1-1-2011

Statistical analysis and interpretation of prenatal diagnostic imaging studies, part 3: Approach to study design

Katherine R. Goetzinger
Washington University School of Medicine in St. Louis

Methodius G. Tuuli
Washington University School of Medicine in St. Louis

Anthony O. Odibo
Washington University School of Medicine in St. Louis

Statistical Analysis and Interpretation of Prenatal Diagnostic Imaging Studies, Part 3

Approach to Study Design

Katherine R. Goetzinger, MD, Methodius G. Tuuli, MD, MPH, Anthony O. Odibo, MD, MSCE

Invited paper

A critical step in planning a successful study is choosing the appropriate design to feasibly answer the clinical question at hand. We provide an overview of common study designs, discuss their advantages and disadvantages, and provide practical examples from the prenatal diagnosis and ultrasound literature. In addition, we highlight specific design considerations that need to be built into the analysis of study results.

Key Words—cohort studies; descriptive studies; prenatal imaging; randomized trials; study design

A critical step in planning a successful study is choosing the appropriate design to feasibly answer the clinical question at hand. Although the research design depends on the question, there are advantages and limitations to each type of design. Furthermore, cost and time constraints as well as sample size considerations may influence the choice of design. Understanding the fundamental principles underlying the most common study designs encountered in the field of clinical research is essential not only for designing studies but also for the critical reading and appraisal of the literature. Recognizing both the strengths and the inherent biases introduced by each particular study design will allow the clinician to more accurately interpret results and, more importantly, determine whether and how these results should be applied to clinical practice. We provide an overview of common study designs, discuss their advantages and disadvantages, and provide practical examples from the prenatal diagnosis and ultrasound literature. In addition, we highlight specific design considerations that need to be built into the analysis of study results.

The Basics of Study Design

Figure 1 illustrates an algorithm for classifying clinical research designs.[1] Study designs can broadly be categorized as observational or experimental. In observational studies, data are gathered by observing events and outcomes over time (either prospectively or retrospectively) without any active manipulation of the exposure. Observational studies can further be divided into descriptive and

Received April 5, 2011, from the Department of Obstetrics and Gynecology, Washington University, St Louis, Missouri USA. Revision requested April 22, 2011. Revised manuscript accepted for publication April 28, 2011.

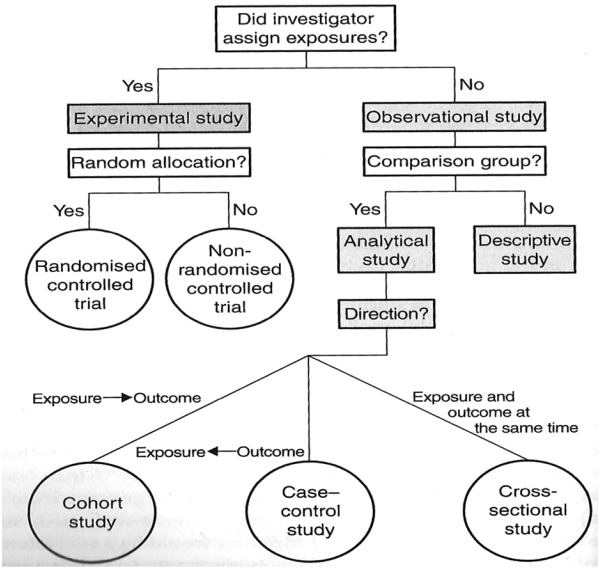
Address correspondence to Anthony O. Odibo, MD, MSCE, Department of Obstetrics and Gynecology, Washington University School of Medicine, 4911 Barnes-Jewish Hospital Plaza, Campus Box 8064, St Louis, MO 63110 USA.

E-mail: odiboa@wudosis.wustl.edu

analytic based on the presence or absence of a comparison group.[2] Inherent to the observational nature of the design, these studies are subject to the effects of bias and confounding. Both bias and confounding may threaten the internal validity of a study, which is a prerequisite for the external validity, or generalizability, of the study results. Although careful design and analysis can minimize their influences on the study results, their effects cannot completely be eliminated. For this reason, causal associations should only be made with caution in observational studies.

