Analytic Issues Related to the Evaluation of Normal Physical Growth As an Indicator of Nutritional Adequacy of New Infant Formulas

A White Paper

Prepared for the Food Advisory Committee on Infant Formula
Food and Drug Administration

by

Edward A. Frongillo, Ph.D.
Division of Nutritional Sciences
Cornell University
Ithaca, New York

November 18, 2002

The purpose of this background document is to discuss five issues related to the questions to be considered by the Food and Drug Administration Food Advisory Subcommittee about the evaluation of normal physical growth as an indicator of nutritional adequacy of new infant formulas. The five issues are:

1. Sensitivity and usefulness of several types of comparisons for comparing test cohorts in clinical growth studies of new infant formulas,
2. Potential for evaluating a meaningful difference in growth increments per day from clinical growth studies of new infant formulas,
3. Impact of transformations of raw data measurements into normalized indices,
4. Advantages and disadvantages of comparing group means and standard deviations and of comparing growth measurements for individual children with various reference data sets,
5. Circumstances where one type of comparison might be preferable to another and the possible impact on study conclusions.

Each of these five issues is discussed in turn, and then conclusions are presented.

Issue 1: Comparing test cohorts in clinical growth studies

Determination of physical growth is an important part of evaluating a new infant formula. The Academy of Pediatrics (1988) suggested that weight gain over the first four months should be examined, with measurements taken on a cohort of infants receiving formula at 14, 60, and 120 days. It was also suggested that rates of weight gain (g/day) should be calculated for the 14-to-60-day, 60-to-120-day, and 14-to-120-day periods.

To evaluate the adequacy of growth on the new infant formula, it is necessary to make a comparison to another group of infants not being fed the new infant formula and

whose growth is deemed to be adequate. The comparison group might be infants: a) randomized to receive a standard, established infant formula, b) whose growth is represented in a reference, c) whose measurements are in a currently available data set, and d) whose measurements are in historical data sets. Each of these choices has advantages and disadvantages.

The advantages of using a randomized comparison group fed a standard, established formula are considerable. Randomized clinical studies are favored because of the theoretical close control of measured and unmeasured factors that might influence the outcome (e.g., weight gain). Design features can be incorporated into randomized studies to minimize known sources of potential bias. For example, infants might be stratified on the basis of gestational age and then randomization done within the strata, thereby ensuring that gestational age cannot be a confounding factor. The probability statement from the statistical test that summarizes the results is justified on the randomized design as well as the statistical model used. The disadvantage of the randomized design is that the sample size is twice as large as the other approaches, because growth must also be measured on the concurrent comparison group.

The advantages of using a comparison to a reference are that the adequacy of growth is compared to a known, established reference sample and only one group of infants is measured. There are four principal disadvantages of such a comparison. First, the cohort fed the new infant formula may differ in some important ways from the reference sample. For example, infants with higher socioeconomic status or lower birth weight might grow somewhat faster. Second, existing references may exhibit somewhat different growth patterns from those of the cohort fed the new (or even a standard,

established infant formula). It is well documented that infants exclusively breast-fed in the first several months grow differently than infants fed formula (WHO Working Group on Infant Growth, 1995; Dewey, 1998; Butte et al., 2000; Frongillo, 2001). For example, Hediger et al. (2000) found a difference of about 120 g between exclusively breast-fed and never breast-fed infants at 4-7 months in the nationally representative third National Health and Nutrition Examination Survey (NHANES III) of the U.S. conducted during 1988-1994. Other studies have found other differences, depending in part on the extent of exclusive breast-feeding in the sample (Dewey, 1998; Frongillo, 2001). Third, the current U.S. reference (Kuczmarski et al., 2000) was constructed as a cross-sectional reference. Consequently, this reference does not represent well the variability in growth increments over the first four months that should be the basis for the statistical test of adequacy of growth. Fourth, reference data are not free of sampling error. For example, the NHANES III data set that is the basis for the early months of the current U.S. growth reference for weight (Kuczmarski et al., 2000) has measured weights on about 100 infants of each sex at both three and four months of age, and about 34 of each sex at two months. Although the U.S. reference is somewhat more precise than implied by this sample size (because of additional data used at birth and smoothing techniques to construct the reference), the reference is not perfectly precise as is sometimes mistakenly assumed when doing one-sample statistical tests comparing the growth of a cohort to a reference.

Currently available data and historical data both refer to existing data that were collected earlier in time. The use of existing, earlier data would minimize data collection. But, the concern would be that the characteristics of the sample of infants on the new infant formula might differ in some important ways from those in the earlier samples.

