OPEN ACCESS

Balancing precision and affordability in assessing infant development in large-scale mortality trials: secondary analysis of a randomised controlled trial

Kristy P Robledo ,[1] Ingrid Rieger,[2] Sarah Finlayson,[1] William Tarnow-Mordi,[1] Andrew J Martin[1,3]

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/archdischild-2024-327762).

[1]NHMRC Clinical Trials Centre, The University of Sydney, Camperdown, New South Wales, Australia
[2]RPA Women and Babies, Camperdown, New South Wales, Australia
[3]The University of Queensland Centre for Clinical Research, Herston, Queensland, Australia

Correspondence to
Dr Kristy P Robledo;
kristy.robledo@sydney.edu.au

WT-M and AJM are joint senior authors.

Received 30 July 2024
Accepted 12 December 2024

To cite: Robledo KP, Rieger I, Finlayson S, *et al*. *Arch Dis Child Fetal Neonatal Ed* Epub ahead of print: [*please include* Day Month Year]. doi:10.1136/archdischild-2024-327762

ABSTRACT

Objective Large-scale mortality trials require reliable secondary assessments of impairment. We compared the Ages and Stages Questionnaire (ASQ-3), a screening tool self-administered by parents, in classifying impairment using the 'gold standard' Bayley Scales of Infant Development (Bayley-III), a diagnostic tool administered by trained assessors.

Design Analysis of 405 children around 2 years corrected age from the Australian Placental Transfusion Study, a trial conducted over 8 years.

Setting Secondary analysis of international, open-label, multicentre randomised trial.

Patients Children born <30 weeks gestation.

Interventions Immediate (<10 s) versus delayed (60 s+) cord clamping.

Main outcomes ASQ-3 and Bayley-III assessments around 2 years corrected age. Impairment (or developmental delay) was defined as <2 SD below the mean (<70) for Bayley-III domains.

Results The area under the receiver operating curve for ASQ-3 domains predicting delay was 0.75–0.99. Sensitivity for predicting delay was 57%–100%, while specificity was 88%–90%.

We modelled the cost and sample size using a less expensive, though less precise, screening assessment for impairment compared with a more costly diagnostic assessment. For detecting a 25% reduction in the relative risk of delay, using ASQ-3 rather than Bayley-III could require double the sample size (15 000 to 30 000), but outcome assessment cost savings would be US$13M (EUR$12M). However, assessment cost savings may be outweighed by upscaling.

Conclusions When measuring developmental outcomes in a large-scale clinical trial, using a more precise diagnostic tool may be financially prohibitive, so increasing the sample size and using a less precise but appropriately calibrated tool may be more affordable.

Trial registration number ACTRN12610000633088.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Large-scale neonatal or perinatal trials evaluating the effects of interventions on mortality require reliable, affordable assessments of impairment. Parent questionnaires (eg, Ages and Stages Questionnaire V.3; ASQ-3) are used for screening for developmental delay, but more precise gold-standard developmental assessments (eg, Bayley-III) are used for assessment of trial outcomes.

WHAT THIS STUDY ADDS

⇒ In 405 infants from the Australian Placental Transfusion Study, a trial of delayed cord clamping, the ASQ-3 successfully predicted developmental delays identified by Bayley-III. We found that the decreased cost of the ASQ-3 offsets the increase in sample size required for the less precise ASQ-3 assessment, compared with Bayley-III.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ If gold-standard assessments are financially prohibitive in large-scale trials, less precise assessments of impairment can be traded off against increased sample size to preserve statistical power.

INTRODUCTION

The best evidence on how to improve health outcomes for preterm infants comes from well-conducted, appropriately powered, randomised controlled trials (RCTs).[1] An adequately powered RCT may need several thousands of participants to detect moderate but clinically important effects,[2–7] especially if mortality is the primary outcome, and disability is a secondary outcome. Understanding the effect of an intervention used in preterm infants on longer term developmental outcomes in childhood is critical for understanding the intervention's overall net-benefit. Practical and affordable approaches for measuring such long-term outcomes are needed in large-scale trials that are designed to detect incremental but important treatment benefits.

The Bayley Scales of Infant Development (Bayley) is a widely used tool for the assessment of infant development[4 8–10] due to each domain correlating well with domain-specific assessments.[11 12] The Bayley has been referred to as the 'gold standard' for assessing behaviours of children aged 1–42 months[12], with moderate correlation with outcomes at 5 years,[13] and more fair correlation at 10 years.[14] While the latest version, Bayley-IV,[15] was released in 2021, it has not yet been widely utilised in clinical trials, unlike the extensively studied Bayley-III.[11] Like several other tools,[16 17] the Bayley-III requires administration by a trained

assessor. This can make it an expensive and logistically challenging option for deriving long-term endpoints in large-scale clinical trials. The cost associated with trained outcome assessors can be prohibitive for large-scale publicly funded phase III trials. For example, assessing developmental outcomes in the 11 976 infants born to women in the ASPIRIN trial using Bayley-III might cost well in excess of US$10M (EUR$9.2M).[6] More practical and affordable options warrant consideration.