Bias is defined as "any trend in the collection, analysis, interpretation, publication, or review of data that can lead to conclusions that are systematically different from the truth."[3] Common types of bias include selection bias and ascertainment bias. Selection bias occurs when there is a systematic difference between the two groups being studied. For example, in a cohort study, because the exposure is not manipulated by the investigator, the exposure group may comprise patients with a particular set of characteristics that differ from those of the nonexposure group, thereby leading to spurious associations. Ascertainment bias occurs when methods for measuring either the exposure or outcome are different between the study groups.[2] For example, in a case-control study evaluating risk factors for aneuploidy, measurement bias may exist if all cases underwent amniocentesis for karyotype results, whereas controls may have only undergone amniocentesis in the presence of suspicious ultrasound findings. This type of bias could potentially misclassify undiagnosed cases as controls.

Figure 1. Algorithm for classifying types of clinical research design. Adapted from Schulz and Grimes.[1]



Finally, the effect of confounding must be considered. A confounder is defined as a third variable that is related to both the exposure and outcome but is not on the causal pathway toward the outcome. The presence of one or more confounders can blur associations found in the study, thereby limiting inferences of cause and effect. The process of randomization in a randomized controlled trial serves to eliminate the effect of both known and unknown confounders. Observational studies, on the other hand, often require more sophisticated methods to handle potential confounders. From a design perspective, methods such as specification and matching may reduce the effect of potential confounders. Specification restricts enrollment to patients with a particular value for a potential confounder, whereas matching selects cases and controls based on the presence or absence of the potential confounder. From a statistical perspective, confounding can be handled by methods including stratification, adjusted analyses, and propensity scores.[4] Despite these methods, residual confounding cannot be eliminated in observational studies.

In their most basic forms, experimental study designs involve an intervention that is applied to patients at the beginning of the study by the investigator, either by random or nonrandom allocation. The patients are then followed forward in time to determine whether they develop the outcome of interest. The double-blind randomized controlled trial is considered the reference standard study design in clinical research because it is the only effective means of avoiding bias and confounding.

The key features of the most commonly used research designs are discussed in the following sections.

Descriptive Study Designs

Descriptive studies are observational studies that lack a comparison group. As is inherent in the name, the goal of the descriptive study is to "describe" features of a condition and the characteristics of those affected. Grimes and Schulz[5] described a pentad of W questions that good descriptive research should be able to answer: Who? What? Why? When? and Where? By answering these questions, descriptive studies can be used to report new diseases or syndromes, evaluate characteristics of affected individuals, and follow trends in health. A major pitfall in the descriptive study design is that, because of the lack of a comparison group, causal associations cannot be evaluated. Despite this drawback, descriptive studies can be extremely useful in generating hypotheses that can then be tested in future analytic or experimental studies.[5,6]

There are 3 main types of descriptive studies encountered in the prenatal diagnosis literature: case reports, case series, and cross-sectional studies.

Case Reports and Case Series

A case report provides a detailed description of a clinical case, typically an unusual or uncommon disease, complication, or constellation of symptoms. Although a case report details findings from an individual patient, a case series describes findings from an aggregate of similar cases.[3] This type of study can be especially helpful when evaluating characteristics of a rare disease.

Cross-sectional Studies

The cross-sectional study design can be described as a "snapshot" view of a sample at one point in time. In this type of study, both exposure and outcome measurements are taken at the same time without a follow-up period. From these measurements, the prevalence of a disease or condition can be estimated.[4,5] The relative prevalence of an outcome can also be calculated in these studies by comparing patients with and without the exposure of interest. For this reason, a cross-sectional study is commonly referred to as a prevalence study. Because there is no longitudinal follow-up incorporated into this study design, the disease incidence cannot be calculated, and temporal relationships cannot be established. Despite these limitations, the associations and disease distributions observed in cross-sectional studies can be hypothesis-generating for future investigations.

