Standard vs population reference curves in obstetrics: which one should we use?

Cande V. Ananth, PhD, MPH; Justin S. Brandt, MD; Anthony M. Vintzileos, MD

Abnormal fetal growth shapes disease risks later in the perinatal period, at infancy, and in childhood stages through chronic diseases and even death later in life.[1] Identification of fetal growth abnormalities, at both extremes of growth and particularly at the lower threshold of fetal size, continues to be debated vigorously.[2] There are at least 2 reasons that have fueled this ongoing debate. First, fetuses grow at different growth velocities, and divergence in growth is a phenomenon that is relatively well-observed at later gestations. It is therefore important that the heterogeneity in any given biometric parameter be taken into consideration when the percentile distributions are estimated. Fortunately, cutting-edge advances in the biostatistical modeling literature have paved ways to address this effectively.[3,4] Second, and arguably a critical issue, is the selection of the right growth curve on which fetal size is assessed. After all, the cohort composition of all growth charts is not the same, and the prevalence of abnormal fetal growth varies across different growth charts.[2,5,6]

In this issue of the *American Journal of Obstetrics and Gynecology*, Hoftiezer et al[7] develop a prescriptive (standard) birthweight chart based on approximately 1.6 million well-dated, singleton infants who were born in the Netherlands (2000−2014). The birthweight charts were constructed after the exclusion of preexisting maternal high-risk conditions and high-risk conditions that developed later in pregnancy. The charts were derived from infants born at 23−42 weeks gestation to "healthy" mothers after uncomplicated pregnancies and spontaneous onset of labor (ie, low-risk subjects). The authors concluded that their "standards" resemble the fetal weight "reference" charts and have greater ability to discriminate between normal and abnormal birthweights.[7]

From Division of Epidemiology and Biostatistics (Dr Ananth) and the Division of Maternal-Fetal Medicine (Dr Brandt), Department of Obstetrics, Gynecology, and Reproductive Sciences, Rutgers Robert Wood Johnson Medical School, New Brunswick, NJ; the Environmental and Occupational Health Sciences Institute, Rutgers Robert Wood Johnson Medical School, Piscataway, NJ (Dr Ananth); and the Department of Obstetrics and Gynecology, NYU Winthrop Hospital, NYU Long Island School of Medicine, Mineola, NY (Dr Vintzileos).

Received Feb. 18, 2019; revised Feb. 18, 2019; accepted Feb. 27, 2019.

The authors report no conflict of interest.

Corresponding author: Cande V. Ananth, PhD, MPH. ananthcv@rwjms.rutgers.edu

0002-9378/free

https://doi.org/10.1016/j.ajog.2019.02.060

Related article, page 383.

What is a "standard," and how does it differ from a "reference" curve? In our view, authors and readers poorly appreciate their distinctions. The goal of this Editorial is focused exclusively on clarification of the similarities and differences between standard and population reference curves and highlighting the strengths and weaknesses of each approach.

A note on terminology: standard vs (population) reference curves

A growth "standard" is constructed by the selection of only a part of the population, often those with no complications or with normal outcomes (ie, "low-risk" subjects). In contrast, a "reference" (population) curve is based on an unselected group of subjects and combines both low- and high-risk subjects and cases with normal and abnormal outcomes, thus being equated with a "population" curve. We also must define the commonly used terms "nomogram" and "normogram." The term *nomogram* is a mathematic term that was derived from the Greek words "νόμος," which means "law," and "γραμμή," which means "line," and describes how the data points can be used (https://www.merriam-webster.com/dictionary/nomogram-SEE), which has nothing to do with the nature of the studied population. The term *normogram*, which is derived from the word "normal," is a graph that depicts the distribution or range of normal values, regardless of the type of population studied. Therefore, both terms can be applied to both standard and reference population curves. The clinical implications of the use of standard vs reference curves depend on the specific clinical scenarios of interest. Thus, an understanding of the pros and cons for each approach is of paramount importance. A glossary of various terms related to the assessment of size and growth is described in the Table.