Parent-administered screening assessments of child development may be an option.[18 19] The Ages and Stages Questionnaire V.3 (ASQ-3) is among the most widely used parent-completed screening tools for developmental delay with age-specific norms derived from a sample of more than 15 000 children. Measured by the intraclass correlation, the reliability between two ASQ-3 assessments by 145 parents 2 weeks apart ranged from 75% to 82%, indicating strong test–retest reliability.[20] However, compared with the gold standard Bayley-III, at least three studies have shown that the ASQ-3 is inadequate for diagnosing developmental delay in clinical practice, especially in low prevalence populations.[21–23]

To be clinically useful as a diagnostic tool, a developmental assessment needs to be unbiased and precise, as demonstrated by high positive and negative predictive value for distinguishing between cases and non-cases. However, in RCTs of infants, the assessment tool need not necessarily match the diagnostic tool performance to allow an unbiased and informative comparison of the treatment arms' average outcomes. Provided appropriate statistical techniques are used to correct for any miscalibration (systematic error), a precise estimate of the expected outcome in a treatment arm of an RCT can be constructed given a sufficient sample size.

Statistical correction for systematic error requires the understanding of the relationship between the scales of the candidate tool (eg, ASQ-3) and those of the criterion tool (eg, Bayley-III). There is a limit to how accurately a recalibrated candidate tool can predict scores from the criterion tool, due to unavoidable measurement error associated with measuring developmental outcomes. For example, with the Bayley-III, 6%–17% of the variation in scores is due to extraneous factors and measurement error, as revealed by Bayley-III test–retest reliability estimates ranging from 83% to 94% at 33–42 months[11]. A well-calibrated but imprecise ASQ-3 assessment would be inappropriate for individual diagnosis of developmental status. In an RCT however, any measurement imprecision from an appropriately calibrated tool can be mitigated by increasing the sample size.

We investigated whether a parental assessment of developmental delay (ASQ-3) is an affordable substitute for the gold-standard evaluation (Bayley-III) as an outcome assessment tool for use in large-scale neonatal trials assessing mortality. Our objectives were to determine the sensitivity and specificity of the ASQ-3 to detect delay as defined by Bayley-III thresholds; to investigate whether using ASQ against the Bayley-III can be improved on with the use of optimised cut-points for delay; and to model the trade-off of using a relatively inexpensive but less precise assessment tool compared with a more expensive and but reliable tool, in terms of trial sample size.

METHODS

Population

These children comprise a preplanned sample from the Australian Placental Transfusion Study (APTS), a multicentre, international, open label, RCT of delayed cord clamping in infants born <30 weeks gestation.[4 24] Fetuses were eligible if specialists considered that they might be delivered before 30 weeks of gestation. Exclusion criteria included fetal haemolytic disease, hydrops fetalis, twin–twin transfusion, genetic syndromes and potentially lethal malformations (see Protocol in online supplemental file 1). Enrolment started in a pilot trial on 21 October 2009 and closed on 6 January 2017, after 1634 fetuses were randomised in 25 centres in six high-income and one low-income country. In the APTS study, clinicians randomised fetuses (1:1) by minimisation, via an interactive voice response system when birth was imminent, to immediate cord clamping (<10 s) or deferred cord clamping (60 s or more), stratified by gestational age (<27, ≥27 weeks) and multiple birth status.[4] In July 2014, before any trial outcomes were known, the APTS childhood follow-up study prespecified the primary outcome of death or major disability in early childhood (2–3 years). The APTS follow-up study had 80% power, assuming alpha of 5% and 30% non-adherence to assigned treatment, to detect differences in the primary outcome ranging from 35% in the immediate-clamping group to 25.4% in the delayed-clamping group (a 27% reduction in relative risk (RR)) with 1350 infants to a difference of 40%–30% (a 25% reduction in RR) with 1450 infants, as detailed in the statistical analysis plan (online supplemental file 1, p67). No interim analyses were planned for the APTS childhood follow-up study, and the primary and other secondary outcomes have previously been reported.[18]

The APTS childhood follow-up study obtained funding for all children to be assessed using the ASQ-3, and a sample of children to be additionally assessed using the Bayley-III, in order to compare the secondary outcome of ASQ-3 and Bayley-III assessments. As Bayley-III is routinely used in many Australian hospitals for assessed of infants born <30 weeks gestation at 2–3 years corrected age, our cohort for this secondary analysis included a random sample, and an opportunistic sample from sites where Bayley-III assessments were routinely performed.