An example of a cross-sectional study design in the ultrasound and prenatal diagnosis literature is in the construction of a fetal growth curve by Alexander et al.[7] They used population-based data from more than 3 million live births in the United States to create a national reference for fetal growth. By using large-scale data from the entire United States, Alexander et al[7] were able to generate a fetal growth curve that was anticipated to be more generalizable than prior growth curves, which had been constructed in various subpopulations and geographic regions. Intrinsic to the cross-sectional design, there was no longitudinal follow-up of estimated fetal weight over time, but instead, birth weight was captured at a single point in time for each patient delivering during the study period. Although growth curves derived from these cross-sectional studies cannot evaluate patterns of fetal growth or causal mechanisms for abnormal growth, they have certainly been used in the generation and testing of multiple hypotheses over the years. An important consideration in cross-sectional studies is that only one measurement per individual patient should be included in the data set to ensure that these measurements are independent. Prenatal diagnostic imaging studies commonly use cross-sectional study designs, and several examples of these have been published in the *Journal of Ultrasound in Medicine*.

Analytic Study Designs

The analytic study design differs from the descriptive study design in that it has a comparison or control group, which allows measures of association such as relative risks and odds ratios to be estimated. Analytic study designs are classically regarded as more powerful than their descriptive counterparts and have been rated as higher-quality clinical evidence by the US Preventive Services Task Force[8] in the hierarchy of study designs. Analytic study designs are divided into cohort and case-control studies based on classification of study groups by exposure or outcome, respectively.

Cohort Studies

The cohort study is a type of analytic study in which patients with an exposure are compared to those without the exposure with respect to one or more outcomes of interest. The defining feature of this type of design is that the patients are followed forward in time from exposure to outcome. From these data, the incidence of the outcome can be calculated because all patients are considered "disease free" at the start of the study. Although conceptually the tracking of data in this type of study design is always moving forward in time, the actual collection of data can occur prospectively or retrospectively. In the prospective cohort study, eligible patients are identified on the basis of the presence or absence of the exposure of interest and are followed for a meaningful length of time with periodic assessments of the outcome. In the retrospective cohort study, assembly of the cohort and outcome determination have already taken place. Existing data, such as medical records and perinatal databases, are frequently used for this type of study. Despite the retrospective data collection and analysis, the cohort itself is still defined by the presence or absence of exposure to the risk factor.[4,9]

Advantages of the cohort study include the ability to calculate the disease incidence and evaluate the temporal relationship between the exposure and outcome. Although causal associations can be suggested, such associations should only be made with caution because of the potential for confounding by other measured and unmeasured variables. Multiple outcomes can also be evaluated in cohort studies, although the statistical implications of multiple

comparisons must be considered. It is generally advisable to determine primary and secondary associations before initiating the study and to report all associations evaluated, both significant and nonsignificant. Performing multiple comparisons may result in significant associations by chance alone. Another advantage to the cohort study is that rare exposures can be studied by assembling a cohort of individuals with uncommon exposures. Although prospective studies allow a higher level of completeness in evaluating the variables of interest, retrospective studies are often of lower cost and much less time-consuming.[4]

Both prospective and retrospective cohort studies have their unique disadvantages. Both are inefficient designs for studying rare outcomes, and both are plagued with the potential for incomplete data due to loss to follow-up. It is important to consider the reason for the loss to follow-up because patients could be dropping out of the study due to a factor that is related in some way to the outcome of interest. In prospective cohort studies, the investigator has control over which variables are measured and how they are defined, whereas retrospective studies are dependent on how the data were previously collected. In many cases these data have not been measured and recorded in ways that are ideally suited to appropriately answer the study question. Important potential confounding variables may not be consistently recorded, thereby limiting the use of these variables in adjusted analyses. Finally, prospective studies may be time-consuming and costly, especially when there is a long lag time between the exposure and disease onset.[4,9] Fortunately, lag time is less of a concern in perinatal research, given the finite and relatively short length of pregnancy.

