BREASTFEEDING MEDICINE
Volume 1, Number 2, 2006

The Design and Analysis of Studies in Premature Infants Using Human Donor Milk or Preterm Formula as Primary Nutrition: A Critique of Schanler et al.

MARTIN L. LEE

ABSTRACT

Nutritional studies of human milk in pre-term infants provide a unique challenge in clinical research. In this paper we review the general tenets of good clinical design and analysis and show how they might be properly applied in these situations. The recommendations are then compared with the approach used in a recent study by Shanler et al. It is concluded that future trials should consider these key statistical design and analytical issues.

INTRODUCTION

With the increased interest in obtaining human donor milk (DM) for potential complete or partial nutrition in babies born premature and, typically, at very low birth weights (<1500 g), there has been a strong push for well-designed clinical trials comparing DM with mother's own milk (MM) and preterm formula (PF). Certainly, whenever it is available, there is little question that MM is the best nutrition available for the newborn infant regardless of the birthweight; however, under conditions of prematurity this may not always be available. As a result, an obvious question for many neonatologists is whether DM (when available) is the best alternative or whether PF is as good under these circumstances? As noted, these questions can be answered properly only by clinical trials that employ proper statistical design, including sample size, and analytical techniques with the subsequently collected data. The purpose of this paper is to outline those principles so that published studies can be assessed properly on their merits and the conclusions drawn from these trials may be accepted. The recent article by Schanler et al.[1] brings these issues into focus, and it is the present author's intent to contrast their approach to these studies with what is believed to be appropriate design and analytical techniques. This paper focuses on this particular study because it is one of the only trials that has attempted to compare MM, DM, and PF. Narayanan et al.[2] evaluated MM (with occasional supplementation by DM and nightly use of PF) versus PF alone in a randomized fashion. Lucas and Cole[3] performed two parallel studies involving DM and PF in one instance (in which MM could be used in either group), and PF versus term formula in the other instance (again, in which MM could be used in either group). In none of these studies was there a specific attempt to compare all three preparations.

DESIGN

First and foremost with respect to the design and analysis of a controlled study, it is fundamental to identify the study arms that will be evaluated within the context of the trial. Clearly, in a study that seeks to use MM as a

Department of Biostatistics, UCLA School of Public Health and Prolacta Bioscience, Monrovia, California.

control group (or "gold standard"), there is no question that participants who intend to provide MM to their infant would not be expected to be part of any study randomization. That leaves the randomized arms of the study to the DM and PF possibilities. This fact raises interesting and unusual design issues. Typically, randomized multigroup trials apply the randomization process to all potential study groups. As a result, the sample size is determined on the basis of the primary goal to compare all of the groups simultaneously (otherwise, why include all of the groups?), and the primary statistical analysis is precisely that comparison. Furthermore, this is the legitimate statistical approach, because the randomization among the groups allows for it. How is one to determine the proper approach under a semi-experimental design in which there is a study with multiple groups, not all of whom are randomized? It might be argued that regardless of the nature of the trial randomization (and, thus, design in this regard) the inclusion of three separate groups of patients (MM, DM, and PF) necessitates that the sample size determination and, in turn, the analysis be performed as if all groups had been randomized. The only issue, of course, is legitimizing any statistical comparison with the nonrandomized arm. In order to do this properly, it is necessary to statistically adjust for potential group differences using a multivariate approach. This particular problem is discussed later.

On the other hand, it is quite possible that the goal of a trial being discussed here is simply to compare primarily only two groups, possibly DM and PF. This is easily done, because the patients who will participate can be randomized to these arms and the methodology for the design and analytical issues is well characterized.[4] However, one other primary issue must be addressed here that concerns whether the goal is to show that these two groups differ or are equivalent with respect to the desired primary endpoint. With respect to the former, most readers are familiar with the clinical trial that attempts to show one treatment is superior to another (usually a control), and the study design centers around the amount of improvement expected with the "new" treatment. (Statisticians usually refer to this improvement as delta, and the sample size for the study is calculated in order to have a high probability of demonstrating a statistical significant difference if delta is real ["power of the study," or 1-β], while minimizing the probability of finding statistical significance if delta is not real [significance level, or α].) However, the design becomes more difficult when the goal is to demonstrate equivalence. Equivalence in a statistical sense implies a clinical difference of less than some predefined amount, rather than a mathematical equivalent value. This is discussed in greater detail below. It is extremely important to recognize, though, that the lack of a significant difference between two groups is not statistical evidence of equivalence. This is easy to see if one designs a trial with only two patients per arm. Obviously, such a trial inevitably will show a lack of statistical significance, but it is doubtful that anyone would be convinced by this "lack of evidence" for a between-group difference.

