Editorial

■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■

Sorting Out the Ways to Sort Out Fetal Growth

Joseph Schulman MD, MS

Thus it is easy to prove that the wearing of tall hats and the carrying of umbrellas enlarges the chest, prolongs life, and confers comparative immunity from disease; for the statistics shew that the classes which use these articles are bigger, healthier, and live longer than the class which never dreams of possessing such things.

— G. B. Shaw, Preface to ''The Doctor's Dilemma'' (quoted by Ellwood, p. 116[1])

We understand much of the world by classifying things. From this perspective, modern understanding of fetal growth might begin with the growth curves of Lubchenco et al.[2] The now decades-old epiphany that birth weight (BW) must be expressed in terms of gestational age (GA) is a special case of broader insights that include stratifying BW by gender and by race/ethnicity.[3,4] More generally, we often learn by observing how outcomes vary for two different groups differing in only one attribute. But our understanding can be distorted when the groups differ in more than one way and each of those differences may be associated with the chosen outcome.

Fetal growth data must be adjusted for known confounders and effect modifiers. Confounding distorts the association between an exposure (a defined gestational period, for example) and an outcome (BW, for example) through the presence of another factor also associated with the same exposure and outcome.[5] Thus, congenital viral infection or maternal hypertension confounds the association between a defined gestational period and BW (and wealth or social class confounds the association between wearing tall hats and living longer). In a prior issue of this journal, Madan et al.[4] stratify BW by race/ethnicity and gender to explore whether these variables behave as effect modifiers. Effect modifiers influence the value of the outcome variable (BW, for example) according to the alternative values of the effect modifier (male/female, for example).[5]

It is important to explore raw data for the existence of subgroups that differ among each other. Combining measurements from such subgroups (unstratified analysis) may obscure differences among them. Differences among subgroups may even cancel each other out

Albany Medical College, Albany, NY, USA; and Clinical Informatics and Outcomes Research, Children's Hospital at Albany Medical Center, Albany, NY, USA.

Address correspondence and reprint requests to Joseph Schulman, MD, MS, Albany Medical College, Children's Hospital at Albany Medical Center, Albany, NY 12208-3479, USA.

when all the measurements are aggregated. The large number of observations that can be collected in state, provincial, and national databases permit ''slicing and dicing'' the data to test whether aggregated summary statistics are appropriate or effect modifiers like racial/ethnic group require stratification instead. That Madan et al. aggregated groups like Blacks, Native Americans, and Hawaiians without reporting stratified results[4] appears to reflect not biologic considerations but numbers in those subgroups too small to stratify.

Other recent work aimed at improving how we classify fetal growth includes new, population-based growth curves,[6,7] customized cut-points[8,9] and a drill-down to subclassify infants categorized as small for gestational age (SGA) on the basis of genetic growth potential — 22% of infants conventionally classified as SGA were merely ''constitutionally small'' rather than having a ''fetal growth restriction (FGR).''[10] The methodological approaches to the fetal growth classification problem are so varied that an overview of some central issues in the design and analysis of these studies may help to consolidate current knowledge and clarify goals.

Collecting complete data for an entire population of individuals is difficult, so we may just sample from the population. When we draw two different samples from a source population we are likely to get two different sample means and standard deviations. But how do we know that the source population is not actually composed of two or more different subgroups? To discriminate sample summary statistics reflecting one or several source populations is the essence of the Madan et al. study.[4] The precision with which the true population mean can be estimated from a sample depends on the degree of homogeneity in the source population. Madan et al.[4] conclude that the differing means among their subgroups probably do not reflect random sampling error but several different source populations. Their study cannot answer whether the hospital sample of each source population corresponds to a random sample. Their study subjects may be unrepresentative of the exposure distribution of factors affecting BW in the source populations. A better test of the hypothesis that fetal growth depends on race/ethnicity would be population based.

Although studying an entire population theoretically removes sampling error, other challenges arise. Population-based infant BW studies provide data for a defined time period — for example, single live births in the US in 1991.[6] However, the population biologic experience may be different in another time interval than the one studied. Moreover, accurate assessment of GA has continually plagued studies of fetal growth.[7] After assessment, the value must be entered accurately into the record. To make inference from inaccurate and incomplete data requires statistical smoothing techniques that

modify the raw population data.[6,7] Electronic databases can prevent some of these problems by automatically checking for particular dimensions of data accuracy and consistency (referential integrity) through built-in field validation.

To date, population-based studies have been cross-sectional. These studies report the weights of different infants according to the GA at which they were born. Longitudinal data, in distinction, would describe the weight of individual infants (fetuses, actually) as they increase in GA. Strictly speaking, growth may only be assessed from longitudinal data.[7] Because factors associated with preterm birth may also be confounders for BW of preterm infants, cross-sectional population-based references may distort some associations.