As with any study, it is important to have clear and measurable definitions of both the exposure and outcome. If the outcome is subjective, then the person assigning the outcome should ideally be blinded to the patient's exposure status. Additionally, all patients in the cohort should be at risk of developing the outcome.[9] For example, in a study of risk factors for postcesarean wound complications, patients who deliver vaginally should not be included because they have no risk of developing the outcome.

In addition to being able to calculate the incidence of an outcome in a cohort study, the strength of association can also be determined and expressed as the relative risk or odds ratio with the 95% confidence interval. Using a classic 2 × 2 table of exposure and outcome, the formula for relative risk is $[a/(a+b)]/[c/(c+d)]$, defined as the ratio of the outcome in exposed patients compared to unexposed patients (Table 1). A relative risk of greater than 1 indicates an increased risk of disease in the exposed group, whereas a relative risk of less than 1 indicates a possible protective effect of the risk factor on the outcome of interest. The further the relative risk is from 0 in either direction, the stronger the association. For example, a relative risk of 2 indicates that patients with a specific risk factor are 2-fold more likely to develop the outcome of interest compared to those without the risk factor.[2,10] When translating reported relative risks into clinical practice, it is important to remember the actual incidence of disease (absolute risk). If the incidence of the outcome is extremely low (eg, 0.01%), then even a large relative risk (eg, 10.0) may not translate into a clinically meaningful increased risk (eg, 0.1%). This interpretation is largely dependent on the severity and implications of the outcome of interest.

Finally, from cohort study data, one can determine the additional risk of the outcome after exposure that is attributable to the risk factor being studied. This measure is known as the attributable risk and is calculated by subtracting the incidence of the disease in the unexposed group from the incidence of disease in the exposed group. The attributable risk is often considered a better measure of individual risk than the relative risk, which is more of a representation of the strength of association.[10]

Cohort studies are very common in the prenatal diagnosis and ultrasound literature. Retrospective studies are often performed using ultrasound and perinatal databases. Such databases may include ultrasound data entered at the time of the examination as well as delivery and neonatal outcome data entered at the time of or soon after delivery. Although data collection is retrospective in nature, such studies are classified as cohort studies according to the definition of study groups based on exposure status. For example, using a perinatal database of 72,373 patients who presented for second-trimester sonography at a single institution, Hua at el[11] performed a retrospective cohort study to estimate the risk association between a single umbilical artery and adverse pregnancy outcomes. This study estimated a relative risk of 1.7 for intrauterine growth restriction, 2.6 for preterm birth before 34 weeks, and 1.8 for

Table 1. Measures of Association in Cohort and Case-Control Studies

	Outcome		
Exposure	**Yes**	**No**	**Total**
Yes	a	b	$a+b$
No	c	d	$c+d$
Total	$a+c$	$b+d$	$a+b+c+d$

Relative risk = risk of an outcome in exposed/risk of an outcome in unexposed = $(a/a+b)/(c/c+d)$. Odds ratio = odds of exposure in cases/odds of exposure in controls = $(a/c)/(b/d) = a \times d/b \times c$.

preterm birth before 37 weeks.[11] In this study, the exposure was defined as the presence or absence of a sonographically diagnosed single umbilical artery, and multiple adverse outcomes were studied. Although these results do not prove causation, they suggest that a single umbilical artery is a risk factor for these adverse perinatal outcomes.

