Statistical design considerations applicable to clinical trials of iodine supplementation in pregnant women who may be mildly iodine deficient[1,2]

*James F Troendle**

Office of Biostatistics Research, Division of Cardiovascular Sciences, National Heart, Lung, and Blood Institute, NIH, Bethesda, MD

ABSTRACT

No large, randomized, placebo-controlled trial of iodine supplementation in pregnant women in a region of mild or moderate iodine deficiency has been completed in which a primary outcome measure was an assessment of the neurobehavioral development of the offspring at age ≥2 y. In this article, I discuss considerations for the design of such a trial in a region of mild iodine deficiency, with a focus on statistical methods and approaches. Exposure and design issues include the ethics of using a placebo, the potential for overexposure to iodine, and the possibility of community randomization. The main scientific goal of the trial is important in determining the follow-up period. If the goal is to determine whether iodine supplementation during pregnancy improves neurobehavioral development in the offspring, then follow-up should continue until a reasonably reliable assessment can be conducted, which might be at age ≥2 y. Once the timing of assessment is decided, the impact of potential loss to follow-up should be considered so that appropriate statistical methods can be incorporated into the design. The minimum sample size can be calculated by using a sample size formula that incorporates noncompliance and assumes that a certain proportion of study participants do not have any outcome observed. To have sufficient power to detect a reasonably modest difference in neurobehavioral development scores using an assessment tool with an SD of 15, a large number of participants (>500/group) is required. The minimum adequate number of participants may be even larger (>1300/group) depending on the magnitude of the difference in outcome between the supplementation and placebo groups, the estimated proportion of the iodine-supplementation group that fails to take the supplement, and the estimated proportion of pregnancies that do not produce outcome measurements. *Am J Clin Nutr* 2016;104(Suppl):924S–7S.

Keywords: clinical trials, iodine supplementation, outcomes measurement, pregnancy, statistical methods

INTRODUCTION

Designing a trial of iodine supplementation in pregnant women is challenging for several reasons. One is the concern that supplementation could lead to iodine-induced hyperthyroidism or hypothyroidism in susceptible women whose dietary iodine intakes are already high (1, 2). This question is made more difficult because urinary iodine concentration (UIC)[3] measured in spot urine samples, although appropriate for population assessment of iodine status, is not a useful measure of iodine status in individuals because it reflects large day-to-day variation in dietary iodine intake (3). Other challenges include how to measure outcome and how long to require follow-up for useful outcome measures. Obtaining outcomes on all subjects is expected to require substantial effort and resources, making such a trial expensive. The inevitable loss to follow-up is another challenge; the protocol for analysis of a long-term trial should include some plan to deal with such loss.

Two placebo-controlled trials of iodine supplementation in pregnant women, Pregnancy Iodine and Neurodevelopment in Kids (PINK) and Maternal Iodine Supplementation and Effects on Thyroid Function and Child Development (MITCH), were recently concluded. The PINK trial, conducted in Australia (4), was stopped early because the Medical Research Council of Australia recommended universal iodine supplementation of pregnant women. The MITCH trial, conducted in India and Thailand (5), has completed recruitment, but the results have not yet been reported. Both trials assessed cognitive and motor development in the offspring at age 2 y using the Bayley scales. This was only one of several outcome measures of the MITCH trial, which set out to determine whether supplementation with iodine at 200 μg/d in regions of mild to moderate iodine deficiency improves maternal and newborn thyroid function, pregnancy outcome, birth weight, and infant and toddler growth, as well as cognitive and motor development. No large, randomized, placebo-controlled trial of iodine supplementation in pregnant women in a region of mild or moderate iodine deficiency has been completed in which a primary outcome measure was the neurobehavioral development of the offspring at age ≥2 y. In the present article, I discuss the primary issues in the design of such a trial, with a focus on statistical considerations.

