OPEN ACCESS

Is survival and neurodevelopmental impairment at 2 years of age the gold standard outcome for neonatal studies?

Neil Marlow

Correspondence to
Professor Neil Marlow, Neonatal Medicine, UCL EGA Institute for Women's Health, 74 Huntley Street, London WC1E 6AU, UK; n.marlow@ucl.ac.uk

Received 17 July 2014
Revised 4 September 2014
Accepted 10 September 2014
Published Online First 10 October 2014

INTRODUCTION

Designing perinatal trials is a continuing challenge. In the process, the choice of primary outcome is one of the critical decisions and one that will determine the necessary trial size and ultimately the success or failure. The primary outcome should be directly causally relevant to the intervention under study and the difference sought must be clinically relevant. Alongside the test of efficacy, there is also a need to ensure that a treatment is safe. Few trials are now designed without a measure of developmental outcome at 18–24 months as a primary or coprimary outcome.[1] This is a complex outcome, being a composite usually of death and four to five domains of developmental impairment, based on value judgments as to severity of impairment in the areas of motor, cognitive, hearing and visual functions, and sometimes communication. I would like to address a range of issues with the use of this as a composite outcome and the extent to which we should rely on it to show benefit when we are assessing research based evidence about interventions.

This paper discusses the relevance of 2-year outcomes in several trials and develops some ideas as to how we might consider 2-year outcomes, in terms of their interpretation and in the implementation of trial data into practice.

ROBUSTNESS OF 2-YEAR OUTCOMES

Recently I reviewed the challenges in measuring 2 year outcomes in research practice.[1] They pose real issues in quality control, diagnostic accuracy and interpretation of the measures themselves in practice. Their predictive value for an individual is relatively poor, except for the most severe impairments.[2] Most studies use a combined severe and moderate outcome classification similar to that published by British Association of Perinatal Medicine.[3] We can improve the predictive accuracy by reducing noise in the measure from variation in assessment technique or by statistically correcting for demographic factors,[4] but the latter is difficult to justify within the confines of a properly designed clinical trial, where such factors should segregate equally between groups. IQ and educational measures at school age may provide better targets for outcome, in terms of being more reliable, but run the risk of large losses to follow-up and extending the duration of each study.

RELEVANCE OF 2-YEAR OUTCOMES

Although we often cite '2-year outcome' as the 'primary' outcome we use a composite outcome that includes death and a range of impairments summated as 'disability' (or survival without disability). The argument runs that we have to account for mortality in a high mortality setting, such as a very preterm population, because an effect on death might mask the effect of the intervention on other outcomes, if the risk of death is different in the two groups. As a byproduct of including deaths, the prevalence of the primary outcome is increased, and thereby the power of the study, rendering the calculated trial size smaller than it might otherwise need to be. Although it is, at first look, a useful end point for a trial, we do see many trials that show little effect on this combined outcome. This leads to the questions—was this composite outcome correctly chosen in the first place, are the effects of the treatment so diluted by other events causing disability that no effect was seen, or is the treatment ineffective?

There are two aspects to this that bear some thought.

First, such a composite outcome may be considered useful if the effects on death and on disability are continuous in their relation to causation in this context (and therefore act in the same direction). The two components are clearly hierarchical, and in a trial in a population where there is high mortality and high rates of morbidity from the condition under treatment, it is difficult to evaluate one without the other. This is the advantage of such a composite outcome: allowing the investigator to get out of the challenge of defining a single end point when the risk studied by the intervention affects the chances of survival and of morbidity (competing risks). Occasionally some interventions may act in different directions for different components: there may be an increase in deaths but a decrease in morbidity, for example. How we cope with this in analysis and in drawing conclusions is challenging and becomes a matter of judgment. One of the first steps is to ensure that the components of the composite are reported alongside the primary outcome, facilitating interpretation and transparency. These component outcomes are then compared between groups after, for example, removing deaths, and after statistical adjustments for the multiple comparisons that are undertaken, something that is often not done. Assessing effects on single components of the composite has considerably less power than the composite and trials are rarely powered on such an analysis. Hence effect size has to be large to show an effect on a single component.

We currently face such a dilemma when evaluating the results of the five trials of oxygen saturation

To cite: Marlow N. *Arch Dis Child Fetal Neonatal Ed* 2015;**100**:F82–F84.

targeting. It seems clear from the first published meta-analysis[5] that mortality is increased in the group with the lower targeting range (relative risk: 1.41 (95% CI 1.14 to 1.74)) but the rate of retinopathy may be increased in the higher targeted group (RR: 0.74 (0.59 to 0.92)), leading to increased potential visual impairment. As it is, for the purposes of the prospective meta-analysis the chosen end point is the composite of death and abnormal neurodevelopment[6]; visual impairment is likely to be a tiny proportion of the latter (and much due to central visual impairment rather than retinal problems), but clearly there is a major tension between directly relevant intermediate outcomes so far. The interpretation of the 2-year outcomes will be difficult.