An example of a prospective cohort study is a recently published study examining the association between the ultrasound appearance of a cesarean hysterotomy scar in nonpregnant women and the outcome of subsequent pregnancies.[12] Three groups of patients were defined according to the nature of their hysterotomy scars on transvaginal sonography (exposure): intact scar, small defect, and large defect. The primary outcome was uterine dehiscence or uterine rupture (outcome). The methods of this study show important principles in the design of prospective cohort studies, including the a priori clear definitions of exposure and outcome and the blinding of clinical staff to the ultrasound findings at the time of delivery to eliminate a biased assessment of the outcome. Finally, only patients with a prior cesarean delivery could be included in the study to fulfill the cohort study criteria that all patients must be at risk for the outcome of interest (uterine scar dehiscence).[12] Although small in numbers, this prospective cohort study suggests a positive association between large hysterotomy defects diagnosed by ultrasound and subsequent uterine dehiscence or rupture. This suggestion is hypothesis-generating and may serve as a basis for future larger-scale studies to definitively determine whether the status of prepregnancy hysterotomy scars on transvaginal sonography can accurately predict uterine dehiscence or rupture.

The cohort study design is particularly suited to many prenatal screening and diagnostic ultrasound studies.

Case-Control Study

In contrast to the cohort study design in which patients are selected according to the presence of the exposure, the case-control study involves identifying patients with (cases) and without (controls) the outcome of interest and then looking backward in time to determine exposure. This design is particularly advantageous when studying rare outcomes or outcomes with long lag times. Although multiple risk factors can be evaluated in a case-control study, analysis is limited to a single outcome. Given that patients are identified when the outcome has already occurred, temporal relationships cannot be established using this study design. Finally, because the number of cases is predetermined by virtue of the study design, the disease prevalence and incidence cannot be calculated.[4]

The odds ratio is the measure of association used in case-control studies. It is defined as the odds of exposure in patients with the disease compared to the odds of exposure in patients without the disease. Using the classic 2×2 table, the odds ratio is given by $a \times d/b \times c$ (Table 1). When the outcome is rare, the odds ratio approximates the relative risk. Alternatively, when the outcome is common, the odds ratio typically represents an exaggeration of the true relative risk. Although relative risks are typically easier to interpret than odds ratios, they cannot be reliably calculated in case-control studies. In other words, the number of cases and controls in a case-control study is fixed by the investigator (usually equal), whereas there are likely far fewer cases than controls in the overall population.[10,13] Even if the odds ratio indicates a strong association between an outcome and a risk factor, it does not necessarily translate into a strong discriminatory or predictive ability of that risk factor. Mathematically, the odds ratio can be written in terms of the true-positive fraction or sensitivity and the false-positive fraction or 1 – specificity as odds ratio = [true-positive fraction/(1 – true-positive fraction)] × [(1 – false-positive fraction)/false-positive fraction]. From this equation, it can be deduced that more than one combination of true- and false-positive fractions can produce the same odds ratio, showing the inconsistency in discrimination. Pepe et al[14] showed that an odds ratio of tremendous magnitude (eg, 171.0) is required to achieve adequate discriminatory ability between true- and false-positives. Therefore, for case-control studies in which prediction is the primary goal, it is prudent to report false- and -negative rates separately for binary risk factors. Alternatively, receiver operating characteristic curves may be more appropriate for continuous risk factors.[14] Ultimately, the principle goal of the study should drive the statistical analysis and reporting of the results.

One of the major limitations of the case-control study is its inherent susceptibility to selection and ascertainment biases. It is important to ensure that both cases and controls are representative of the population of interest and that patients designated controls are representative of the population at risk for developing the outcome of interest. It is particularly important to carefully consider the selection of controls. If controls did not undergo a diagnostic test or procedure to rule out the outcome of interest, it is possible that undiagnosed cases may actually be misclassified as controls.[4,13] The most common form of ascertainment bias in the case-control study is recall bias. Cases may be more likely to recall an exposure compared to controls. The classic example of recall bias in obstetrics is in teratogen research, in which mothers of neonates born with birth

defects are much more likely to have carefully considered every possible exposure during pregnancy compared to mothers with healthy neonates. This situation potentially leads to false conclusions of positive associations between exposures and outcomes.