The pros and cons of the use of standard curves

We illustrate a few key considerations that guide the construction of a standard curve. First, given the exclusion of high-risk subjects, as was done in the study by Hoftiezer et al,[7] an important consideration is the potential for applicability of the nomogram to the general population. These standard curves are developed for a specific purpose, and the clinical impact of the application to pregnancies outside of their intended study population remains uncertain. Extreme caution should be exercised when such standard curves are applied clinically to birthweight references. This was the conclusion in a large Canadian study[8] that compared the INTERGROWTH 21st standards[9−11] to Canadian birthweight reference curves and the conclusion of others[12] who have cautioned regarding the premature adoption of the standards.

TABLE
Glossary of terms to define growth and size

Term	Definition
Reference	Statistical summary of the frequency distribution of fetal size of a reference population (descriptive); this descriptive depiction of fetal size across gestational age is based on an unselected population.
Standard	Nomogram of the frequency distribution of fetal size of an ideal population (prescriptive); this prescriptive depiction of fetal size across gestational age is based on a selected, low-risk population and reflects aspirational fetal size.
Nomogram	A graphic representation that consists of several lines marked off to scale and arranged in such a way that, by the use of a straightedge to connect known values on 2 lines, an unknown value can be read at the point of intersection with another line (Merriam Webster's definition); a nomogram can be applied to both standards and references.
Growth chart	Reference or standards of fetal size, not velocity; the data may be based on cross-sectional or longitudinal ascertainment of ultrasound or birthweight data.
Descriptive chart	Reference of fetal size that is based on a specific unselected population.
Prescriptive chart	Standard of fetal size that is based on a selected, low-risk population.
Birthweight chart	Reference of newborn infant size based on birthweights that are assumed to correlate with gestational age at delivery; because preterm infants are more likely to be pathologically small, birthweight-for-gestational-age charts generally underdiagnose small for gestational age at preterm gestations.
Ultrasound-based chart	Reference or standards of fetal size based on sonographic biometric parameters of fetal size that correlate with birthweight.
Individualized chart	Individualized assessment of fetal size based on early ultrasound assessments and projected trajectory of fetal growth.
Customized chart	Statistical summary of the frequency distribution of fetal size that incorporates maternal and fetal physiologic parameters (such as maternal race, parity, body mass index, and fetal sex).
Size	Quantitative assessment of fetal size or estimated weight at a specific gestational age (usually from cross-sectional assessments).
Velocity	Quantitative assessment of the change in fetal size over time, which reflects fetal growth (usually from a longitudinal study).
Growth	Quantitative assessment of velocity (rate of change) that is ascertained longitudinally.
Fetal growth restriction	Estimate of fetal size at a specific gestational age that is below a predefined threshold (usually the bottom percentile) based on a specific reference or nomogram; although this distinction is intended to identify fetuses who are at risk for adverse perinatal outcomes, some of these fetuses are not at risk for these complications; the prevalence of fetal growth restriction is dependent on the reference or nomogram.
Small for gestational age	Birthweight at a specific gestational age that is below a predefined threshold (usually the bottom percentile) based on a specific reference or nomogram; although this distinction is intended to identify neonates who are at risk for adverse neonatal outcomes, some small-for-gestational-age neonates are not at risk for these complications; the prevalence of small for gestational age is dependent on the reference or standard.
Pathologic small size	Distinction of small fetal size associated with adverse perinatal outcomes; optimally, only these fetuses would be identified as fetal growth restricted; however, in clinical practice, some may be characterized as normally grown and not exposed to antenatal surveillance.
Constitutional small size	Distinction of small fetal size that denotes normal fetal size; optimally, none of these fetuses should be identified as fetal growth restriction because they are not at risk for adverse perinatal outcomes; however, in clinical practice, some of these fetuses may be characterized as fetal growth restricted and exposed to potential, iatrogenic risks that are associated with false-positive antenatal surveillance and associated interventions.

