ACCEPTED VERSION

Lisa N. Yelland, Maria Makrides, Andrew J. McPhee, Julie Quinlivan, Robert A. Gibson
Importance of adequate sample sizes in fatty acid intervention trials
Prostaglandins, Leukotrienes and Essential Fatty Acids, 2016; 107:8-11

Final publication at http://dx.doi.org/10.1016/j.plefa.2016.02.003

http://hdl.handle.net/2440/99443

Importance of adequate sample sizes in fatty acid intervention trials

Lisa N Yelland[1,2], Maria Makrides[1,3,4], Andrew J McPhee[5], Julie Quinlivan[6], Robert A Gibson[7]

[1]Women's and Children's Health Research Institute, The University of Adelaide, Australia

[2]School of Public Health, The University of Adelaide, Australia

[3]South Australian Health and Medical Research Institute, Australia

[4]School of Paediatrics and Reproductive Health, The University of Adelaide, Australia

[5]Neonatal Medicine, Women's and Children's Health Network, Australia

[6]Institute for Health Research, University of Notre Dame, Australia

[7]School of Agriculture, Food and Wine, The University of Adelaide, Australia

Author for Correspondence: Dr Lisa Yelland, School of Public Health, Mail Drop DX650 511, The University of Adelaide, SA 5005, Australia, Email: lisa.yelland@adelaide.edu.au, Phone: +61 8 8313 3215, Fax: +61 8 8223 4075. Reprints not available.

Short title: Sample sizes in fatty acid trials

Summary

Randomised controlled trials are the ideal way to assess the effects of interventions. Small trials are useful for generating pilot data to determine samples sizes for larger trials, but can produce unreliable or biased results if they are considered in their own right. We investigate the impact of small sample sizes due either to inadequate recruitment targets or high attrition rates on the results of fatty acid intervention trials. Data from our large trial of DHA supplementation during pregnancy with minimal attrition are used for illustration. Our findings demonstrate that recruiting fewer participants or neglecting to follow up difficult participants can lead to substantially different results and alter conclusions about the effectiveness of the intervention. Developing strategies for overcoming these inadequacies should be a top priority in fatty acid intervention trials.

1. Introduction

Randomised controlled trials (RCTs) are generally considered to be the gold standard for judging the effectiveness of an intervention [1]. However, it can be difficult to draw meaningful conclusions from RCTs when the number of participants in the trial (the sample size) is inadequate. Statisticians would describe such a trial as underpowered to detect clinically important treatment effects, if they are present. Underpowered trials are problematic as, despite the enthusiasm of many investigators, they are more likely to produce significant findings that are the result of chance, rather than a real effect of the intervention [2, 3]. Underpowered trials have also been deemed unethical, as they expose participants to interventions with little chance of providing a clear answer regarding their effectiveness [4]. Trials that become underpowered due to high attrition rates (participant losses due to various causes) are even more problematic, since they can produce biased estimates of the treatment effect. Such bias is likely to occur when the participants who withdraw or are lost to follow up have different characteristics or outcomes than those participants who provide complete data, or when the attrition rate differs between treatment groups [5-7]. Attrition can also reduce the generalisability of the trial results [5]. It is therefore crucial to ensure that RCTs have adequate sample sizes with sufficient power to detect clinically important treatment effects, by choosing appropriate recruitment targets and minimising attrition rates.

Conducting adequately sized trials in fatty acid research can be especially challenging. Since background levels of the fatty acid of interest are present in the control group due endogenous synthesis and background diet intake, the difference between treatment groups in the fatty acid of interest can be reduced [8]. As a result, the difference in outcomes that is achievable with a fatty acid intervention can be smaller than for other types of interventions, and hence a larger sample size may be needed. Despite this, many RCTs of fatty acid interventions suffer from small sample sizes due to inadequate recruitment targets and/or high attrition rates, as highlighted by systematic reviews of fatty acid interventions [e.g. 9, 10]. A reminder of the importance of conducting adequately sized trials in this setting is therefore warranted.

The aim of this article is to demonstrate how trials involving small numbers of participants can lead to questionable results, using data from our large-scale Docosahexaenoic acid (DHA) to Optimize Mother Infant Outcome (DOMInO) trial with minimal attrition for illustration [11].