[1] Presented at the workshop "Maternal Iodine Supplementation: Clinical Trials and Assessment of Outcomes" held by the NIH Office of Dietary Supplements in Rockville, MD, 22–23 September 2014.

[2] The author reported no funding received for this study.

*To whom correspondence should be addressed. E-mail: jt3t@nih.gov.

[3] Abbreviations used: MITCH, Maternal Iodine Supplementation and Effects on Thyroid Function and Child Development; PINK, Pregnancy Iodine and Neurodevelopment in Kids; UIC, urinary iodine concentration.

First published online August 17, 2016; doi: 10.3945/ajcn.115.110403.

ISSUES RELATED TO EXPOSURE AND DESIGN

The first question to ask when designing a trial of iodine supplementation is whether it is ethically acceptable to use a placebo as a control. In regions of mild iodine deficiency, at least, there is no compelling evidence that iodine supplementation of mothers during gestation improves clinical outcomes in their infants (6). Thus, it would seem that there is no ethical problem with using a placebo control in such regions. Furthermore, when iodine deficiency is observed in pregnant women in the US population, it can nearly always be described as mild (7). If it is ethically acceptable to use a placebo control, then feasibility must be established; the feasibility of the use of a placebo control in regions of mild iodine deficiency seems to have been demonstrated by the MITCH and PINK trials.

The next question is about the potential for overexposure to iodine. Upper-limit thresholds for daily iodine intake in individuals on the basis of UIC have not been established. The day-to-day (i.e., within-subject) variability of UIC is high; for example, the within-subject CV was reported to be 38% in 52- to 77-y-old Swiss women (8). For that reason, it is difficult to establish thresholds for excessive iodine intake in individuals on the basis of single spot urine samples. If there is concern about excessive intake in susceptible women, then either a more robust measure than UIC is needed (e.g., the serum concentration of thyroid-stimulating hormone, a direct measure of thyroid function) or sequential testing of study subjects with a UIC value above a given cutoff in ≥ 1 sample should be considered. Of course, sequential testing would increase the cost of the trial according to the number of extra urine samples analyzed; therefore, the UIC value that triggers retesting should be chosen so as to satisfy safety concerns without adding unnecessary expense. If sequential testing is implemented, then the study design could incorporate reduction in or withdrawal of iodine supplementation in subjects with consistently high UIC.

A third question related to study design is whether it might be possible to randomly assign communities instead of individuals. Community randomization would require some form of community consent and, in principle, could save much of the cost associated with obtaining individual consent and managing both the randomization process and the conduct of the trial. However, cluster-randomized trials (including those in which communities are the clusters) require a much larger total number of subjects than individual-randomized trials because each cluster is considered a single unit for the purposes of statistical analysis (9). Therefore, if the outcome requires intense follow-up, the larger total sample size required would likely make a community-randomized trial more expensive than an individual-randomized trial.

ISSUES RELATED TO OUTCOME AND ANALYSIS

The primary outcome measures and the timing of assessment can be decided once the scientific goal of the trial is established. Is the goal solely to determine whether the thyroid function of pregnant women with mild or moderate iodine deficiency is altered by iodine supplementation? If so, then follow-up can end at delivery. Is the goal instead (or also) to determine whether some aspect of neurobehavioral development is affected by maternal iodine supplementation during pregnancy? If so, then follow-up should continue until an age when a reasonably reliable assessment of the outcome measures of interest can be conducted. As discussed elsewhere in this supplement issue, research is needed to establish which specific neurodevelopmental domains and corresponding cognitive, behavioral, and/or motor tasks specific to infants and toddlers are likely to be the most sensitive to maternal iodine supplementation in pregnant women with mild to moderate iodine deficiency (10–12). Although the performance of infants and toddlers on a standardized test of neurobehavioral development is often the primary outcome measure of cohort studies addressing possible effects of maternal exposures, confounding by maternal intelligence, the home environment, and other covariates is often an issue affecting the interpretation of results (13, 14). Because a more even distribution of covariates across treated and nontreated participants can be achieved in a clinical trial, the finding of an effect (or no effect) of maternal iodine supplementation on neurobehavioral development scores in infants and young children could be more definitive than in incompletely controlled cohort studies.