Within the context of the previous discussion, consider the design of Schanler et al. It was intended as a three-arm study, although patients, of course, were only randomized to the DM and PF groups. How was sample size handled in their situation? They state in their paper that the primary analysis for the study was to compare the DM and PF groups. Yet paradoxically the sample size calculations were based on the comparison between either the DM or PF groups (or both) with MM. As noted in the preceding, a fundamental premise of any comparative clinical trial is to determine the sample size for the study on the basis of the primary study analysis. In this case, presumably that would mean a demonstration that the DM and PF groups were statistically equivalent, because the paper explicitly states that these products are expected to be equal and, ultimately, pooled together. On the other hand, if the goal of the study is to compare the three nutritional products and, presumably, demonstrate that MM is superior to both DM and PF, then the sample size calculation must be based on this premise. Schanler et al. also espouse this goal. As a result, there are two objectives for the trial, which are at cross purposes from a design perspective. It is fundamental to the sample size calculation and analysis that the groups to be compared be

treated equally in these regards. In other words, as noted, a basic approach to analyzing multigroup studies is to do a global analysis (i.e., evaluate all of the arms simultaneously) and then subanalyze any significant global finding using a multiple comparison approach that controls for the overall significance level of the testing. This procedure is basically espoused in any fundamental statistical textbook.[5] Thus, with respect to sample size calculation, the first step is to determine the effect size for the three group differences. Schanler et al. suggested that for their primary outcome variable (incidence of LOS and/or NEC) the MM group would have a rate of 30%, whereas the other two groups would have rates of 55% each. They perform their calculation with a two-group comparison between MM and either DM or PF (which is not in keeping with the spirit of the primary analysis of the study—the DM/PF comparison) using the mentioned rates and conclude that 70 subjects per group are needed, in spite of the fact that all three study arms were to be compared ultimately. (It is interesting to note that the rates found in their study were nowhere near these predictions. In the Shanler study, the MM group had a 29% rate, whereas the DM and PF groups had a 39% rate each. Given these figures [which yield a chi-square effect size of 0.098] and based on the proper analysis using the chi-square test for homogeneity, the study only had about 21% power to detect such a difference.)

It is very informative to determine the sample size needed in this or any trial in which the aim is to statistically demonstrate the equivalence of two therapies. As noted, in order to perform this calculation, one must first define what is meant by "equivalence." Clearly, this is not precisely meant to be the exact same frequency of the outcome for both groups, because that is unrealistic. Thus, some delta (in the previous terminology) is required. This quantity represents the largest difference between clinically indifferent treatment outcomes. For instance, in a comparison between DM and PF, one might be willing to clinically accept the equivalence of these if the primary outcome (say late-onset sepsis and/or NEC) rates were within ±10% of each other (and assuming that the true proportion for this outcome for both groups was on the order of 55%). With this definition and using 80% power with a two-tailed 5% significance level, then 424 patients per group would be required (calculation based on the PASS 2005 software program, Kaysville, UT). Even if the definition of equivalence were expanded to ±15%, the sample size needed would still be 189 per group. Of course, larger sample sizes would be needed for 90% power. For the 70 subjects per group used in the Schanler et al. paper, a definition of equivalence of an absolute difference of ±25% would have been required to have the requisite power to statistically demonstrate this. In terms of the primary outcome (late-onset sepsis and/or NEC), if a baseline rate of 55% is assumed, then equivalence would equate to rates of 30% to 80%, or as few as 21 cases and as many as 56, if the comparator group (say PF) had approximately 38 cases. There is little doubt that this definition goes far beyond a reasonable value for delta. (After all, the sample size calculation used by Schanler et al. is based on the goal of showing a 25% difference between MM and DM/PF.) Thus, the consequence of such results is to make it quite difficult to argue that DM and PF can be demonstrated legitimately as equivalent, and thus pooled.