Ananth. Standard vs population reference curves in obstetrics. Am J Obstet Gynecol 2019.

Second, should the curves that are generated be based on the exclusion of a particular set of high-risk conditions and not others? For instance, the standard curves by Hoftiezer et al[7] were based on the exclusion of maternal high-risk conditions only and not fetal or neonatal factors. The authors stated that the "final low-risk study population consisted of live-born singleton infants, born to 'healthy' mothers after uncomplicated pregnancies and spontaneous onset of labor." However, their study included preterm births that started at 23 weeks gestation. Are such newborn infants really low-risk or are births at preterm gestations an abnormal phenomenon? It would seem that any growth chart based on preterm

birthweights would have to be "descriptive" in nature; yet, Hoftiezer et al[7] have created a standard that appears to be more effective at the identification of growth abnormalities than the Dutch birthweight charts. Another approach, which is mainly relevant to fetal growth curves, could have excluded not only all mothers with complications or high-risk conditions but also all neonates who experienced complications at or after birth. But again, is the prospective clinical application of such curves appropriate when the final outcome is not yet known? In short, the inclusion criteria for a standard curve may reflect selection bias that limits the clinical applicability of the standards across specific populations.

The selection bias that remains inherent in the development of each standard also prevents appropriate comparisons between the standards. This point is emphasized by the recent confusion and debate regarding the adoption of the newly developed growth standards—the INTERGROWTH 21st project,[9-11] WHO Multicentre Growth Reference Study,[13] or the NICHD's Fetal Growth Studies[14]—into clinical practice. Because the scientific community has grappled with the uncertain clinical utility of these standards, researchers have started to compare them with established reference curves. As we have demonstrated earlier, the dramatically different assumptions and the inherent selection bias complicate these comparisons.

For example, INTERGROWTH 21st was applied to all singleton live births in Canada (excluding Quebec) from 2002−2012. With the use of the birthweight-for-gestational age, the frequency of small for gestational age (SGA) and associated neonatal morbidity/mortality rates was determined and compared with the Canadian birthweight reference. The study found important differences in the frequencies of SGA and neonatal morbidity and mortality rates that were associated with specific percentile categories. Although it is possible that the difference reflects real "biologic" differences, it is more likely that these differences are the consequence of varying cohort composition.

Third, the prospective clinical application of standard curves that are created based on exclusion of preexisting high-risk conditions may be problematic because they do not correct for complications that may occur later in gestation. In real life, a proportion of women who are recruited early in pregnancy inevitably will experience (pathologic) complications that are likely to affect fetal size and growth. These pathologies, depending on the gestational age at which they occur or are diagnosed, will have important implications in the identification of newborn infants that are SGA or large for gestational age for pathologic vs constitutional reasons.[15] Should these patients be later excluded because of the development of high-risk conditions such as preeclampsia or gestational diabetes mellitus? Thus, the practicability of standard curves should be examined cautiously. It is important to emphasize that, no matter which curves are used, we may never be able to capture all at-risk fetuses with a single assessment of fetal size and that longitudinal assessments of fetal growth that include individualized and customized growth charts may be reasonable alternatives.[16-18]

Fourth, and perhaps the most underappreciated, issue, pertains to the purpose for which a nomogram (standard or reference) is intended to identify. For a given fixed percentile cutoff, the use of standard curves will increase the sensitivity in the classification of fetuses as growth-restricted or newborn infants as SGA; inevitably, the specificity will be decreased. Thus, standard curves may be viewed as "screening" tests ("Clinical implications").

The pros and cons of using population reference curves

Population reference curves are developed for a defined population and without any subject exclusions. The advantage of the use of population reference curves is that these can be used prospectively to evaluate all patients, low- and high-risk, without the need to know the outcome. The disadvantage of the use of population curves is that these may not be comparable from population to population, given the differences in racial and ethnic composition and other biologic factors that affect growth. Thus, investigators are "forced" to construct and use fetal and birthweight reference curves from their own population.