Bayley-III and ASQ-3

The ASQ-3 is a validated set of 21 age-appropriate questionnaires administered from 1 month to 5.6 years. The ASQ-3 was completed by primary caregivers either on paper, online or over the phone with research staff. If the child's age was outside the ASQ-3 at 24 months corrected age time window (23 months to 25 months and 15 days), the alternative age-appropriate ASQ-3 was completed. The ASQ assesses five areas of development: communication, gross motor, fine motor, problem-solving and personal-social. Additionally, an overall domain asks open-ended questions about the child's development. Each of the five domains are evaluated by 6–7 questions, forming an overall domain score. If the score is below the domain cut-off (two SD below the mean), further professional assessment may be needed. In the APTS study, the ASQ-3 questionnaires ranged from 22 months to 42 months. For each domain, mean and SD are available in the ASQ-3 manual for each age-appropriate questionnaire[20] and have been used to age standardise the ASQ domains to a mean of 100 and SD of 15. This enables ASQ-3 assessments performed over time to be pooled and compared with the age-standardised Bayley-III.

The Bayley-III was administered by a certified psychologist, paediatrician or other trained assessor, within 3 months of the ASQ-3 completion. The Bayley-III items cover cognitive, language and motor skills as well as social–emotional and adaptive behaviour. The motor skill domain comprises a fine motor and gross motor subscale, and the language domain comprises an expressive language and receptive language subscale. For each of

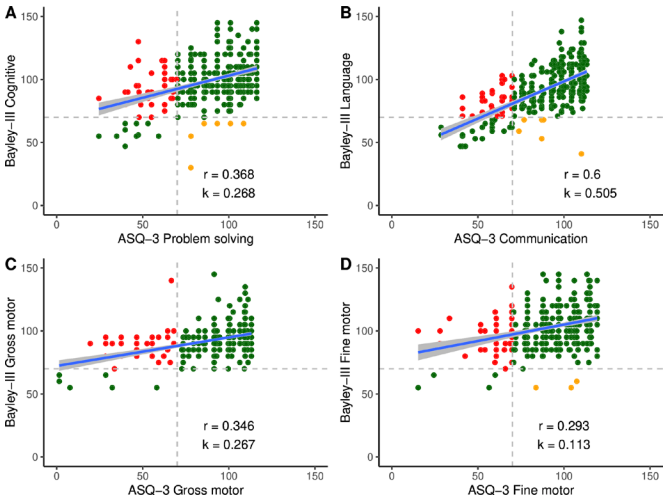

Figure 1 Scatterplots of ASQ-3 versus Bayley-III, for Problem solving versus Cognitive (A), Communication versus Language (B), Gross motor versus Gross motor (C) and Fine motor versus Fine motor (D), with regression lines shown in blue with grey 95% CI. Dotted lines denote 70 (2 SD below the mean). Green dots denote agreement of the two tools, red denotes a false-positive and orange denotes a false-negative, considering Bayley-III as the gold standard. Spearman correlations are given as r and kappa agreement as k. ASQ-3, Ages and Stages Questionnaire V.3.

these developmental areas, the Bayley-III yields a score, that is age-adjusted and standardised to a mean of 100 and an SD of 15. Additionally, the fine and gross motor subscales were rescaled to have mean 100 and SD of 15. The completion rates for the social–emotional domain are not reported as it is not routinely administered.

Statistical methods

The strength of association between the corresponding scales (online supplemental eTable 1) was quantified by Spearman correlation coefficients. Given the standardisation of the ASQ-3 data, both ASQ-3 and Bayley-III domain scores 2 SD below the norm (scores <70) were considered indicative of developmental impairment or delay. Kappa agreement statistics quantified the agreement of these two categorical delay definitions. As the Bayley-III is known to underestimate delay,[5 8 9 25–28] sensitivity analyses explored alternative cut-points of <80.[25 26 29]

Using the Bayley-III as the criterion measure, the performance of the ASQ-3 was evaluated in terms of its sensitivity and specificity (95% CI) and receiver operating characteristic (ROC) curves were prepared. The ROC curves plot the sensitivity and specificity of the ASQ-3 across all possible score cut points. That with the highest sensitivity and specificity was chosen[30] as the optimised ASQ-3 cut-point for each domain and contrasted with the conventional cut-point of 2 SDs below the norm (<70). The area under the ROC curves (AUC) provides an overall summary of the predictive performance of the ASQ-3.

We analysed the financial and sample size implications of using a more affordable, although less precise, screening tool for impairment/developmental delay versus a more expensive, criterion-based assessment. We constructed 30 hypothetical RCT scenarios that varied across the following three parameters: accuracy of the assessment (sensitivity and specificity varied from 80% to 100%), prevalence of delay (5% or 10%) and the experimental treatment effect size relative to control treatment (RR of (0.25, 0.5 or 0.75). The sample size needed for each of the 30 hypothetical RCTs to have 90% power to identify an effect at the 5% level of significance was calculated using standard calculations (see online supplemental eMethods). We also compared the cost of outcome assessment assuming that the ASQ-3 was used in preference to the Bayley-III.[31] For the ASQ-3 online system, an annual fee of US$850 (EUR$781) was assumed for access plus US$0.50 per child (EUR$0.46), along with US$295 per site for kits (EUR$271) and US$100 per child (EUR$92) for site payment.[32] For Bayley-III, we assumed US$997 per child (EUR$917) for the assessment[31] and US$100 per child (EUR$92) for site payment. Further details are provided in the online supplemental eMethods. Analyses were performed in SAS V.9.4 (Cary, USA) and R V.4.1.3.[33] Ethics approval was obtained for all participating hospitals. This trial is registered with the Australian and New Zealand Clinical Trials Registry (ACTRN12610000633088).