PRIMARY ANALYSIS

These comments concerning the sample size calculations in these types of studies lead to the issue of how a three-group comparison should be dealt with statistically. As noted and suggested routinely as the method for a primary analysis under these circumstances, all three groups should be compared simultaneously. With the rate or proportion data, this analysis is basically the familiar chi-square test (for homogeneity of rates) that is addressed in basic statistical textbooks. (A special adjustment for the *p*-value calculation should be used if there are particularly small rates, for example, a proportion of patients with more than one episode of LOS and/or NEC that are less than a few percent). With continuous data (e.g. the growth parameters), the one-way analysis of variance model is appropriate. In Schanler et al. the approach used was to pool the DM and PF groups

if they were not statistically significant and then compare this pooled group with MM. As recounted, the study was not designed and, in particular, powered for this comparison. As a result, a lack of statistical significance between these two groups was almost inevitable and reiterates the common statistical fallacy that a lack of significance implies equality, when in fact it merely suggests and should be correctly stated as a lack of statistical evidence for a difference—a totally different conclusion.

With these ideas in mind, it is informative to see the consequences of a more reasonable reanalysis of the Schanler et al. data. For instance, looking at the three-group comparison of the LOS and/or NEC primary endpoint, one finds an overall p-value of 0.30. For just the one-episode subjects $p = 0.61$ is obtained. (For the more than one–episode patients, $p = 0.03$, although these data should more properly be analyzed using a statistical [Poisson] model, which evaluates the number of episodes per subject and not just whether a subject had multiple episodes. The latter approach has the effect of equating a patient with two episodes with one who had, say five. This type of analysis was done in a study of intravenous immunoglobulin in this same patient population.)[6] It is also interesting to note with these data that there is no statistical difference between DM and MM ($p = 0.15$); therefore, there is no "justification" for simply pooling only DM and PF. In fact, if MM and DM are pooled and compared with PF, $p = 0.52$!

Similarly, with respect to LOS alone, the overall p-value is 0.13 with a p-value of 0.58 for just the single episode subjects and 0.14 (based on an exact calculation because of small cell sizes) for the multiple episode subjects. The same lack of significance is found for a comparison of the rates of blood isolates across all isolates, $p = 0.12$ (three-group comparison). There is also a conclusion in the paper concerning the lack of effect of DM with respect to hospital stay, yet the comparison of DM and MM in this regard shows no significant difference ($p = 0.11$).

Finally, with respect to the growth parameters, it is interesting to note, for instance, that for the length increment when the infant has achieved 150 mL/kg per day until the end of the study, the authors claim a significant difference between the pooled DM and PF groups versus MM ($p = 0.03$). Yet the DM and PF groups are significantly different from each other ($p = 0.007$), suggesting that both DM and MM are superior to PF and illustrating the dangers of pooling groups. Indeed, the only place in which a significant inter-group difference is found that concurs with the authors findings in this set of parameters was with the length increment for the entire study.

Essentially, from a reasonable analysis of the study data, one must conclude that there are virtually no differences among the three groups in total, leaving an overall negative result for the comparison of MM, DM, and PF. This may be attributed in part to the lack of sample size for the trial.

ANALYTICAL ISSUES ARISING FROM THE LACK OF COMPLETE RANDOMIZATION

Other methodological issues with these types of studies are worth pointing out. As noted, inevitably complete randomization is understandably impossible in studies that choose to have MM as the primary comparator. However, this tends to create a statistical problem that full randomization attempts to avoid, namely, the reasonable possibility that the study arms are not comparable with respect to important characteristics of the study subjects (i.e., covariates). First, one must *a priori* define these covariates, and such choices are based on clinical knowledge of which factors would be expected to correlate with the primary study outcome.[7] Next, these covariates are then incorporated into the statistical analysis (typically using regression models) as a way of "adjusting" for the group differences in the covariates in order to better conduct the primary comparison of the study groups with regard to the main clinical outcome. Ideally, this adjustment model should be used as the primary analysis, although very frequently it is not, and only the unadjusted comparison of the treatments is evaluated. It is certainly informative to evaluate model approaches and determine whether the addition of covariates to the statistical analysis does matter.