Issue 2: Evaluating a meaningful difference in growth increments per day

When designing a study to compare growth increments of groups of infants, it is important to plan for a sample size that will give sufficient statistical power for the smallest meaningful difference. The smallest meaningful difference is not the difference expected or the difference that others have found. It is the smallest difference that would be substantively important, which begs the question as to whether substantive importance should be considered primarily on individual or population grounds. That is, should we be considering the differential growth of individual infants or groups of infants? Given that a new infant formula will likely be used by a large number of infants, the population perspective is the most salient.

It has previously been recommended that the smallest meaningful difference in growth increment is 3 g/day, or 318 g for the period of 14 to 120 days (Academy of Pediatrics, 1988). The basis for this recommendation is not established in that report. If two different infant formulas resulted in a difference of 318 g of weight at 4 months of age, that means that the entire distribution of weight would be shifted by that amount. To gauge whether 318 g is a small or large difference, it is helpful to reflect both on the distribution of infant growth as captured in a reference and on differences due to other factors that we know about and accept as being important. The difference of 3 g/day is about the difference between the 25^{th} and 50^{th} (and 50^{th} and 75^{th}) percentiles of the increments in the Iowa and Iowa/Fels data (Nelson et al., 1989; Guo et al., 1991). The difference of 318 g is nearly as big as the difference in birth weight between low and very high altitude (Haas et al., 1982) and is about 50% larger than the effect of smoking during

gestation on birth weight of about 200 g (Institute of Medicine, 1990). From these perspectives, a difference of 3 g/day is meaningful.

The next question, then, is whether differences smaller than 3 g/day are also meaningful. In the end, this is a matter of perspective and judgement, with little empirical research available for guidance. An investigation of the criteria for judging growth faltering in infants suggested that a difference of about 0.5 standard deviations was clinically meaningful (Frongillo et al., 1990; Frongillo and Habicht, 1997). For weight gain during the first four months, this would correspond to a difference of about 2.5 g/day (about one-half of 5.3 g/day, see next paragraph). The direct applicability of these results to the evaluation of differences in weight gain of new infant formulas is somewhat open to question because: 1) the criterion of 0.5 standard deviations referred to cross-sectional variability (i.e., at an age) and not variability in increments, 2) the investigation was of infant growth after six months, not before, and 3) what is important at the individual level may not be indicative of what is important at the population level. Nevertheless, it seems that the smallest meaningful difference is probably less than 3 g/day, perhaps 2.5 g/day, and possibly as low as 2.0 g/day or 212 g from 14 to 120 days. Another perspective on this can be obtained by examining results that have been found previously when comparing different formulas. Table 1 below presents the results on weight gain from birth to four months for about 263 infants (Roche et al., 1993).

	Breast-fed	*Good Start*	*Isomil*	*Similac*
Male infants	3220	3440	3170	3650
Female infants	3000	3010	3020	3120

The differences in weight gains among the three formulas for males of 210, 270, and 480 g seem important, but the largest difference for females of 120 does not.

The choice of the smallest meaningful difference has implications for the sample size needed. The Academy of Pediatrics (1988) estimated the sample size of 28 needed per group on the basis of a one-tailed test at $p<0.05$ with 80% power. The smallest meaningful difference used was 3 g/day and the standard deviation used was 4.5 g/day. This estimate of the standard deviation was attributed to a personal communication with S.E. Nelson, but Nelson et al. (1989) subsequently reported a standard deviation of 5.3 g/day (calculated by the current author by pooling the squares of the male and female standard deviations of 5.6 and 4.9). This standard deviation is consistent with that reported by Guo et al.. (1991) for weight gain over months 1 to 4. With this standard deviation, the sample size needed per group is 40, not 28. But, an argument can be made that the power of 80% is too low because it means that the test fails to reject the null hypothesis of no difference one-fifth of the time for which there truly is a difference of at least 3 g/day. Furthermore, we should be concerned about differences in weight gain in either direction. If the power is set at 90% and the test is two-tailed, then the sample size needed per group is 67. With these parameters, if the smallest meaningful difference is 2 (rather than 3) g/day, then the sample size needed per group is 149. Table 2 (appended at end of paper) summarizes these results, along with other options for the choices of parameters (assuming a standard deviation of 5.3 g/day).

Issue 3: Transformations of raw data measurements into normalized indices

Normalized indices are statistics generated by matching raw data measurements with reference values on sex and age. The most commonly used indices are z-scores. Z-

scores are obtained by taking the value of a growth measurement, subtracting the age- and sex-specific median taken from the reference, and dividing by the age- and sex-specific standard deviation taken from the reference. Z-scores for weight-for-age (and sex), height-for-age (and sex), and weight-for-height (also sex-specific) are commonly used with infants.

The primary purpose for using z-scores is descriptive. Z-scores allow a set of measurements from infants in a sample who may vary in age and sex to be combined together. The use of z-scores has greatly facilitated comparing the growth of groups of infants from one place to another and is also helpful for assessing the growth of individual infants (WHO, 1995).