Population curves based on birthweights have been used widely to identify fetuses with pathologic growth. As a consequence of including preterm deliveries, these references underdiagnose impaired fetal growth at preterm gestations when there are high rates of growth restriction. On the other hand, ultrasound curve of Hadlock et al,[19] which remains the most widely used fetal growth reference in the United States, was based on on-going pregnancies, thus including normal fetuses. This is the most likely reason for resembling the standard curves of Hoftiezer et al.[7]

In contrast to a standard curve, for a given fixed percentile cutoff, the use of population reference curves will increase specificity but may result in lower sensitivity in the detection of fetuses or newborn infants with abnormal growth. Thus, population reference curves may be viewed as "diagnostic" tests[20] ("Clinical implications").

Clinical implications

What are the clinical implications for the adoption of a standard vs a population reference curve? Although the answer may be seemingly trivial, an important, yet underappreciated, implication pertains to an understanding of the aims and scope of the generated curves. Charts from unselected (reference curves) vs selected (standard) populations each have a distinct role with respect to clinical implications. Application of a chart from a standard curve to assess "newborn infant size" invariably will render the newborn infant relatively "large" (ie, higher percentile distribution) in comparison with a reference curve. Although such distinctions may be obvious for well-grown newborn infants, the problem surfaces at the extremes of growth (percentiles at <5 or >95). It is in these percentile ranges where the application of an appropriate chart matters.

To put this in perspective, consider the evaluation of the effectiveness of a new test to identify subjects with a disease.[21]

If the test is used as a screening tool and applied to the general population of unselected subjects, then the intent of the test is to maximize specificity. In contrast, if the test is applied to a select (high-risk) population, then the test may be classified as a "diagnostic" test. The intent is to maximize the sensitivity.

An understanding of the distinctions between a screening vs a diagnostic test is akin to the development of reference curves in an unselected population (reference curves) vs a selected population of low-risk subjects (standards). If reference curves are developed in an unselected population, then the intent (albeit implicit) is more so to identify normal newborn infants that are not classified as being small or growth-restricted. This approach will yield increased specificity and lower sensitivity of the reference curve. In contrast, if the curve is developed in a selected population of low-risk women, then the intent is to identify most, if not all, fetuses or newborn infants who are small or growth-restricted. This approach will yield increased sensitivity in the identification of small or growth-restricted newborn infants. Of course, the sensitivity and specificity trade-offs can be modified by the choice of different percentile cut-offs for the biometry.

Conclusions

The philosophies that underlie fetal growth standard and reference curves are different, and an understanding of these differences for clinical or research purposes should guide the selection of the appropriate growth curves. For a fixed percentile cutoff, the use of standard curves as screening tests will result in the maximum sensitivity in the identification of fetuses or neonates with pathologic growth abnormalities, but at the expense of more false-positive results. The reverse is true for reference population curves. The appropriate percentile cutoff depends on the balance between sensitivity and false-positive rates, which can be dictated only by the clinical scenarios and the objective of the research. We caution that any comparison of growth standard with population reference curves will remain confounded by differences in cohort compositions, which may explain the differences in test characteristics and the underlying biologic differences. ■