Second, using this specific composite 2-year outcome we need to have a strong view that death and disability are on a direct causal pathway from the intervention under test. Although it is true that if a baby has died they cannot develop impairment, often cited as an argument for using combined outcomes, failing to ensure both are on a direct causal pathway will run the risk of missing an important treatment effect. Furthermore to use a more general composite outcome reduces the risk of finding any effect, as other influences that determine the proportion of children who die or have developmental, neurological or sensory impairment may overwhelm or dilute a small effect from the treatment. Ideally, among those children with death or disability we need to determine the proportion for whom the outcome is directly the result of the target of our intervention (true positive) or due to other causes (false positives). In any trial the higher the rate of 'false positives', the greater the likelihood of a null result.[7 8] Thus, among the myriad events that led to brain injury, the proportion that were directly related to the trial intervention (and what were due to other events), becomes an important question, as it is for deaths and sensory impairments.

In the example of a recent large trial of magnesium sulfate for neuroprotection[9] the chosen end point was death or moderate/severe cerebral palsy rather than a more general impaired outcome; this was chosen because it was thought to be related better to the concept of neuroprotection. This composite was not significantly different between groups (RR: 0.97 (0.77 to 1.23); p=0.80). Death was considered important because of the results of a smaller earlier trial that claimed to show significantly increased mortality risk, although there appeared to be no direct causative pathway.[10] In this trial of antenatal magnesium sulfate, there was a non-significant increase in deaths in the magnesium group (RR: 1.12 (0.85 to 1.47)) but a significant reduction in the proportion with cerebral palsy (RR: 0.55 (0.32 to 0.99)). Further secondary analysis revealed that many of the deaths were ascribed to the presence of a congenital anomaly (ie, false positives), removal of which reduced the risk to 1.03. Hence it seems reasonable to conclude that magnesium sulfate infused during preterm labour reduces the risk of cerebral palsy without increasing mortality appreciably. Within the same trial they also carried out developmental testing but found no difference in proportions of children meeting conventional cut-offs for impairment (<70 or <85). Twenty children in the magnesium sulfate group and 24 children in the placebo group did not have Bayley results for reasons that are not stated. Previous experience has shown that those not followed may be the ones with greatest impairment, thus interpreting these findings is challenging—particularly if there was an excess of children with cerebral palsy who did not receive developmental testing, which was carried out in that study at a second assessment visit. Had the authors chosen an even more complex end point—death or multicomponent disability—results may have driven very different conclusions. The findings of this trial are reinforced by confirmation in a systematic review.[11]

One way to get around the high rate of false positives is to target a high-risk group in the study—to individualise trial entry. For example the two largest trials of high frequency oscillation are subtly different though often combined in meta-analysis. The UK Oscillation Study accepted all comers and started as soon after birth as possible. This was irrespective of the severity of lung disease; no effects on neonatal or on long-term outcomes to 2 years was found.[12] In the second trial[13] trial entry was delayed until it was clear that surfactant had been ineffective; high frequency ventilation was effective at reducing bronchopulmonary dysplasia (BPD) in this study. Much has been made of the conflicting results between the two trials but this fundamental difference is rarely acknowledged, making it difficult to combine the two studies.

2-YEAR OUTCOME AS AN APPROPRIATE TARGET

Finally we might consider whether 2-year outcomes are in fact the most appropriate efficacy outcome. Many interventions in the neonatal period are targeted on short-term benefit and, as such, their combined benefit may in time lead to improved neurodevelopmental outcomes. However using neurodevelopmental outcomes as primary outcomes for single agent trials, as indicated above, may hide significant effects if the neonatal benefits are not weighed in the assessment. For example, most would agree that indometacin is a useful treatment for patent arterial ducts. Prophylactic indometacin as studied in the Trial of Indomethacin Prophylaxis in Preterms (TIPP) trial is associated with reduced need for rescue treatment with further drug or surgery and, as a by-product, with a smaller proportion of babies with a large intraventricular haemorrhage—overall 12% of the population.[14] There was, however, no benefit in reducing the proportion with death or disability at 18 months, which was present in 46.5% of the population. This does not mean that these neonatal benefits should be ignored, simply that other factors leading to death or impairment may overwhelm the signal from a reduction in a less prevalent risk factor. Thus the data would seem to show that prophylactic indomethacin to reduce problems associated with patent arterial ducts is a safe therapy. Indeed the investigators recently posited the rhetorical question: "why would a sane clinician not prescribe prophylactic indomethacin?"[15]

In several trials of interventions in the immediate or early neonatal period currently being proposed, the primary outcome is death or impairment at 2 years. Is this sensible? Would we eschew the use of the intervention on the premise it didn't improve 2-year outcomes, despite clear neonatal benefits? In this we might use as examples two current interventions where trials are being planned.