The use of z-scores for analytic purposes, however, can be problematic, especially when the pattern of growth in the reference, both in terms of central tendency (i.e., median) and variability (i.e., standard deviation), differs from the sample at hand over time because of differential feeding mode and other factors (WHO Working Group on Infant Growth, 1995; Frongillo, 2001). Given that the current U.S. growth reference was constructed from a mixture of infants with regard to feeding mode, some differences in patterns will likely be present. Furthermore, as stated earlier, the medians and standard deviations in the U.S. reference (or any cross-sectional reference) are useful for assessing growth status at one time, but are less useful for assessing growth status over time.

In the context of comparing infant formulas, the main rationale for using normalized indices—that sex and age are adjusted—is not truly an advantage. Since infants in a typical study to compare the growth associated with different infant formulas would be measured at predefined, common ages, it is not necessary to adjust for variation

in ages when analyzing increments expressed in weight gain per day (i.e., weight gain over an interval divided by days in the interval) unless the ages of measurement are far from the nominal ages. Even if age adjustment is needed, this adjustment would be made either by the inclusion of covariates for age in the analysis or by interpolation and extrapolation using the time series of measurements. Furthermore, it is well established that the growth of male and female infants should be assessed separately because there is often a sex differential in the response of growth to different infant formulas (Academy of Pediatrics, 1988; Nelson et al., 1989; Roche et al., 1993; see Table 1 above). Since sex is an effect modifier, growth should be assessed separately for male and female infants and there is no need to adjust for sex as a covariate or through use of z-scores.

Issue 4: Comparing with various reference data sets

A growth reference is a tool providing a common basis for purposes of comparison. The reference population should reflect the growth expected for children. About ten years ago, both National Center for Health Statistics/Centers for Disease Control and Prevention (NCHS/CDC) and the World Health Organization (WHO) independently but cooperatively began processes to extensively review the uses and interpretations of growth information. Included in these reviews was consideration of the limitations of the then-current references. Although these reviews differed somewhat in process and outcomes, both reviews reached a consensus that there were a number of important limitations of the 1978 NCHS/CDC and 1983 WHO references. As a result of the recommendations from these reviews, both NCHS/CDC and WHO set out to construct new reference growth charts.

In May 2000, NCHS/CDC released revised reference growth charts for the United States (Kuczmarski et al., 2000). These charts were created with improved data and statistical curve smoothing techniques. Data were taken from five national health examination surveys collected from 1963 to 1994 and five supplementary sources. These were combined into one analytic data set to produce the reference growth charts. These data better represent racial and ethnic diversity in the United States than the previous reference, and contain a mixture of growth data from infants who were breast- and formula-fed. The new reference growth charts were largely constructed using a descriptive approach, meaning that the reference portrays the growth that is expected of children in the population typically or on average, according to sex and age. Some aspects of a prescriptive approach were also taken, meaning that the reference portrays the growth that is expected of children who are healthy and well-nourished, and who have received proper care. Specifically, some data were excluded to avoid the influence of an increase in body weight that was observed in the more recent data.

The international infant growth reference currently used (i.e., the 1983 WHO growth reference) was constructed primarily from the growth of formula-fed infants. Motivated in part by evidence that the growth patterns of breast- and formula-fed infants differ, especially in the first year of life, WHO and its member states have strongly endorsed the construction of a new reference, and that it should reflect the growth expected of a population defined on the basis of having followed widely endorsed health and nutritional recommendations. To achieve this, WHO is conducting an intensive six-country study of children living in healthy environments to collect the necessary data. The new reference growth charts from this study will likely be completed in 3 years. An

important part of the rationale for constructing this new international growth reference is that it will improve the nutritional management of infants and lead to better support for breast-feeding and other accepted health practices.

Another possible reference for use in the evaluation of new infant formulas is the Iowa data (Nelson et al., 1989) or combined Iowa and Fels data (Guo et al., 1991). The Iowa data are from normal, term white infants born to faculty and students at the University of Iowa between 1965 and 1987 and who were breast- or formula-fed. The Fels data are from white, normal birth weight infants born to families with a wide range of socioeconomic status in Ohio between 1930 and 1987 and who were formula-fed. Whereas the current U.S. reference is useful for comparing the attained weight of infants to the U.S. population of infants, the Iowa and Fels data are useful for comparing the weight gain of infants because the longitudinal data have allowed the calculation of standard deviations and percentiles for weight gain. The measurement ages for the Iowa data best match the recommended timing of measurements for evaluating new formulas.

Issue 5: Circumstances favoring one type of comparison to another

There is one circumstance where it may be particularly advantageous (and therefore compelling) to use a comparison to currently available data. If an organization wished to test one new infant formula and also intended to test other new formulas near enough in time, then it would be efficient to sample infants from the same population (e.g., same geographic area, neighborhoods, or pediatric practices) for the whole series of studies, but without the need to repeat the sampling of a comparison group. As discussed earlier, the concern would be that the characteristics of the later samples might differ in some important ways from the earlier samples.