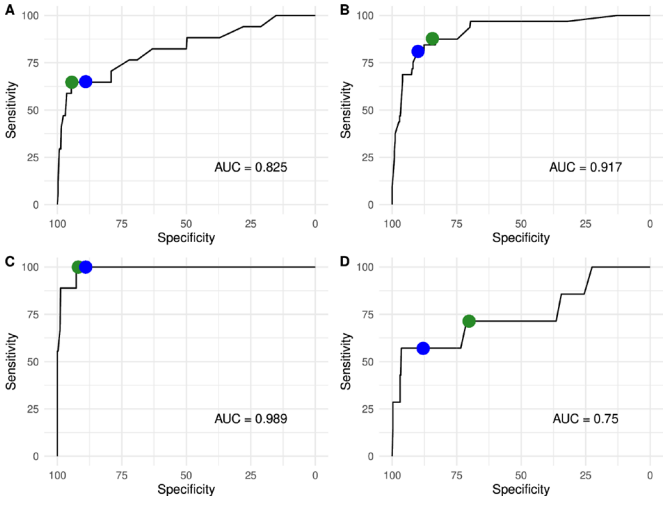

Figure 2 ROC for ASQ-3 domains: (A) cognitive, (B) language, (C) gross motor and (D) fine motor. Blue indicates conventional cut-points, and green indicates optimal cut-points. ASQ-3, Ages and Stages Questionnaire V.3; AUC, area under the ROC curves; ROC, receiver operating characteristic.

RESULTS

A total of 405 children had a Bayley-III performed within 3 months of an ASQ-3 assessment (online supplemental eFigure S1). The median age corrected for prematurity at the assessment was 24 months (range: 22–42). The mean gestational age at birth was 27.6 weeks (SD 1.6) with mean birth weight 1016 g (SD 262). 50% (203/405) were randomised to deferred cord clamping, 58% (235/405) were men and 79% (318/405) were singleton births. 71% (287/405) were Australian-born children, 20% (80/405) New Zealand and 9% (38/405) French-born (online supplemental eTable 2). This cohort of infants was generally representative of the APTS trial, both in terms of characteristics at birth and at 2 years, although all infants in this cohort were from Australia, New Zealand or France and alive at 2 years (online supplemental eTable S3).

Comparing the ASQ-3 and Bayley-III

ASQ-3 scores are plotted against the Bayley-III scores for each domain in figure 1. The agreement was highest for Bayley-III language and ASQ-3 communication scores, with Spearman correlation of 0.6 and Kappa agreement of 0.5, indicating weak to moderate correlation.

Table 1 Sensitivity and specificity for two ASQ-3 cut-points, conventional and optimal cut-points, according to delay defined by Bayley-III

	Conventional*				Optimal†			
Domain	Sensitivity	Specificity	Likelihood ratio (positive)	Likelihood ratio (negative)	Sensitivity	Specificity	Likelihood ratio (positive)	Likelihood ratio (negative)
Cognition	65% (38–86)	89% (86-92)	5.98 (3.8–9.4)	0.40 (0.21–0.75)	65% (38–86)	95% (92–97)	11.96 (6.94–20.6)	0.37 (0.20–0.71)
Language	81% (64–93)	90% (86-93)	8.10 (5.68–11.56)	0.21 (0.1–0.43)	84% (67–95)	88% (84-91)	6.85 (4.99–9.40)	0.18 (0.08–0.40)
Fine motor	57% (18–90)	88% (85-91)	4.91 (2.44–9.85)	0.49 (0.21–1.14)	71% (29–96)	71% (66–76)	2.47 (1.51–4.05)	0.40 (0.12–1.30)
Gross motor	100% (66–100)	89% (86-92)	9.21 (6.94–12.21)	0 (0–NaN)	100% (66–100)	93% (90–95)	13.66 (9.62–19.38)	0 (0–NaN)

*Conventional cut-points correspond to<70 (2SD below the norm).
†Optimal cut-points correspond to 60 (2.7SD), 75 (1.7SD), 59 (2.7SD) and 85 (1SD), for cognitive, language, gross, and fine motor delay, respectively.
ASQ-3, Ages and Stages Questionnaire V.3; NaN, 'Not a Number' as it cannot be defined.