The nested case-control study design is a modification of the case-control design that may reduce bias. In this design, cases and controls are drawn from a defined cohort of patients. By definition, all patients in the cohort are disease free at entry into the study. Those who go on to develop the outcome of interest become the cases, and a random sample of the remaining patients who do not develop the outcome of interest become the controls. To reduce confounding, controls are often matched to cases on the basis of the presence or absence of one or more variables. This unique study design eliminates the potential selection bias of controls coming from a population that is different from that of cases. This study design is also useful when measurement of variables may be costly or time-consuming. Rather than performing this measurement on all patients in the cohort, archived samples or images can later be analyzed only in patients selected as cases and controls.[4,15]

A recently published study on placental morphometric characteristics in pregnancies affected by intrauterine growth restriction and preeclampsia[16] is an example of this type of study design. In this study, placental morphometric assessment was performed on 13 patients with preeclampsia, 7 patients with gestational hypertension, 7 patients with intrauterine growth restriction, and 20 uncomplicated control pregnancies. Both cases and controls were selected from a cohort of patients who were enrolled in a prospective study evaluating the association between first-trimester measures of placental dysfunction and adverse pregnancy outcomes.[16] Given the time- and labor-intensive nature of placental morphometric assessment, it would not be feasible to perform this test on all patients in the cohort. Instead, the nested case-control study design provided a large cohort of representative patients from which both cases and a subset of controls could be identified.

Experimental Studies

Unlike observational studies in which the investigator has no control of the exposure, experimental studies involve assignment of the exposure by the investigator. The ability to assign exposure provides the investigator control over experimental studies that cannot be achieved in observational studies. However, interventional studies may not be feasible for all research questions. For example, it may not be ethical for an investigator to expose patients to a factor likely to cause a deleterious outcome. Similarly, outcomes with long lag times and the high costs associated with such studies may make the experimental design unsuitable for a particular question.

Interventional studies involving human patients are termed clinical trials. Depending on whether patients are randomly or nonrandomly assigned to the comparison groups, clinical trials may be randomized or nonrandomized.

Randomized Controlled Trials

The randomized controlled trial is considered the reference standard for clinical research. It is the only known effective means of avoiding selection bias and confounding. However, it must be noted that randomized trials have a number of drawbacks. First, whereas internal validity (measurement of what it sets out to measure) is likely if a randomized controlled trial is well designed and conducted, external validity (ability to generalize results to the broader population) is not guaranteed. The strict inclusion criteria in many trials result in groups that may not be representative of the broader population. Furthermore, trial participants who are often volunteers tend to differ from nonparticipants.[17] Second, ethical considerations make the randomized controlled trial design inappropriate in many instances. Finally, randomized controlled trials can be extremely expensive, limiting their use.

A number of key elements characterize randomized controlled trials: randomization, allocation concealment, blinding, analysis and reporting, and validity.

Randomization is the hallmark of randomized controlled trials. It is the method of assigning patients to groups in such a way that the characteristics of the patients do not affect the group to which they are assigned. To achieve this end, the investigator allows chance to decide to which group each patient is assigned. Randomization ensures that differences in outcomes between comparison groups are attributable to the intervention alone and not to known or unknown confounding characteristics. Although randomization does not guarantee that the groups will be identical in all baseline characteristics, it does ensure that any differences between them are due to chance alone. Randomization also facilitates concealment of an intervention from patients and investigators to further reduce bias. Finally, randomization leads to treatment groups that are random samples of the study population, permitting the use of standard statistical tests, which are based on probability theory.

The most basic means of randomization is "toss of the coin." More commonly used methods include random tables and computer-generated randomization sequences. Blocking, stratification, and minimization are modifications to simple randomization. Blocking is used to ensure that numbers of patients in each group are similar at all times. In block randomization, a chosen number of patients (blocks) are randomized at a time. Stratification is aimed at ensuring that groups of patients receiving different interventions are similar with respect to important prognostic factors. Because simple randomization does not guarantee that groups would be similar with regard to all factors (especially in a small trial), stratified randomization may be used. This process involves the creation of separate randomization for each subgroup (stratum) based on an important prognostic factor. For example, in a trial of treatments for reducing preterm birth, the investigator may want to take multiple gestations into account. Randomization may therefore be performed separately for singleton and multiple pregnancies. Minimization is used in small studies with several important prognostic variables, in which stratified allocation may not be feasible for all of the important prognostic variables. It involves assigning the next patient entering the trial in a way that minimizes the overall imbalance between the groups at that stage of the trial based on specified prognostic variables.