Another classification strategy uses "customized" cut-points to test associations with specific outcomes.[8,9] This approach invites reflection on what measurement variable might be most informative of fetal growth. For example, "ponderal index" may be better than BW because it is a function of both weight and length.[11] Might some significant differences Madan et al.[4] reported in BW among racial groups disappear after accounting for the shorter length that accompanied lower BW? The point is that our conceptual imprecision does not allow equating BW *classification* with fetal growth *diagnosis*. To progress from classification to diagnosis we need to understand better the causal linkages between a chosen measurement variable and associated outcomes — prerequisite for establishing a "gold standard" to evaluate the discriminative power of a diagnostic test.

Classifying BW according to a particular cut-point, customized or not, converts continuous data to categorical data, with associated information loss. Madan et al. converted the data set used for Table 1 to produce a problematic Table 2.[4] Essentially, Table 2 interrogates the same sampling distribution problem of Table 1. More importantly, based on the analysis of Table 1, the statistical hypothesis testing of Table 2 seems altogether not warranted. To compute a *p* value for a particular difference among groups is to estimate the probability of obtaining a value as extreme or more extreme than that observed, *assuming the null hypothesis — no real difference* (Pagano and Gauvreau,[12] p. 212). But the analysis of Table 1 *rejected* the null hypothesis and concluded that the groups *were* different.

Categorization, appropriate or sometimes not, is widespread in healthcare. If the blood glucose concentration in a patient exceeds a particular value, the patient has diabetes mellitus. Above a particular serum cholesterol concentration a patient has hypercholesterolemia. If BW falls below a particular cut-point, an infant is SGA. The notion in each example is: either you've got "it" or you don't. The categorical model may deceive us into thinking we understand a condition more clearly than we actually do. Not all conditions are

best represented as categorical variables. Although a woman is either pregnant or she is not, many conditions may not be categorical variables so much as continuous variables (think: gradient of dose–response). Some of the confusion and conflict among reports on BW classification may stem from thinking categorically.

As first observed by Geoffrey Rose and more recently articulated by Marmot, instead of a model in which "either you've got it or you don't," maybe we should think about "how much of it you've got".[13] For evaluating fetal growth, this amounts to patient-specific risk profiling. Population-based profiling for an array of outcome risks, adjusted for confounders and effect modifiers, could discriminate the infant with FGR in need of medical intervention from the constitutionally small infant of identical BW in need only of parental love.

References

1. Ellwood JM. Critical Appraisal of Epidemiological Studies and Clinical Trials. Oxford: Oxford University Press; 1998.
2. Lubchenco LO, Hansman C, Dressler M, Boyd E. Intrauterine growth as estimated from liveborn birth-weight data at 24 to 42 weeks of gestation. Pediatrics 1963;32:793–800.
3. Thomas P, Peabody J, Turnier V, Clark RH. A new look at intrauterine growth and the impact of race, altitude, and gender. Pediatrics 2000;106:e21.
4. Madan A, Holland S, Humbert J, Benitz WE. Racial differences in birth weights of term infants in a Northern California population. J Perinatol 2002;22:230–5.
5. Hennekens CH, Buring JE. Epidemiology in Medicine. Boston: Little, Brown and Company; 1987.
6. Alexander GR, Himes JH, Kaufman RB, Mor J, Kogan M. A United States national reference for fetal growth. Obstet Gynecol 1996;87:163–8.
7. Kramer MS, Platt RW, Wen SW, et al. A new and improved population-based Canadian reference for birth weight and gestational age. Pediatrics 2001;108:e35.
8. Clausson B, Gardosi J, Francis A, Cnattingius S. Perinatal outcome in SGA births defined by customized versus population-based birthweight standards. Br J Obstet Gynaecol 2001;108:830–4.
9. De Jong CL, Francis A, Van Geijn HP, Gardosi J. Customized fetal weight limits for antenatal detection of fetal growth restriction. Ultrasound Obstet Gynecol 2000;15:36–40.
10. Mamelle N, Cochet V, Claris O. Definition of fetal growth restriction according to constitutional growth potential. Biol Neonate 2001;80:277–85.
11. Charlton V. Fetal growth: nutritional issues (perinatal and long-term consequences). In: Taeusch HW, Ballard RA, editors. Avery's Diseases of the Newborn. Philadelphia: WB Saunders; 1998. p. 45–55.
12. Pagano M, Gauvreau K. Principles of Biostatistics. Belmont, CA: Duxbury Press; 1993.
13. Marmot MG. Improvement of social environment to improve health. Lancet 1998;351:57–60.