Once the primary outcome measure and the timing of assessment are decided, the impact of potential loss to follow-up must be considered. In studies that require the assessment of the offspring, there will likely be loss before birth, dropout at birth, and dropout thereafter. In the case of dropout after some outcome values have been obtained, the analysis must account for the missing outcome data. Although there are several statistical methods that can be used to try to account for missing outcome (including inverse probability weighting and multiple imputation) (5), it is probably wise to plan on using a repeated-measures model (linear mixed model) (15). The repeated-measures model allows for the direct influence of the mean of an unobserved final outcome measurement from observed interim measurements through the correlation between the interim and final measurements. Although requiring interim measurements increases the cost of the trial, they could be crucial to reducing the influence of missing data on the final analysis. If the unobserved values at the final time point are missing at random, conditional on the observed values at interim time points, then maximum likelihood analysis of a repeated-measures model of the observed values is valid (16–18). This assumption is more tenable the more interim values one obtains. Thus, the use of a repeated-measures model (along with a design that includes many interim measurements) can help minimize the potential uncertainty in the final trial results attributable to whatever assumptions were needed to account for missing data in the analysis.

For any outcome based on the offspring, there will be some pregnancies that are lost to follow-up as a result of fetal loss or stillbirth and thus no outcome values are obtained. How can this be addressed? A repeated-measures model will not include such pregnancies in the analysis because there is no outcome available. Multiple imputation could still be used if some other covariates can predict this event. Even a poorly informative imputation model would be preferable to not including such pregnancies in the analysis (19). However, a poorly informative imputation model could still leave the final analysis prone to informative missingness and bias if the fetuses that are lost might be expected (had they lived) to have poor outcomes. If fetal loss is also positively or negatively correlated to maternal iodine supplementation, then there is the possibility of additional bias. In this case, one could try modeling the dropout mechanism (20). A sensitivity analysis to test the robustness of the result to

potentially informative missing data would then consist of trying a reasonable selection of models for the dropout mechanism. Another approach might be a nonparametric analysis based on ranks that gives the worst rank to those with the earliest dropout time (21).

ISSUES RELATED TO SAMPLE SIZE CALCULATION

Once the method of analysis is chosen, sample size can be considered. Here I assume a placebo-controlled trial to determine the effect of maternal iodine supplementation during pregnancy on neurobehavioral development in infants and toddlers. I assume the primary analysis is based on a repeated-measures model in which repeated assessments of neurobehavioral development are performed up to a certain final age. The primary comparison is the model-predicted mean difference in the final assessment among the treatment groups. For simplicity, the sample size is based on just the final assessment's comparison between the iodine and placebo groups. The actual analysis will have larger power, and the power will increase in tandem with the correlation between the interim assessments and the final assessment (9).

The assessment used is assumed to be a continuous ratio scale (i.e., it describes how much or how many), although this is not expected to be exactly true if the neurobehavioral development score is an ordinal sum of the scores for individual tasks or subcomponents. However, if many subcomponents are assessed, it seems reasonable to regard the assessment as an approximation of an underlying continuous attribute. The underlying attribute will also be approximately normally distributed in the population, making sample size calculations simple. Let the true mean difference in assessment scores (iodine group – placebo group) be δ and the SD be σ. Furthermore, let α be the desired significance level and let $1-\beta$ be the desired power to detect a difference in mean assessment between the groups. The standard formula for the sample size required in each group is then as follows:

$$n = \frac{2\left[\Phi^{-1}\left(1-\frac{\alpha}{2}\right) - \Phi^{-1}(\beta)\right]^2}{\left(\frac{\delta}{\sigma}\right)^2} \qquad (1)$$

where Φ is the cumulative distribution function of the standard normal distribution.