2. Patients and Methods

2.1 The DOMInO Trial

The DOMInO trial was a double-blind, multicenter RCT conducted in five maternity hospitals in Australia between 2005 and 2009 (Australian and New Zealand Clinical Trials Registry Identifier ACTRN12605000569606; anzctr.org.au) and has been described in detail previously [11]. Briefly, women with a singleton pregnancy between 18 and 21 weeks' gestation who were not already taking a prenatal supplement containing DHA were eligible to participate. Women providing written informed consent were randomised to the DHA or control group and were asked to consume three DHA-rich fish oil capsules or vegetable oil capsules per day, respectively, from trial entry until delivery. All procedures were conducted in accordance with the approval of the relevant Human Research Ethics Committees at each maternity hospital.

The aim of the DOMInO trial was to determine the effect of DHA supplementation during pregnancy on postnatal depression in the women, and cognitive and language development in the infants. Postnatal depression was assessed at six weeks and six months postpartum using the Edinburgh Postnatal Depression Scale (EPDS) [12] and a score of 12 or more was used to indicate probable depression. Cognitive and language development were assessed at 18 months of age (corrected for premature birth) using the Bayley Scales of Infant and Toddler Development, Third Edition (Bayley-III) [13]. The Bayley-III produces standardised scores with a mean of 100 and a standard deviation of 15, where lower scores represent poorer performance.

High recruitment targets were set for the DOMInO trial [11]. Sample size calculations indicated that 2280 women (1140 per group) should be recruited to detect a clinically important 4.2% absolute reduction in the prevalence of postnatal depression from 16.9% in the control group with 80% power. A 4.2% reduction was chosen on the basis of epidemiological data that

suggested a 7-8% reduction [14] and the possibility that part of this effect may be due to residual confounding, while the control group prevalence of 16.9% was estimated from Australian population data [15]. Only 630 infants (315 per group) were required to detect a clinically meaningful 5 point improvement in the mean cognitive and language scores separately for boys and girls, based on 80% power and the known standard deviation of the Bayley-III standardised scores of 15 [13]. A 5 point improvement was of interest as previous studies showing differences of 4-5 points had prompted changes in health policy [16, 17]. Developmental assessment was therefore planned for only a subset of infants to minimise both the cost of the trial and the burden on participants. A number of strategies for minimising attrition were implemented, including collecting up to four alternative contacts for participants at trial entry, and keeping in contact via mailing a regular newsletter that included a change of address slip.

2.2 Statistical Methods

A post-hoc exploratory analysis of data from the DOMInO trial was performed to determine how the results might have changed if the sample size had been smaller due to recruiting fewer participants or neglecting to follow up difficult participants. Language scores at 18 months of age and postnatal depression at six months postpartum were used for illustration.

To investigate the potential impact of small sample sizes due to inadequate recruitment targets, we estimated the effect of DHA supplementation on language scores in three different groups of DOMInO participants: (i) all infants who completed the developmental assessment; (ii) 100 different random samples of 50 infants (25 per group) selected from the infants who completed the developmental assessment; and (iii) after every 50 infants had completed the developmental assessment. In each case, the mean language score was compared between the DHA and control groups using a two-sample t-test. The percentage of random samples of 50 infants where the estimated treatment effect was expected to exceed the clinically important difference of 5 points, just by chance, was calculated based on properties of the normal distribution.

To investigate the potential impact of small sample sizes due to high attrition rates, we estimated the effect of DHA supplementation on postnatal depression at six months postpartum in two

different subsets of DOMInO participants: (i) all women who completed the EPDS; and (ii) after excluding women who were difficult to follow up, defined as completing the EPDS more than 30 days after it was due, or requiring telephone follow up based on incomplete responses to the questionnaire. For each of these subsets, the proportion of women who had probable depression (EPDS>12) was compared between the DHA and control groups using a chi-square test.

3. Results

Recruitment targets were exceeded for the DOMInO trial to ensure adequate sample sizes would remain after any attrition. A total of 2399 women (1197 DHA, 1202 control) were enrolled in the trial and 726 infants (351 DHA, 375 control) were selected for the developmental assessment Attrition rates for the trial were kept to a minimum. EPDS scores at six months postpartum were obtained from 2341 (97.6%) women, while 692 (95.3%) infants completed the language assessment at 18 months of age. The primary findings of the trial have been reported in detail elsewhere [11].