Summary

Studies to compare new infant formulas with existing formulas should use a design with a randomized, concurrent comparison group for the primary analysis. The sample size per group of 28 that has been recommended since 1988 does not have sufficient power to detect meaningful differences in growth of even 3 g/day, and a larger sample size is needed. Furthermore, the smallest meaningful difference in growth might be 2 g/day. If this difference is accepted, than a sample size substantially larger than 28 per group is needed. The recommendation for the smallest meaningful difference should be based on our best understanding of the biology and also on the regulatory, clinical, and public health decisions that are to be made. A comparison of attained weight for all groups at each measured age with the current U.S. reference, and rates of weight gain with the Iowa data, would also be useful for descriptive purposes.

Acknowledgement

I thank Anna Milman for assistance with the preparation of this paper.

References

Butte NF, Wong WW, Hopkinson JM, O-Brian Smith E, Ellis KJ. Infant feeding mode affects early growth and body composition. Pediatrics 106:1355-1366, 2000.

Dewey KG. Growth characteristics of breast-fed compared to formula-fed infants. Biol Neonate 74:94-105, 1998.

Frongillo EA, Rothe GE, Lambert JKJ. Determining growth faltering with a tracking score. American Journal of Human Biology 2:491-501, 1990.

Frongillo EA, Habicht JP. Investigating the weanling's dilemma: Lessons from Honduras. Nutrition Reviews 55:390-395, 1997.

Frongillo EA. Growth of the breast-fed child. In: Martorell R, Haschke F (eds) Nutrition and Growth. 47th Nestlé Nutrition Workshop. Philadelphia: Lippincott Williams & Wilkins, 2001, pp. 37-52.

Guo S, Roche AF, Fomon SJ, Nelson SE, Chumlea WC, Rogers RR, Baumgartner RN, Ziegler EE, Siervogel RM. Reference data on gains in weight and length during the first two years of life. J Pediatr 119:355-362, 1991.

Haas JD, Moreno-Black G, Frongillo EA, Pabon A J, Pareja L G, Ybarnegaray U J, Hurtado G L. Altitude and infant growth in Bolivia: A longitudinal study. American Journal of Physical Anthropology 59:251-262, 1982.

Hediger ML, Overpeck MD, Ruan WJ, Troendle JF. Early infant feeding and growth status of US-born infants and children aged 4-71 mo: Analyses from the third National Health and Nutrition Examination Survey, 1988-1994. Am J Clin Nutr 72:159-167, 2000.

Institute of Medicine. Nutrition During Pregnancy. Washington, D.C.: Subcommittee on Nutritional Status and Weight Gain During Pregnancy, Institute of Medicine, National Academy of Sciences, 1990.

Nelson SE, Rogers, RR, Ziegler EE, Fomon SJ. Gain in weight and length during early infancy. Early Human Development 19:223-239, 1989.

Kuczmarski RJ, Ogden CK, Gummer-Strawn LM, et al. CDC Growth Charts: United States. Advance data from vital and health statistics; no. 314. Hyattsville, MD: National Center for Health Statistics, 2000. See World Wide Web address: < http://www.cdc.gov/growthcharts/>.

Roche AF, Guo S, Siervogel RM, Khamis HJ, Chandra RK. Growth comparison of breast-fed and formula-fed infants. Canadian Journal of Public Health 84:132-135, 1993.

WHO. Measuring change in nutritional status. Guidelines for assessing the nutritional impact of supplementary feeding programmes for vulnerable groups. Geneva: World Health Organization, 1983.

WHO. Report of WHO Expert Committee. Physical Status: The Use and Interpretation of Anthropometry. WHO Technical Report Series No. 854. Geneva: World Health Organization 1995.

WHO Working Group on Infant Growth. An evaluation of infant growth. Geneva: World Health Organization, 1994.

WHO Working Group on Infant Growth. An evaluation of infant growth: The use and interpretation of anthropometry in infants. Bulletin of the World Health Organization 73:165-174, 1995.

Table 2. Sample size needed per group for assessing growth over the first four months assuming a standard deviation of 5.3 g/day (Nelson et al., 1989) and testing at $p<0.05$, depending on the smallest meaningful difference, power, and one- or two-tailed test.

Smallest Meaningful Difference	*Power (%)*	*Tails for test*	*Sample size per group*
1.5	80	1	156
1.5	80	2	197
1.5	90	1	245
1.5	90	2	264
2.0	80	1	88
2.0	80	2	112
2.0	90	1	121
2.0	90	2	149
2.5	80	1	57
2.5	80	2	72
2.5	90	1	78
2.5	90	2	96
3.0	80	1	28
3.0	80	2	50
3.0	90	1	40
3.0	90	2	67