Predicting delay with the ASQ

Based on a Bayley-III domain score cut-point of <70 (ie, 2 SDs below the mean), 4.2% (n=17) had cognitive delay, 7.9% (n=32) had language delay, 2.2% (n=9) had gross motor delay and 1.7% (n=7) had fine motor delay.

Using the Bayley-III definition of <70 for delay as the criterion, the ROC AUC for the ASQ-3 domains were high, at 0.825, 0.917, 0.989 and 0.750 for cognitive, language, gross and fine motor, respectively (figure 2). The sensitivity and specificity for conventional ASQ-3 cut-points are shown in table 1. Optimal cut-points are also given to maximise the sensitivity and specificity for detecting delay, particularly language and fine motor. These cut-points correspond to 60 (2.7SD), 75 (1.7SD), 59 (2.7SD) and 85 (1SD), for cognitive, language, gross and fine motor delay. Results using the Bayley-III <80 definition are slightly lower (online supplemental eTable 4 and eFigure 2).

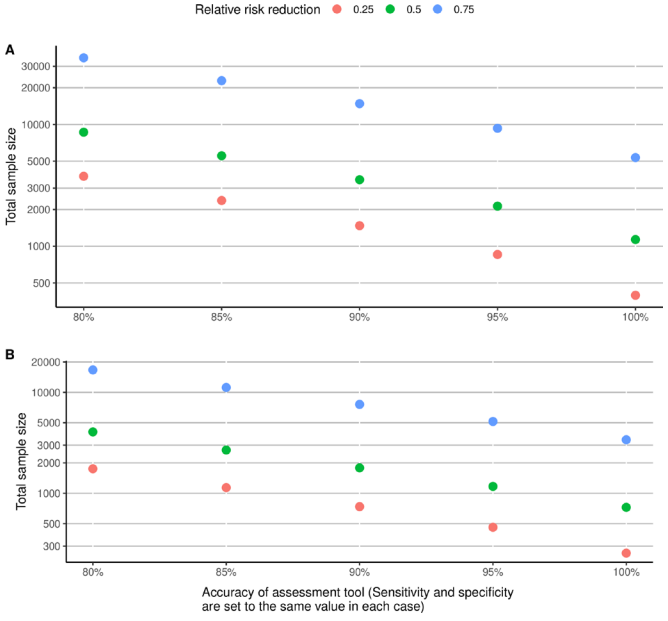

Figure 3 Scenarios to show that the higher the sensitivity and specificity (ie, accuracy) of an assessment tool, the lower the sample size that is required. To limit the number of examples, the sensitivity and specificity have been set to the same value in each scenario and the power is set to 90%. (A) and (B) show that a larger sample size is needed with a control rate of 10% (A), than with a control rate of 15% (B). Larger sample sizes are also required for smaller reductions in relative risk (eg, 25%) than for larger reductions (eg, 75%).

Predictive accuracy and sample size

Figure 3 summarises the sample size required for the 30 hypothetical RCTs and reveals three findings. First, the higher the sensitivity and specificity (ie, accuracy) of an assessment tool, the lower the sample size that is required. Second, a larger sample size is required when the prevalence of delay in the control arm is lower. Finally, smaller reductions in RR of delay require larger sample sizes.

The ASQ-3 and Bayley-III were compared in terms of sample size and cost (online supplemental eFigure 3 and table 2), assuming a 10% prevalence of delay in the control arm, a range of effect sizes (RR 0.25, 0.5 and 0.75) and that the sensitivity and specificity of the Bayley-III were greater (both 90%) than the ASQ-3 (65% and 89%, respectively). Although the ASQ-3 required around two times as many patients to detect the same effect size, it was more affordable. For a RR reduction of 25%, an extra 13 224 patients are required with the ASQ-3 assessment compared with the Bayley-III, however the ASQ-3 assessment results in over US$13M saved in trial-related assessment costs. Even when detecting a large RR reduction of 75%, the extra 1384 patients required with the use of the ASQ-3 is offset with the affordability of the ASQ-3 and over US$1.3M (EUR$1.2M) are saved.

DISCUSSION

This study found weak to moderate correlations[34] between the ASQ-3 and Bayley-III across various domains. Using the Bayley-III as the criterion (ie, 'gold standard'), the sensitivity of the ASQ-3 was higher in the language and gross motor domains than in fine motor and cognition. Using optimised ASQ-3 cut-points slightly improved the sensitivity, with the most significant increase noted in fine motor sensitivity but led to a decrease in specificity. This finding, along with others,[21–23] confirms the ASQ-3's suitability primarily as a screening tool for impairment, rather than as a diagnostic tool.