3.1 Impact of Inadequate Recruitment Targets

Based on all 692 infants who completed the language assessment, the mean (SD) language score from the Bayley-III was 96.5 (13.6) for the DHA group and 98.2 (15.3) for the control group. The difference in means (DHA minus control) was -1.7 points (95% confidence interval, -3.9 to 0.5), indicating that there was insufficient evidence to support the hypothesis that the mean language score differed between the treatment groups (P=0.13). Similar results have been reported for this outcome elsewhere using more complex statistical methods [11]. A high degree of confidence can be placed in these results, due to the large sample size and minimal attrition rate.

In order to demonstrate the effect of inadequate recruitment targets on these results, we estimated the effect of DHA supplementation on language for 100 random samples of 50 infants (25 per group). Treatment effect estimates ranged from a reduction in the mean score of 8.4 points, to an increase in the mean score of 7.8 points, depending on the random sample chosen (Figure 1).

There was one random sample where DHA supplementation had a significant positive effect and three random samples where it had a significant negative effect on the mean language score. These small samples therefore could have led to conclusions that differed from the main trial findings when all infants were included in the analysis. None of the other random samples showed a statistically significant effect of DHA supplementation on language. For 20% of random samples, the estimated treatment effect exceeded the clinically important difference in mean language scores of 5 points. If DHA supplementation had no effect on language, a difference of 5 points or more would be expected just by chance in 24% of random samples. These findings are therefore consistent with the lack of effect of DHA supplementation on language seen among all infants.

Another way to demonstrate the effect of inadequate recruitment targets on the trial results is to examine the cumulative effect of DHA supplementation on language after every 50 infants completed the developmental assessment. Treatment effect estimates varied greatly in the early stages of the DOMInO trial, ranging from an increase in the mean score of 3.1 points when only 50 infants had been assessed, to a reduction of 3.6 points when 300 infants had been assessed (Figure 2). The treatment group differences achieved statistical significance in favour of the control group when the sample size was 300, 350, 450 or 500 infants and the smallest p-value observed was 0.02. As the trial progressed and the sample size increased to the target required to ensure adequate power, the estimated effect of DHA supplementation stabilised and the mean language scores did not differ significantly between the treatment groups.

3.2 Impact of High Attrition Rates

Based on all 2341 women who completed the EPDS at six months postpartum, 9.6% of women in the DHA group and 11.2% of women in the control group had scores indicating probable depression. Despite the large sample size and minimal attrition rate, there was no evidence to suggest that DHA supplementation reduced the prevalence of postnatal depression (P=0.21). Similar findings have previously been reported for this outcome using more complex statistical methods [11].

As with all trials, some DOMInO participants were easier to follow up than others. There were 784 (33.5%) women who were difficult to follow up for assessing postnatal depression at six months postpartum. These women had different characteristics compared with women who were easy to follow up; they were more likely to be younger, less educated, smokers, have a history of depression and belong to the DHA group (Table 1). Such differences in observed (or unobserved) characteristics can introduce bias into estimated treatment effects when these women are excluded from the analysis. In the subset of 1557 (66.5%) women who were easy to follow up, DHA supplementation was found to reduce postnatal depression, with a prevalence of 8.2% observed in the DHA group and 11.1% in the control group (P=0.05). This result contrasts with the null finding when all women were included in the analysis and should be considered unreliable due to the potential biases.

4. Discussion and Conclusions

This report highlights the importance of achieving adequate sample sizes in fatty acid intervention trials. The large recruitment targets and high follow up rates attained in our DOMInO trial of DHA supplementation in pregnancy provided an opportunity to illustrate what can go wrong when trials are too small. While problems with inadequate sample sizes have been discussed in general previously [e.g. 2, 3-7], small trials continue to be conducted in fatty acid research and a review of the issues in this context is timely. The results of our exploratory analyses demonstrate that less confidence can be placed in the results of a small trial, since recruiting more participants or improving follow up rates could lead to substantially different results. Small trials can still be valuable for contributing to meta-analyses or providing pilot data to help plan future trials with adequate power, but are unlikely in themselves to provide a clear answer regarding the effectiveness of the intervention.