First, trials of immediate or delayed cord clamping: in these there is a drive to show that delaying cord clamping improves 2-year outcomes, if successful this would settle any lingering concerns about the practice. Presumably this effect is via a route whereby deaths are avoided because of improved early condition, brain outcomes are improved because of less low blood pressure (or the interventions used to correct it), etc. Many other factors compete for these casual pathways and I would suggest that the route from intervention to effect on 2-year outcomes is so complex that it is impossible to disentangle the effect of one intervention from other causes of what are false positive outcomes without a huge trial in which the competing risks are well balanced. Second, trials of non-steroidal

interventions for BPD may use 2-year outcomes on the basis that BPD is related to developmental outcome. The effect of the intervention in reducing BPD is important and clear, whereas the effect on neurodevelopment is likely to be diluted with so many other influences (false positives) that this effect is lost.

Should these trials not be designed primarily on short-term end points that are directly related to delaying cord clamping or to respiratory outcomes, respectively, and determine whether these factors, intermediate in the potential causal pathways, have been affected? Subsequently longer-term outcomes might then be considered, but I would view the 2-year outcomes as primarily a safety assessment.

CONCLUSION

Thus in conclusion, we have invested in the use of 2-year neurodevelopmental outcomes as markers of our care and as results to use in research. We need to be aware of the nature and dangers of using such a complex composite outcome in research and to be prepared to use it as evidence of safety in situations where the neonatal signal from an intervention is likely to be overwhelmed by other influences that determine outcome. We should tailor the outcome to the causal pathway we are testing in order to reduce to a minimum the false positive outcomes to maximise our ability to detect effects of treatments. This may only mean using certain components of the impairment classification. Finally we should not be afraid to regard 2-year outcomes as proof of safety rather than efficacy, and therefore be reassured in using the treatment. Multiple interventions will combine to produce incremental benefit in outcomes, such as those seen in the EPICure studies,[16] but individual interventions may produce benefits that are undetectable in terms of 2-year outcomes. This does not mean they are worthless.

Acknowledgements The author receives part funding from the Department of Health's NIHR Biomedical Research Centre's funding scheme at UCLH/UCL. The author is grateful to Professor Donald Peebles for reading the manuscript and making suggestions.

Funding UCL/UCLH Biomedical Research Centre.

Competing interests None.

Provenance and peer review Commissioned; externally peer reviewed.

REFERENCES

1 Marlow N. Measuring neurodevelopmental outcome in neonatal trials: a continuing and increasing challenge. *Arch Dis Child Fetal Neonatal Ed* 2013;98:F554–8.
2 Marlow N, Wolke D, Bracewell MA, *et al*. Neurologic and developmental disability at six years of age after extremely preterm birth. *N Engl J Med* 2005;352:9–19.
3 Report of a BAPM/RCPCH working group. *Classification of Health Status at 2 years as a perinatal outcome*. London: BAPM, 2008.
4 Avon Premature Infant Project. Randomised trial of parental support for families with very preterm children. Avon Premature Infant Project. *Arch Dis Child Fetal Neonatal Ed* 1998;79:F4–11.
5 Saugstad OD, Aune D. Optimal oxygenation of extremely low birth weight infants: a meta-analysis and systematic review of the oxygen saturation target studies. *Neonatology* 2014;105:55–63.
6 Askie LM, Brocklehurst P, Darlow BA, *et al*. NeOProM: Neonatal Oxygenation Prospective Meta-analysis Collaboration study protocol. *BMC Pediatr* 2011;11:6.
7 Kessler KM. Combining composite endpoints: counterintuitive or a mathematical impossibility? *Circulation* 2003;107:e70.
8 Prieto-Merino D, Smeeth L, Staa TP, *et al*. Dangers of non-specific composite outcome measures in clinical trials. *BMJ* 2013;347:f6782.
9 Rouse DJ, Hirtz DG, Thom E, *et al*. A randomized, controlled trial of magnesium sulfate for the prevention of cerebral palsy. *N Engl J Med* 2008;359:895–905.
10 Mittendorf R, Covert R, Boman J, *et al*. Is tocolytic magnesium sulphate associated with increased total paediatric mortality? *Lancet* 1997;350:1517–18.
11 Doyle LW, Crowther CA, Middleton P, *et al*. Magnesium sulphate for women at risk of preterm birth for neuroprotection of the fetus. *Cochrane Database Syst Rev* 2009;(1):CD004661.
12 Johnson AH, Peacock JL, Greenough A, *et al*. High-frequency oscillatory ventilation for the prevention of chronic lung disease of prematurity. *N Engl J Med* 2002;347:633–42.
13 Courtney SE, Durand DJ, Asselin JM, *et al*. High-frequency oscillatory ventilation versus conventional mechanical ventilation for very-low-birth-weight infants. *N Engl J Med* 2002;347:643–52.
14 Schmidt B, Davis P, Moddemann D, *et al*. Long-term effects of indomethacin prophylaxis in extremely-low-birth-weight infants. *N Engl J Med* 2001;344:1966–72.
15 DeMauro SB, Schmidt B, Roberts RS. Why would a sane clinician not prescribe prophylactic indomethacin? *Acta paediatrica* 2011;100:636.
16 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.