We have illustrated the trade-off between assessing developmental delay as an outcome in a large-scale mortality RCT using a more affordable, less precise assessment tool (ASQ-3) compared with a more expensive but more accurate tool (Bayley-III). The use of the ASQ-3 rather than the Bayley-III will result in a less precise estimate due to a high number of false negatives and false positives, therefore requiring a larger sample size to detect the same effect size at the same statistical power. The sensitivity and specificity of the given assessment tool can have an exponential effect on the sample size calculation, particularly at low prevalence rates of impairment. For example, given 90% power

Table 2 Comparison of sample sizes and costs for use of Bayley-III compared with ASQ-3 to assess cognitive delay, given 90% power, 5% alpha and 10% control rate of delay

Relative risk reduction	Sample size for cognitive domain (Bayley-III)*	US$ cost for Bayley-III assessments†	Sample size for cognitive domain (ASQ-3)‡	US$ cost for ASQ-3 assessments§	US$ difference in cost (Bayley-III minus ASQ-3)	EUR$ difference in cost (Bayley-III minus ASQ-3)
0.75	1 474	$1 615 504	2858	$294 829	$1 320 675	$1 213 898
0.50	3 518	$3 855 728	6726	$683 563	$3 172 165	$2 915 886
0.25	14 814	$16 236 144	28 038	$2 825 419	$13 410 725	$12 327 273

*Assuming sensitivity and specificity of 90% for Bayley-III.
†Assuming Bayley-III assessment cost of USD$996 per child (EUR$917), and USD$100 per child (EUR$92) as the site payment for completion.
‡Assuming sensitivity and specificity of 65% and 89% for ASQ-3.
§Assuming 20 sites performing follow-up over 2 years, with ASQ-3 costs of USD$850 per year (EUR$781) for ASQ online subscription plus USD$0.50 per child (EUR$0.46), USD$295 per site (EUR$271) for the ASQ-3 kit, and USD$100 per child (EUR$92) as site payment for ASQ-3 completion.
ASQ-3, Ages and Stages Questionnaire V.3.

and 5% significance, cognitive delay according to the ASQ-3 (a tool with 65% sensitivity and 89% specificity), would result in two times as many participants required compared with a more precise tool such as the Bayley-III (assuming 90% sensitivity and specificity). If the significance level was reduced from 5% to 1%, this would result in even larger sample sizes.[35–37] Despite requiring more participants, the ASQ-3's affordability results in an overall cost saving in trial-related costs for assessing impairment. Even with small effect sizes, the cost savings with ASQ-3 can be substantial (US$1.3M to US$13M (EUR$1.2-$12M), depending on the effect size). However, some of these savings in assessment costs may be offset by the costs of recruiting, and treating, more participants. As performed in our clinical trials,[4 10] it would be advantageous to increase precision by exploring a hybrid approach, where routine Bayley-III assessments are utilised if available and otherwise an ASQ-3 assessment is used. The required sample size will decrease from the ASQ-3 scenario, depending on the proportion of Bayley-III assessments available.

An important consideration in designing an RCT is missing data. As there is no foolproof manner to analyse data with large amounts of missing data, trialists should aim for 100% completion rates when designing clinical trials.[38] Researchers should consider that parental assessments like the ASQ-3 may have higher completion rates than the Bayley-III, as the ASQ-3 can be completed online at a parent's convenience, while the Bayley-III requires a lengthy, in-person paediatric appointment. However, in some settings, participants may have these assessments performed as a part of routine care. This highlights the importance of involving a collaborative group of experienced trialists and clinicians when choosing an assessment tool and designing a clinical trial.[37 39] Expert clinical trial statisticians will understand the relationship between an assessment tool's precision, sample size and cost; clinical researchers provide essential clinical insights, and parent representatives ensure that the trial is patient-centred.

This study has several limitations. First, as the Bayley-III is known to underestimate delay in Australian cohorts,[8 25 26] and while we attempted to investigate alternative cut-points for delay according to the Bayley-III, we are not comparing it to known impairment. Additionally, the Bayley-III is not a perfect instrument with a reported interobserver kappa of 0.77, which is a measure of the agreement between two different observers scoring the Bayley-III on the same child.[21] Third, we are reporting Bayley-III results, not the current version of the assessment, the Bayley-IV. Finally, we have estimated the costs for the two given assessments only, due to the variation in both the costs of administering treatments and recruiting participants in neonatal trials. However, in an era of larger, streamlined trials that emphasise registry-based outcomes, the assessment costs are drivers of these trial budgets.

When designing large-scale mortality trials in future, neonatal trialists could consider further strategies to assess impairment or development in survivors. These include obtaining consent for data linkage with state or national registries of educational outcomes[2 3] alongside an approach informed by value of information analyses,[40] which may reveal benefits in administering 'gold standard' assessments in a subset of participants.

In conclusion, in large-scale clinical trials assessing developmental outcomes, a trade-off exists. If criterion assessments prove to be financially prohibitive, it may be cost-effective to utilise a less precise, but affordable tool that is well calibrated. This approach may require a larger sample size to detect the same effect size, but the savings in assessment costs may be substantial.

X Kristy P Robledo @kristyrobledo and William Tarnow-Mordi @williamotm

Acknowledgements We would like to thank all the families and site staff who made the study possible, and the trial management committee, our collaborators and the data and safety monitoring committee.