We found that DHA supplementation could have a significant positive effect, a significant negative effect, or no significant effect on language development in different random samples of infants who completed the language assessment in the DOMInO trial. If DHA supplementation really has no impact on language, significant positive and negative effects can be expected in some samples of infants, just by chance. Although this is true in samples of any size, the likely

magnitude of chance treatment effects decreases as the sample size increases. This means that any false positive findings are expected to be smaller in larger samples, and are therefore less likely to be considered clinically important. As an example, for a trial with a total sample size of only 50 infants, we found there is a 24% chance of observing a clinically meaningful difference in mean language scores between the treatment groups if DHA supplementation actually has no effect on language. This can be reduced to less than a 1% chance if the total sample size for the trial is increased to 250 infants. These calculations emphasize the importance of replicating findings reported from trials with few participants in adequately powered trials before considering changing public health policy or clinical practice guidelines.

The target sample size for a trial should depend on a number of factors, including the minimum clinically important difference to be detected between the treatment groups and the desired power [4, 18, 19]. In fatty acid intervention trials, the presence of the fatty acid of interest in the control group due to endogenous synthesis and dietary intake should also be considered, as this can reduce the difference between treatment groups in the fatty acid of interest and hence the achievable difference in the primary outcome, which increases the sample size required [18]. Non-compliance is another important factor to consider in sample size calculations for fatty acid intervention trials. While participants in the intervention group can fail to take the intervention in any trial, the risk of participants in the control group taking the intervention is particularly high in fatty acid research when readily available foods and dietary supplements contain the fatty acid of interest. Compliance with the assigned treatment should be encouraged throughout the trial and the sample size can be increased to account for non-compliance [18]. By choosing a target sample size that provides high power to detect a clinically meaningful treatment effect, should one exist, and low risk of finding such an effect just by chance, researchers are more likely to be able to draw meaningful conclusions from their trials.

Our analysis of the postnatal depression data from the DOMInO trial illustrates the problems with high attrition rates in fatty acid intervention trials. We found that women who were difficult to follow up had different characteristics than those who were not difficult to follow up, and that excluding the difficult to follow up women from the analysis resulted in different conclusions about the effectiveness of the intervention. This emphasizes the need to achieve high follow up

rates. In fatty acid intervention trials, the maximum follow up rate that can realistically be attained depends on the population being studied, the nature and intensity of the intervention, and the length of follow up planned. While attrition rates of up to 5% have traditionally been considered acceptable [20], greater emphasis is now being placed on the potential implications of any losses to follow up, rather than defining acceptable attrition rates [5]. Attrition is likely to introduce bias because the participants who withdraw or are lost to follow up inevitably have different characteristics or outcomes than those participants who provide complete data [5-7]. Complex statistical methods are available for dealing with missing data due to attrition [21], but these rely on untestable assumptions about the missing values and should not become a substitute for time and effort spent following up trial participants. Strategies for minimising attrition are therefore crucial, and these have been discussed elsewhere [7, 22]. By achieving high follow up rates, researchers are more likely to draw appropriate conclusions from their trials.

In summary, our findings highlight the importance of ensuring RCTs have adequate sample sizes by recruiting sufficient participants and minimising attrition rates. Achieving an adequate sample size should be considered a top priority in fatty acid intervention trials.

Acknowledgements

Conflict of Interest

The authors have no conflicts of interest to declare.

Authorship

LNY, MM and RAG designed the research; MM, AJM, JQ and RAG designed the DOMInO trial; LNY performed the statistical analysis and drafted the initial manuscript; all authors critically revised the manuscript; LNY and RAG had responsibility for final content. All authors read and approved the final manuscript.

References

[1] D.L. Sackett, W.M.C. Rosenberg, J.A.M. Gray, R.B. Haynes, W.S. Richardson, Evidence based medicine: what it is and what it isn't, British Medical Journal, 312 (1996) 71-72.

[2] J.P.A. Ioannidis, Why most published research findings are false, Plos Medicine, 2 (2005) 696-701.

[3] J.A.C. Sterne, G.D. Smith, Sifting the evidence - what's wrong with significance tests?, British Medical Journal, 322 (2001) 226-231.