Blinding attempts to avoid bias by keeping patients, investigators, or both unaware of the group to which the patients are assigned. A patient's response may be affected by knowing to which group the patient is assigned, either through a belief that a particular treatment is or is not beneficial or through a desire to please the investigator. Furthermore, an investigator's evaluation may be affected by knowing to which group a patient is assigned. Thus, it is best if neither the patient nor the investigator knows to which group the patient is assigned (double blinding). Sometimes it is only possible for one party to be unaware of the assignment, in which case the trial is single blind. Most often, only the patient is blinded because it is often impossible for the investigator to be unaware of the treatment. In some cases, blinding is not possible for either patients or investigators, and the study is said to be unblinded. Even when blinding is not possible, assessment of outcomes may still be done in a blinded fashion. For example, in the Management of Myelomeningocele Study,[18] in which outcomes in fetuses with myelomeningocele undergoing prenatal or postnatal surgery were compared, the patients and investigators could not be blinded. However, assessment of the level of the lesion on magnetic resonance imaging and the mental and motor function of the patients was determined in a blinded fashion by individuals who were unaware of the group allocation of the patients.[18]

Allocation concealment has a related but conceptually different goal from blinding. It refers to techniques used to implement a randomization sequence, in which the interventions are made indistinguishable to the patients and investigators. This situation may be achieved through the use of sequentially numbered opaque sealed envelopes, pharmacy control, numbered or coded containers, or central randomization. Allocation concealment has an impact on trial results. For example, trials with inadequate allocation concealment have been shown to yield up to 40% larger estimates of effect size and are associated with greater heterogeneity in results.[19]

In terms of analysis and reporting, it is crucial that randomized controlled trials are not only well designed and conducted but analyzed and reported in sufficient detail to enable readers to assess the quality of their conduct and validity of the study results. The Consolidated Standards of Reporting Trials is a concerted attempt to improve the quality of reporting of randomized trials.[20] Readers are referred to the most recent update of the Consolidated Standards of Reporting Trials statement for details.[21] We highlight 4 elements: intention-to-treat analysis, analysis of baseline characteristics, adjusted analysis, and subgroup analysis.

Intention-to-treat analysis involves analysis of patients according to the group in which they were originally assigned regardless of whether they received or adhered to the intervention. This approach is in contrast to "per-protocol analysis" or "as-treated analysis," in which patients are analyzed by the intervention actually received. The primary results of a randomized control trial always should be analyzed according to the intention-to-treat principle. This process avoids bias that may result from noncompliance with the intervention due to unpleasant side effects or lack of a benefit. In addition, whereas intention-to-treat analysis often results in a more conservative estimate of effect size, it allows a better estimate of the effectiveness and implications of a practice change rather than pure efficacy alone.[22]

Baseline characteristics are collected at enrollment in trials. They are used to determine whether the randomization procedure successfully results in comparable groups, to adjust treatment effects for variables strongly related to the outcome, to perform subgroup analysis, and to assess the generalizability of trial results. Investigators often use formal statistical tests to compare baseline characteristics between intervention groups. Although a detailed description of trial participants is important, significance tests for baseline differences are considered inappropriate, given that the hypothesis that the groups come from different populations is known to be false.[23] Because the groups are obtained by randomization, any differences are due to

chance or flawed randomization. The multiplicity of comparisons in baseline characteristics also inflates the type I error; thus, significant results may occur by chance alone. Finally, it is argued that nonsignificant differences in strong predictors will have a more significant effect on the effect size than significant differences in factors unrelated to the outcome.[23] For these reasons, it is recommended that formal statistical comparison of baseline characteristics be limited to a few variables known to be strong predictors of the primary outcome.[10]