Contributors WT-M, AJM, KPR and IR conceived this secondary analysis and WT-M was the chief investigator of the clinical trial. KPR performed the statistical analyses and verified the data. KPR, WT-M and AJM were responsible for writing the first version of the paper. All authors critically reviewed and approved the final version, and vouch for the accuracy and completeness of the data. All authors had full access to the data and had final responsibility for the decision to submit for publication. KPR acts as the guarantor and confirms that the manuscript is an honest, accurate and transparent account of the analyses being reported.

Funding APTS was funded by NHMRC 571309. APTS childhood follow-up study was funded by NHMRC APP1086865. The funders had no role in the design, collection, analysis or interpretation of the data, including the writing of the report and the decision to submit for publication.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval As stated in previous publications, an ethics board at each centre approved the study. These are listed below. Presented in the format: HREC/IRB reference number—site(s) covered. Women and Newborn Health Service Ethics Committee—1879EW—KEMH. ACT Health HREC—ETH.3.11.049—Canberra Hospital. Northern Sydney Local Health District HREC—1007-272M—RPA, RNSH, JHH, Liverpool, RHW, Nepean. Children's Health Services HREC—HREC/11/QRCH/12 Townsville, Mater Mothers', RBWH. Mercy Health HREC—R11/16- Mercy Hospital for Women. Monash Health HREC—B 11031B—Monash. Southern Adelaide Clinical HREC—006.11—FMC. Northern B HDEC—MEC/11/12/102/AM08—Auckland, Christchurch, Dunedin, Waikato, Wellington. Office for Research Ethics Committees Northern Ireland—12/NI/0164—Royal Jubilee Maternity, Craigavon. University of Vermont Committees on Human Subjects—CHRMS: 14-356—Fletcher Allen Aga Khan University Ethical Review Committee—2451-Obs-ERC-13—Aga Khan Baylor College of Medicine Office of Research—H-34236—Texas Children's Hospital IWK Health Centre Research Ethics Board—1018451—IWK Health. CPP Ile de France XI Comite de Protection des Personnes—14052—Antoine-Beclere. Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. Individual de-identified participant data from the results reported in this article will be available for 5 years after publication. Researchers will need to provide a methodologically sound proposal to apts.study@sydney.edu.au and this will be reviewed by the trial management committee. Researchers will need to sign a data access agreement.

