Discrepancies were identified and corrected through a formal data clarification process with input from Bayley-III examiners at the respective study sites. A small number of queries (n = 36) could not be resolved directly with the clinical centers and were referred to an adjudication committee consisting of a developmental pediatrician (D.M.) and a neurodevelopmental consultant (K.P.).

Post Hoc Analysis. In this study, 1 assessor (J.D.) reexamined all Bayley-III source documents and classified the errors that were identified during the original review process into 5 categories: calculation of the child’s corrected age, documenting or applying scoring rules, raw score addition, look-up of scaled scores and composite scores in normative tables, and electronic data entry. A 15% random sample was independently classified by a second assessor (L.C.) to ensure consistent classification of error types. Only 2 disagreements were found and resolved through consensus. The error categories are hierarchical in nature. An early mistake—for example, calculation of the corrected age—may affect all subsequent steps. We counted such errors only once, assigned them to the category in which they first occurred, and summarized the frequency of independent errors in each of the 5 categories.

Results | During COT follow-up, the source documents for 936 of 954 (98.1%) Bayley-III assessments were submitted to the coordinating center. Eighteen children could not be evaluated because of severe developmental delay or autism. Of 936 source documents, 576 (61.5%) contained no errors. The remaining 360 (38.5%) contained at least 1 error, and the total number of independent errors was 387 (Table). None of the 25 clinical centers were completely error free. The best and worst center-specific error rates were 8 of 61 (13.1%) and 14 of 16 (87.5%), respectively. Had they not been detected during the original review process, 41 of 387 (10.6%) incorrectly reported composite scores would have changed the determination of the composite primary outcome in COT.

Discussion | Experienced Bayley-III examiners made numerous administrative, scoring, and reporting errors in this inter-

Table. Type and Frequency of Errors in Bayley-III Assessments

<table>
<thead>
<tr>
<th>Type of Error</th>
<th>Total Independent Errors</th>
<th>Changed Outcome of Cognitive or Language Delay</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corrected age calculation</td>
<td>165 (42.6)</td>
<td>20 (12.1)</td>
</tr>
<tr>
<td>Documenting/applying scoring rules</td>
<td>47 (12.1)</td>
<td>5 (10.6)</td>
</tr>
<tr>
<td>Raw score addition</td>
<td>83 (21.4)</td>
<td>2 (2.4)</td>
</tr>
<tr>
<td>Table look-ups</td>
<td>48 (12.4)</td>
<td>10 (20.8)</td>
</tr>
<tr>
<td>Data entry</td>
<td>44 (11.4)</td>
<td>5 (11.4)</td>
</tr>
<tr>
<td>Total</td>
<td>387 (99.9%)</td>
<td>42 (10.9)</td>
</tr>
</tbody>
</table>

national multicenter trial. Errors were detected and corrected in real time by comprehensive and rigorous central source document verification. The use of computer-assisted scoring software may reduce calculation and table look-up errors but cannot correct the effects of administrative or clerical errors that are made before data are entered into the program.

Conclusions | This study underscores the importance of diligent administration and recording of psychometric assessments. We recommend central source document verification of all psychometric tests that contribute to the primary outcome in large multicenter trials of perinatal and neonatal therapies.

Lorrie Costantini, BA
Judy D’Ilario, RN
Diane Moddemann, MD
Karen Penner, MSc
Barbara Schmidt, MD, MSc

Author Affiliations: Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada (Costantini, D’Ilario, Schmidt); Department of Pediatrics and Child Health, University of Manitoba, Winnipeg, Manitoba, Canada (Moddemann, Penner); Division of Neonatology, Children's Hospital of Philadelphia and University of Pennsylvania, Philadelphia (Schmidt).

Corresponding Author: Lorrie Costantini, BA, Department of Clinical Epidemiology and Biostatistics, McMaster University, Neonatal Trials Group, DBCVSRI Hamilton General Hospital Campus, 237 Barton St East, Hamilton, ON L8L 2X2, Canada (costant@mcmaster.ca).


Author Contributions: Ms Costantini and D’Ilario had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Study concept and design: Costantini, Schmidt.

Acquisition, analysis, or interpretation of data: Costantini, D’Ilario, Moddemann, Penner.

Drafting of the manuscript: Costantini.

Critical revision of the manuscript for important intellectual content: All authors.

Statistical analysis: Costantini.

Administrative, technical, or material support: Costantini, D’Ilario, Moddemann, Penner.

Study supervision: Schmidt.

Conflict of Interest Disclosures: None reported.

Funding/Support: The Canadian Oxygen Trial was supported by grant MCT-79217 from the Canadian Institutes for Health Research (Dr Schmidt and Moddemann).

Role of the Funder/Sponsor: The funding source had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.


COMMENT & RESPONSE

Risk Model for Renal Scarring: Validation and Implications Still Needed for Primary Care Practice

To the Editor We read the article by Shaikh et al recently published in JAMA Pediatrics with interest and were captivated by the concept of a model that could be used to determine the risk for renal scarring based on 3 noninvasive components (temperature, causal organism, and ultrasonography). However, based on limitations inherent in the methods, we caution against unreserved acceptance of these results. As the authors point out (and we agree), their findings require validation and further discussion as to implications for primary care practice.

Because this was a retrospective analysis that relied on raw, individual patient data reassembled many years after the initial studies, there is question as to the reliability and validity of the data. Shaikh et al included studies published 5 to 16 years ago in their analysis.

Also, we found it interesting that the proposed primary model was not additionally analyzed with the inclusion of the most important risk factor identified by the data, grade IV or V reflux (odds ratio = 23.70).

Determining which factors increase the risk for renal scarring is important and preventing renal scarring is the key objective in diagnosis and treatment. Moreover, not every patient should or needs to be assessed for vesicoureteral reflux. Vesicoureteral reflux remains an important element in the pathophysiology and, as this study reinforces, a significant risk factor for renal scarring.

Tracy Bunting-Early, PhD
T. Ernesto Figueroa, MD

Author Affiliations: Outcomes Positive Inc, Newark, Delaware (Bunting-Early); Division of Urology, Alfred I. duPont Hospital for Children, Wilmington, Delaware (Figueroa).

Corresponding Author: Tracy Bunting-Early, PhD, Outcomes Positive Inc, 364 E Main St, Newark, DE 19711 (tbe@outcomespositive.org).

Conflict of Interest Disclosures: None reported.


In Reply We thank Drs Bunting-Early and Figueroa for their interest in our article. Although they correctly point out that some of the data included are relatively old and that our analysis is retrospective (albeit using prospectively collected data), they do not proffer any reasons why these characteristics make the data unreliable. Individual-patient data meta-analysis is considered the best available study design to address risk prediction because it uses all available data and avoids confounding...