[4] J.B. Carlin, L.W. Doyle, Statistics for clinicians: 7: sample size, J Paediatr Child H, 38 (2002) 300-304.

[5] M.S. Fewtrell, K. Kennedy, A. Singhal, et al., How much loss to follow-up is acceptable in long-term randomised trials and prospective studies?, Archives of Disease in Childhood, 93 (2008) 458-461.

[6] V.M. Montori, G.H. Guyatt, Intention-to-treat principle, Can Med Assoc J, 165 (2001) 1339-1341.

[7] K.F. Schulz, D.A. Grimes, Sample size slippages in randomised trials: exclusions and the lost and wayward, Lancet, 359 (2002) 781-785.

[8] M.J. James, T.R. Sullivan, R.G. Metcalf, L.G. Cleland, Pitfalls in the use of randomised controlled trials for fish oil studies with cardiac patients, British Journal of Nutrition, (2014).

[9] N. Stratakis, M. Gielen, L. Chatzi, M.P. Zeegers, Effect of maternal n-3 long-chain polyunsaturated fatty acid supplementation during pregnancy and/or lactation on adiposity in childhood: a systematic review and meta-analysis of randomized controlled trials, Eur J Clin Nutr, 68 (2014) 1277-1287.

[10] R.E. Cooper, C. Tye, J. Kuntsi, E. Vassos, P. Asherson, Omega-3 polyunsaturated fatty acid supplementation and cognition: A systematic review and meta-analysis, J Psychopharmacol, 29 (2015) 753-763.

[11] M. Makrides, R.A. Gibson, A.J. McPhee, L. Yelland, J. Quinlivan, P. Ryan, Effect of DHA supplementation during pregnancy on maternal depression and neurodevelopment of young children: a randomized controlled trial, Journal of the American Medical Association, 304 (2010) 1675-1683.

[12] J.L. Cox, J.M. Holden, R. Sagovsky, Detection of postnatal depression - development of the 10-item Edinburgh Postnatal Depression Scale, British Journal of Psychiatry, 150 (1987) 782-786.

[13] N. Bayley, Bayley Scales of Infant and Toddler Development, Third Edition, Pearson Education Inc., San Antonio, TX, 2006.

[14] J. Golding, C. Steer, P. Emmett, J.M. Davis, J.R. Hibbeln, High levels of depressive symptoms in pregnancy with low omega-3 fatty acid intake from fish, Epidemiology, 20 (2009) 598-603.

[15] S. Brown, J. Lumley, Maternal health after childbirth: results of an Australian population based survey, BJOG, 105 (1998) 156-161.

[16] D.O. Carpenter, Effects of metals on the nervous system of humans and animals, International journal of occupational medicine and environmental health, 14 (2001) 209-218.

[17] T. Walter, I. Deandraca, P. Chadud, C.G. Perales, Iron-deficiency anemia: adverse-effects on infant psychomotor development, Pediatrics, 84 (1989) 7-17.

[18] A. Kirby, V. Gebski, A.C. Keech, Determining the sample size in a clinical trial, Medical Journal of Australia, 177 (2002) 256-257.

[19] B. Rohrig, J.B. du Prel, D. Wachtlin, R. Kwiecien, M. Blettner, Sample size calculation in clinical trials part 13 of a series on evaluation of scientific publications, Dtsch Arztebl Int, 107 (2010) 552-556.

[20] D.L. Sackett, S.E. Strauss, W.S. Richardson, W. Rosenberg, R.B. Haynes, Evidence based medicine : how to practice and teach EBM, 2nd ed., Churchill Livingstone, Edinburgh ; New York, 2000.

[21] R.J.A. Little, D.B. Rubin, Statistical Analysis with Missing Data, 2nd ed., Wiley, New York, 2002.

[22] D.F. Polit, B.M. Gillespie, Intention-to-treat in randomized controlled trials: recommendations for a total trial strategy, Res Nurs Health, 33 (2010) 355-368.

Table 1. Characteristics of women by whether they were easy or difficult to follow up for assessing postnatal depression at six months postpartum

Characteristic	Easy (n=1557)	Difficult (n=784)
Age at trial entry (years)[*]	29.8 (5.3)	27.3 (6.0)
Completed secondary education	1066 (68.5)	414 (52.8)
Completed further education	1098 (70.5)	498 (63.5)
Smoked at trial entry or leading up to pregnancy	420 (27.0)	329 (42.0)
Previous or current medical diagnosis of depression	346 (22.2)	220 (28.1)
Assigned to DHA supplementation group	758 (48.7)	420 (53.6)

Values are number (percent) unless otherwise indicated.