In Schanler et al., the choice of covariates for

their adjustment models (which were the secondary analyses of the data) were based on the finding of baseline significance for any variable observed. Although this is a common approach, there is a basic flaw to such a choice, particularly with regard to randomized study groups. In theory, randomization is supposed to eliminate any baseline differences among the groups. As a result, any statistical differences are essentially Type 1 errors. Therefore, if a variable is selected this way, it needs to have a sound clinical basis for selection, as noted. Taking this approach, it is curious that the investigators did not include any of the significant social characteristics (e.g., household income, education) as covariate adjusters. Could one not argue that such socioeconomic variables might have both positive and negative impact on the mother's nutrition and, in turn, the quality of the milk? Finally, it is also considered appropriate practice to include the stratification variables (gestational age and receipt of prenatal steroids in Schanler et al.'s study) as part of the adjustment model. They used the latter, but only because it remained significant at baseline in spite of the stratification.

THE INTENT-TO-TREAT PARADIGM

Last, another important concept in the analysis of clinical data is that of the intent-to-treat (ITT) paradigm. This has been defined in various ways in the clinical and statistical literature. Simply put, the ITT paradigm that has been universally adopted, particularly by the regulatory agencies, demands that all data collected be evaluated.[8] The simple notion behind ITT is that the primary analysis should reflect clinical practice "warts and all." Thus, an ITT analysis provides the most conservative comparison of the study data, making statistical significance with this approach that much more meaningful. The definition of intent-to-treat (ITT) used in the Schanler et al. study is worth noting. Cases of LOS and/or NEC that occurred before a milk intake of 50 mL/kg was achieved were dropped. Although the authors' argument that they wanted these outcomes to relate to milk exposure is understandable, this is not in keeping with the ITT concept.

CONCLUSION

In this brief exposition of the design and analytical issues associated with studies of human milk and preterm formula, a number of methodological and analytical issues have been noted. In particular, the difficulties in using a standard, completely randomized approach create unusual statistical problems that must be handled in a thoughtful manner. This paper has illustrated these issues through the evaluation of the study recently reported by Schanler et al. Although this was an earnest attempt to provide clinical answers to nutritional questions in premature infants, the statistical problems with this study bring its conclusions into question. As noted, it appears that this trial was not able to demonstrate any important clinical differences among MM, DM, and PF, particularly with regard to their primary outcome of infection-related events. Although theoretically and practically MM is the nutrition of first choice for premature infants, this study could not adequately show that. Furthermore, there is no real evidence that DM is substantially worse, and necessarily equivalent to PF. With regard to this latter point, a trial directly comparing DM and PF needs to be undertaken with a reasonable sample size. (A study merely looking at NEC rates would require at least 900 subjects to be adequately powered.) Furthermore, from a practical perspective, the formulation of DM must be done under carefully controlled conditions, particularly requiring consistency in the caloric content and constituents of the milk.

REFERENCES

1. Schanler RJ, Lau C, Hurst NM, O'Brian Smith E. Randomized trial of donor human milk versus preterm formula as substitutes for mother's own milk in the feeding of extremely premature infants. *Pediatrics* 2005;116: 400–406.
2. Narayanan I, Prakash K, Bala S, et al. Partial supplementation with expressed breast-milk for prevention of infection in low-birth-weight infants. *Lancet* 1980;ii: 561–563.
3. Lucas A, Cole TJ. Breast milk and neonatal necrotizing enterocolitis. *Lancet* 1990;336:1519–1523.
4. Rosner B. *Fundamentals of Biostatistics*, 6th ed., Chapter 8. Thomson Higher Education, Belmont, CA, 2006.

5. Rosner B. *Fundamentals of Biostatistics*, 6th ed., Chapter 12. Thomson Higher Education, Belmont, CA, 2006:573.
6. Baker CJ, Melish MB, Hall RT, et al. Intravenous immune globulin for the prevention of nosocomial infection in low birth weight neonates. *NEJM* 1992;327: 213–219.
7. Piantadosi S. *Clinical Trials: A Methodological Perspective*. John Wiley & Sons, New York, 1997.
8. Chow S-C, Liu J-P. *Design and Analysis of Clinical Trials: Concepts and Methodologies*, 2nd ed. John Wiley & Sons, Hoboken, NJ, 2004.

Address reprint requests to:
Martin L. Lee, Ph.D., C.Stat.
Department of Biostatistics
UCLA School of Public Health and
Prolacta Bioscience
605 E. Huntington Drive
Monrovia, CA 91016

E-mail: martin.l.lee@att.net