ORCID iD

Kristy P Robledo http://orcid.org/0000-0003-0213-7652

REFERENCES

1 Barton S. Which clinical studies provide the best evidence? *BMJ* 2000;321:255–6.
2 Yusuf S, Collins R, Peto R. Why do we need some large, simple randomized trials? *Stat Med* 1984;3:409–22.
3 Tarnow-Mordi W, Kumar P, Kler N. Neonatal trials need thousands, not hundreds, to change global practice. *Acta Paediatr* 2011;100:330–3.
4 Robledo KP, Tarnow-Mordi WO, Rieger I, *et al*. Effects of delayed versus immediate umbilical cord clamping in reducing death or major disability at 2 years corrected age among very preterm infants (APTS): a multicentre, randomised clinical trial. *Lancet Child Adolesc Health* 2022;6:150–7.
5 Moore T, Hennessy EM, Myles J, *et al*. Neurological and developmental outcome in extremely preterm children born in England in 1995 and 2006: the EPICure studies. *BMJ* 2012;345:e7961.
6 Hoffman MK, Goudar SS, Kodkany BS, *et al*. Low-dose aspirin for the prevention of preterm delivery in nulliparous women with a singleton pregnancy (ASPIRIN): a randomised, double-blind, placebo-controlled trial. *Lancet* 2020;395:285–93.
7 Magpie Trial Follow-Up Study Collaborative Group. The Magpie Trial: a randomised trial comparing magnesium sulphate with placebo for pre-eclampsia. Outcome for children at 18 months. *BJOG* 2007;114:289–99.
8 Spittle AJ, Spencer-Smith MM, Eeles AL, *et al*. Does the Bayley-III Motor Scale at 2 years predict motor outcome at 4 years in very preterm children? *Dev Med Child Neurol* 2013;55:448–52.
9 Vohr BR, Stephens BE, Higgins RD, *et al*. Are outcomes of extremely preterm infants improving? Impact of Bayley assessment on outcomes. *J Pediatr* 2012;161:222–8.
10 Martin A, Ghadge A, Manzoni P, *et al*. Protocol for the Lactoferrin Infant Feeding Trial (LIFT): a randomised trial of adding lactoferrin to the feeds of very-low birthweight babies prior to hospital discharge. *BMJ Open* 2018;8:e023044.
11 Albers CA, Grieve AJ, Bayley N. Test Review: Bayley, N. (2006). Bayley Scales of Infant and Toddler Development– Third Edition. San Antonio, TX: Harcourt Assessment. *J Psychoeduc Assess* 2007;25:180–90.
12 Del Rosario C, Slevin M, Molloy EJ, *et al*. How to use the Bayley Scales of Infant and Toddler Development. *Arch Dis Child Educ Pract Ed* 2021;106:108–12.
13 Schmidt B, Anderson PJ, Doyle LW, *et al*. Survival without disability to age 5 years after neonatal caffeine therapy for apnea of prematurity. *JAMA* 2012;307:275–82.
14 Taylor GL, Joseph RM, Kuban KCK, *et al*. Changes in Neurodevelopmental Outcomes From Age 2 to 10 Years for Children Born Extremely Preterm. *Pediatrics* 2021;147:e2020001040.
15 Albers CA, Grieve AJ. *Bayley scales of infant and toddler development: 4th edition. Technical manual*. NCS Pearson, 2019.
16 Griffiths R. *The abilities of young children: a comprehensive system of mental measurement for the first eight years of life*. 1st edn. Child Development Research Centre, 1970.
17 Frankenburg WK, Dodds JB. The Denver developmental screening test. *J Pediatr* 1967;71:181–91.
18 Martin AJ, Darlow BA, Salt A, *et al*. Identification of infants with major cognitive delay using parental report. *Dev Med Child Neurol* 2012;54:254–9.
19 Martin AJ, Darlow BA, Salt A, *et al*. Performance of the Parent Report of Children's Abilities-Revised (PARCA-R) versus the Bayley Scales of Infant Development III. *Arch Dis Child* 2013;98:955–8.
20 Squires J, Bricker D, Twombly E. *The ASQ-3 user's guide*. 3rd edn. Brookes, 2009.
21 Steenis LJP, Verhoeven M, Hessen DJ, *et al*. Performance of Dutch children on the Bayley III: a comparison study of US and Dutch norms. *PLoS One* 2015;10:e0132871.
22 Veldhuizen S, Clinton J, Rodriguez C, *et al*. Concurrent validity of the Ages And Stages Questionnaires and Bayley Developmental Scales in a general population sample. *Acad Pediatr* 2015;15:231–7.
23 Yue A, Jiang Q, Wang B, *et al*. Concurrent validity of the Ages and Stages Questionnaire and the Bayley Scales of Infant Development III in China. *PLoS One* 2019;14:e0221675.
24 Tarnow-Mordi W, Morris J, Kirby A, *et al*. Delayed versus Immediate Cord Clamping in Preterm Infants. *N Engl J Med* 2017;377:2445–55.
25 Anderson PJ, De Luca CR, Hutchinson E, *et al*. Victorian Infant Collaborative Group the. Underestimation of developmental delay by the new Bayley-III scale. *JAMA Pediatr* 2010;164:352–6.
26 Spencer-Smith MM, Spittle AJ, Lee KJ, *et al*. Bayley-III Cognitive and Language Scales in Preterm Children. *Pediatrics* 2015;135:e1258–65.
27 Lowe JR, Erickson SJ, Schrader R, *et al*. Comparison of the Bayley II Mental Developmental Index and the Bayley III cognitive scale: are we measuring the same thing? *Acta Paediatr* 2012;101:e55–8.
28 Serenius F, Blennow M, Maršál K, *et al*. Intensity of perinatal care for extremely preterm infants: outcomes at 2.5 years. *Pediatrics* 2015;135:e1163–72.
29 Yi YG, Sung IY, Yuk JS. Comparison of Second and Third Editions of the Bayley Scales in Children With Suspected Developmental Delay. *Ann Rehabil Med* 2018;42:313–20.
30 Youden WJ. Index for rating diagnostic tests. *Cancer* 1950;3:32–5.
31 Doyle LW, Clucas L, Roberts G, *et al*. The cost of long-term follow-up of high-risk infants for research studies. *J Paediatr Child Health* 2015;51:1012–6.
32 ASQ product packages. 2023 Available: https://brookespublishing.com/asq-product-packages/
33 R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing; 2022. Available: https://www.R-project.org/
34 Mukaka MM. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi Med J* 2012;24:69–71.
35 Ioannidis JPA. The Proposal to Lower P Value Thresholds to .005. *JAMA* 2018;319:1429–30.
36 Benjamin DJ, Berger JO, Johannesson M, *et al*. Redefine statistical significance. *Nat Hum Behav* 2018;2:6–10.
37 Tarnow-Mordi WO, Robledo K, Marschner I, *et al*. To guide future practice, perinatal trials should be much larger, simpler and less fragile with close to 100% ascertainment of mortality and other key outcomes. *Semin Perinatol* 2023;47:151789.
38 Little RJ, D'Agostino R, Cohen ML, *et al*. The prevention and treatment of missing data in clinical trials. *N Engl J Med* 2012;367:1355–60.
39 Ioannidis JPA. How to make more published research true. *PLoS Med* 2014;11:e1001747.
40 Tuffaha H. Value of Information Analysis: Are We There Yet? *Pharmacoecon Open* 2021;5:139–41.