[*] Values are mean (standard deviation).

Figure Legends

Figure 1. Difference in mean language score (DHA minus control) for 100 different random samples of 50 infants. Stars indicate statistically significant differences (P<0.05). Dots indicate non-significant differences.
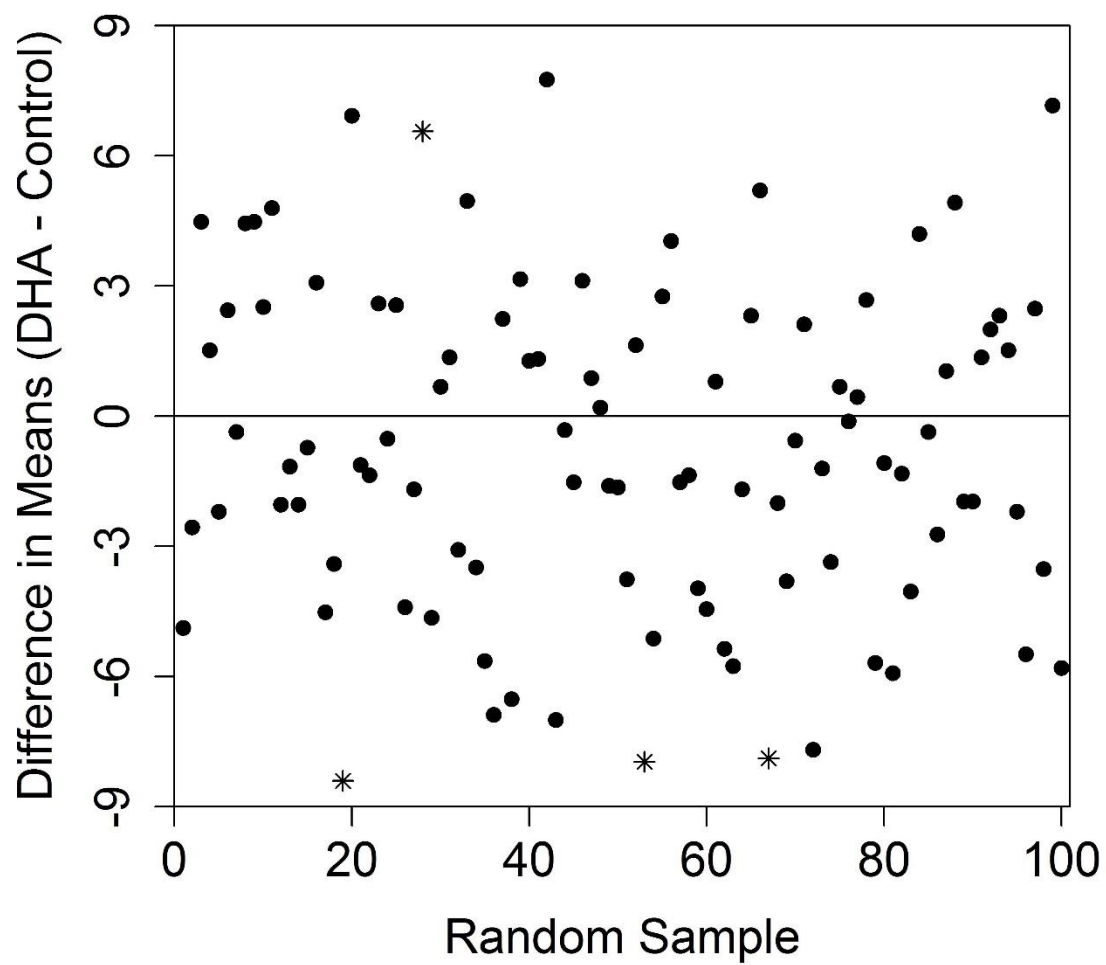
Figure 2. Difference in mean language score (DHA minus control) after every 50 infants completed the developmental assessment. Stars indicate statistically significant differences (P<0.05). Dots indicate non-significant differences.