REFERENCES

1. Crispi F, Miranda J, Gratacos E. Long-term cardiovascular consequences of fetal growth restriction: biology, clinical implications, and opportunities for prevention of adult disease. Am J Obstet Gynecol 2018;218(suppl):S869–79.
2. Grantz KL, Hediger ML, Liu D, Buck Louis GM. Fetal growth standards: the NICHD fetal growth study approach in context with INTERGROWTH-21st and the World Health Organization Multicentre Growth Reference Study. Am J Obstet Gynecol 2018;218(suppl):S641–55.e28.
3. Ohuma EO, Altman DG, for the International Fetal and Newborn Growth Consortium for the 21st Century (INTERGROWTH-21st Project). Design and other methodological considerations for the construction of human fetal and neonatal size and growth charts. Stat Med 2018. https://doi.org/10.1002/sim.8000 [Epub ahead of print].
4. Ohuma EO, Altman DG, for the International Fetal and Newborn Growth Consortium for the 21st Century (INTERGROWTH-21st Project). Statistical methodology for constructing gestational age-related charts using cross-sectional and longitudinal data: The INTERGROWTH-21st project as a case study. Stat Med 2018. https://doi.org/10.1002/sim.8018 [Epub ahead of print].
5. McCowan LM, Figueras F, Anderson NH. Evidence-based national guidelines for the management of suspected fetal growth restriction: comparison, consensus, and controversy. Am J Obstet Gynecol 2018;218(suppl):S855–68.
6. Sovio U, Smith GCS. The effect of customization and use of a fetal growth standard on the association between birthweight percentile and adverse perinatal outcome. Am J Obstet Gynecol 2018;218(suppl):S738–44.
7. Hoftiezer L, Hof MHP, Dijs-Elsinga J, Hogeveen M, Hukkelhoven CWPM, van Lingen RA. From population reference to national standard: new and improved birthweight charts. Am J Obstet Gynecol 2019;220:383.e1–17.
8. Liu S, Metcalfe A, Leon JA, et al. Evaluation of the INTERGROWTH-21st project newborn standard for use in Canada. PLoS One 2017;12:e0172910.
9. Villar J, Altman DG, Purwar M, et al. The objectives, design and implementation of the INTERGROWTH-21st Project. BJOG 2013;120(suppl 2):9–26, v.
10. Villar J, Cheikh Ismail L, Victora CG, et al. International standards for newborn weight, length, and head circumference by gestational age and sex: the Newborn Cross-Sectional Study of the INTERGROWTH-21st Project. Lancet 2014;384:857–68.
11. Papageorghiou AT, Kennedy SH, Salomon LJ, et al. The INTERGROWTH-21st fetal growth standards: toward the global integration of pregnancy and pediatric care. Am J Obstet Gynecol 2018;218(suppl):S630–40.
12. Zeitlin J, Vayssiere C, Ego A, Goffinet F. More validation is needed before widespread adoption of INTERGROWTH-21st fetal growth reference standards in France. Ultrasound Obstet Gynecol 2017;49:547–8.
13. WHO Multicentre Growth Reference Study Group. WHO child growth standards based on length/height, weight and age. Acta Paediatr Suppl 2006;450:76–85.
14. Grewal J, Grantz KL, Zhang C, et al. Cohort profile: NICHD fetal growth studies-singletons and twins. Int J Epidemiol 2018;47:25-25l.
15. Ananth CV, Vintzileos AM. Distinguishing pathological from constitutional small for gestational age births in population-based studies. Early Hum Dev 2009;85:653–8.
16. Francis A, Hugh O, Gardosi J. Customized vs INTERGROWTH-21st standards for the assessment of birthweight and stillbirth risk at term. Am J Obstet Gynecol 2018;218(suppl):S692–9.
17. Gardosi J. Counterpoint. Am J Obstet Gynecol 2019;220:74–82.
18. Gardosi J, Francis A, Turner S, Williams M. Customized growth charts: rationale, validation and clinical benefits. Am J Obstet Gynecol 2018;218(suppl):S609–18.
19. Hadlock FP, Harrist RB, Martinez-Poyer J. In utero analysis of fetal growth: a sonographic weight standard. Radiology 1991;181:129–33.
20. Burkhardt T, Schaffer L, Zimmermann R, Kurmanavicius J. Newborn weight charts underestimate the incidence of low birthweight in preterm infants. Am J Obstet Gynecol 2008;199:139.e1–6.
21. Ananth CV, Smulian JC, Vintzileos AM. Epidemiology of antepartum fetal testing. Curr Opin Obstet Gynecol 1997;9:101–6.