There are two further considerations that affect calculation of the required sample size. First, noncompliance will be expected for a certain proportion of the study entrants. This is always a concern in clinical trials, but in regions without obvious iodine deficiency it might be especially large. Let p_{nc} be the sum of the estimated proportion of subjects randomly assigned to receive iodine supplementation who do not take the supplements and the estimated proportion of those randomly assigned to placebo who do take supplements. The effect of such noncompliance is to effectively reduce δ to $(1 - p_{nc})\delta$. A second consideration is to account for those without any measurement of outcome. Although one should also use imputation or some other method to adjust for these subjects in the analysis, if the proportion is relatively large it can significantly affect power and so should also be accounted for in the sample size calculation to ensure a sufficient number of subjects with outcome measurements; here, the word "subjects" is intended to convey either the mothers or their offspring, depending on the outcome measure. Let p_{no} be an estimated proportion of subjects for which there are no outcome assessments. To obtain a sufficient number of subjects with the observed outcome, one should randomly assign $n/(1 - p_{no})$ to each group. Thus, to account for noncompliance and also to account for the lack of outcome measurement in some subjects, the number of subjects randomly assigned to each group should be:

$$n = \frac{2\left[\Phi^{-1}\left(1-\frac{\alpha}{2}\right) - \Phi^{-1}(\beta)\right]^2}{\left(\frac{(1-p_{nc})\delta}{\sigma}\right)^2 (1-p_{no})} \qquad (2)$$

As an example of how the formula could be used to plan a trial of maternal iodine supplementation during pregnancy, consider a trial in which 90% power with a 5% significance level is desired and the primary outcome measure is neurobehavioral development in the offspring at the final age. Here we assume that the SD of the assessment instrument is 15 because that is the SD of the composite score equivalents of each of the 5 Bayley subscales in the general population (22). It is understood that, in any particular trial population, the SD of the outcome measure chosen could be larger or smaller than 15. **Table 1** gives the required sample size in each group for several true mean differences (δ) and true proportions (p_{nc} and p_{no}). As shown in the table, a trial must be large (>500/group) to have sufficient power to detect a reasonably modest difference in neurobehavioral development under the specified conditions.

CONCLUSIONS

Large studies of iodine supplementation in pregnant women in areas of mild iodine deficiency can be safely conducted with a placebo control. Establishing assessment of neurobehavioral development as the primary outcome measure calls for a follow-up period long enough to ensure minimal validity. Large trials (>500/group) are needed to have sufficient power to detect modest, exposure-related differences in neurobehavioral development outcomes, assuming an assessment tool with an SD of 15. Although expensive, such studies could nevertheless be of

TABLE 1
Required sample size per group for a trial in women individually randomly assigned to receive iodine supplementation or placebo during pregnancy[1]

δ	p_{nc}	p_{no}	n
4	0.2	0.2	579
4	0.3	0.2	753
3	0.2	0.2	1025
3	0.3	0.2	1339

[1]The analysis assumes that the outcome measure is the score on an assessment of cognitive development (or other component of neurodevelopment) with an SD of 15. The table shows the required sample size per group (n) based on 90% power. The table shows values of n calculated for 2 possible true mean differences in assessment score (δ) and 2 possible values of the true sum of proportions of noncompliant subjects (p_{nc}). The true proportion of subjects without an observed outcome (p_{no}) is assumed to be the same in each group.

societal importance if a modest increase in neurobehavioral development outcomes at the age of final assessment is positively associated with later achievement.

I thank Gay Goodman, Iodine Initiative Consultant to the NIH Office of Dietary Supplements, for her contributions in the course of providing expert scientific and technical review.

JFT wrote the manuscript and read and approved the final manuscript. The author reported no conflicts of interest related to the study.

REFERENCES

1. Zhou SJ, Anderson AJ, Gibson RA, Makrides M. Effect of iodine supplementation in pregnancy on child development and other clinical outcomes: a systematic review of randomized controlled trials. Am J Clin Nutr 2013;98:1241–54.
2. Leung AM, Braverman LE. Iodine-induced thyroid dysfunction. Curr Opin Endocrinol Diabetes Obes 2012;19:414–9.
3. Rasmussen LB, Ovesen L, Christiansen E. Day-to-day and within-day variation in urinary iodine excretion. Eur J Clin Nutr 1999;53:401–7.
4. Australian New Zealand Clinical Trials Registry. Pregnancy Iodine and Neurodevelopment in Kids (PINK) [cited 2014 Nov 14]. Available from: https://www.anzctr.org.au/Trial/Registration/TrialReview.aspx?ACTRN=12610000411044.
5. ClinicalTrials.gov [Internet]. Maternal Iodine Supplementation and Effects on Thyroid Function and Child Development (MITCH) [cited 2014 Nov 14]. Available from: https://clinicaltrials.gov/ct2/show/NCT00791466.
6. Zimmermann MB. The effects of iodine deficiency in pregnancy and infancy. Paediatr Perinat Epidemiol 2012;26(Suppl 1):108–17.
7. Caldwell KL, Makhmudov A, Ely E, Jones RL, Wang RY. Iodine status of the U.S. population, National Health and Nutrition Examination Survey, 2005–2006 and 2007–2008. Thyroid 2011;21:419–27.
8. König F, Andersson M, Hotz K, Aeberli I, Zimmermann MB. Ten repeat collections for urinary iodine from spot samples or 24-hour samples are needed to reliably estimate individual iodine status in women. J Nutr 2011;141:2049–54.
9. Murray DM. Design and analysis of group-randomized trials. New York: Oxford University Press; 1998.
10. Bauer PJ, Dugan JA. Suggested use of sensitive measures of memory to detect functional effects of maternal iodine supplementation on hippocampal development. Am J Clin Nutr 2016;104(Suppl):935S–40S.
11. Bell MA, Ross AP, Goodman G. Assessing infant cognitive development after prenatal iodine supplementation. Am J Clin Nutr 2016;104(Suppl):928S–34S.
12. Ershow AG, Goodman G, Coates PM, Swanson CA. Research needs for assessing iodine intake, iodine status, and the effects of maternal iodine supplementation. Am J Clin Nutr 2016;104(Suppl):941S–9S.
13. Winneke G. Appraisal of neurobehavioral methods in environmental health research: the developing brain as a target for neurotoxic chemicals. Int J Hyg Environ Health 2007;210:601–9.
14. Mink PJ, Goodman M, Barraj LM, Imrey H, Kelsh MA, Yager J. Evaluation of uncontrolled confounding in studies of environmental exposures and neurobehavioral testing in children. Epidemiology 2004;15:385–93.
15. Vonesh EF, Chinchilli VM. Linear and nonlinear models for the analysis of repeated measurements. New York: Marcel Dekker; 1997.
16. Carpenter JR, Kenward MG. Missing data in randomized controlled trials—a practical guide. Birmingham (United Kingdom): National Institute for Health Research; 2007.
17. Carpenter J, Pocock S, Lamm CJ. Coping with missing data in clinical trials: a model-based approach applied to asthma trials. Stat Med 2002;21:1043–66.
18. Little R, Yau L. Intent-to-treat analysis for longitudinal studies with drop-outs. Biometrics 1996;52:1324–33.
19. Graham JW. Missing data analysis: making it work in the real world. Annu Rev Psychol 2009;60:549–76.
20. Little RJA. Modeling the drop-out mechanism in repeated-measures studies. J Am Stat Assoc 1995;90:1112–21.
21. Finkelstein DM, Schoenfeld DA. Combining mortality and longitudinal measures in clinical trials. Stat Med 1999;18:1341–54.
22. Bayley N. Bayley Scales of Infant and Toddler Development - Third Edition. J Psychoeduc Assess 2007;25:180–90.