Adjustment of the effect size for baseline variables is controversial. Although an adjusted analysis may produce more precise estimates, results are more difficult to interpret than a simple unadjusted analysis. Adjustment for covariates in clinical trials is recommended only when there is clear a priori evidence that some baseline factors are strong predictors of the outcome, when the outcome is numeric and its baseline value is measured, and when the trial is small and imbalances are sufficiently large to bias the effect size.[10]

Subgroup analysis in clinical trials is common but often flawed, with results commonly overinterpreted.[23] Because trials are often not designed to test subgroup effects, it is recommend that subgroup analysis only be conducted if there is a clear a priori hypothesis and biological plausibility that intervention effects differ between different subgroups. It is also important that subgroup analysis include formal tests of interaction rather than assuming that a significant effect in one subgroup and not the other is conclusive evidence that the treatment effect differs between the subgroups.[10] Finally, results of subgroup analysis should be interpreted with caution, unless they are supported by strong evidence of biological plausibility.

A classic example of a randomized clinical trial in prenatal imaging is the Routine Antenatal Diagnostic Imaging With Ultrasound trial.[24] This clinical trial of 15,530 women was designed to test the hypothesis that screening sonography in low-risk pregnancies would improve the perinatal outcome. Pregnant women without any specific indication were randomized to receive either 2 screening prenatal ultrasound examinations or routine obstetric care. The trial found no clinically significant benefit from routine screening ultrasound examinations.[24] The trial still remains the largest devoted to the study question but also illustrates some of the potential limitations of randomized trials discussed above.

Variations of Randomized Controlled Trials

There are variations of simple parallel-group randomized controlled clinical trials in which the groups of patients are studied concurrently.

The crossover randomized trial is one in which patients serve as their controls. Each patient receives interventions in sequence, and the order of interventions is decided at random for each patient. A major advantage of the crossover design is that comparisons are more efficiently made within patients rather than between patients, which is particularly useful for outcomes that are highly variable between patients and when the effect of the intervention can be assessed quickly. The analysis of results must take into account this particular design by using statistical methods for paired data. Crossover trials have limitations. For example, they cannot be used for conditions that can be cured by the intervention. There is also concern of a "carryover effect," in which there is a residual influence of the intervention from one period to the other. This situation may require a "wash-out" period between interventions.

Cluster randomized trials involve allocating entire groups (clusters) rather than individual patients to interventions. The clusters must be taken into account in study design and data analysis. For example, cluster trials require larger sample sizes because the use of clusters is associated with reduced statistical power compared to the traditional parallel group randomized clinical trial design. Analysis must also include a cluster coefficient, which takes into account correlations between patients within the clusters.

The factorial clinical trial design is one in which several exposure factors are compared at the same time. In this design, each patient receives a combination of exposures such that all combinations are received by some patients. Factorial designs are particularly suited to the investigation of factor interactions, and they are often powered to detect interactions rather than main factor effects. Factorial clinical trial designs are sometimes proposed to optimize statistical power when the number of patients available for a study and resources are limited. However, the use of the factorial design for this purpose is based on the assumption of no interactions between the factors that may not be true. In addition, main factor effects from factorial trials represent averages of all combinations of the other factors, which may not be meaningful. These variations of randomized clinical trials have rarely been used in prenatal imaging studies but are included here to be complete.

Nonrandomized Trials

In some cases, practical, ethical, and other considerations may preclude the use of a randomized controlled trial. In such cases, a number of alternatives are available to evaluate interventions. Nonrandomized trials may be performed, in which patients are assigned to groups based on

reasons other than chance. Inference regarding the impact of the intervention in such designs is difficult to isolate from the effects of other factors. Analysis of data from such trials will need to be adjusted for potential confounders. Pre-post comparisons involve comparing rates of outcomes in a group before (preintervention) and after (postintervention) implementation of an intervention. This design technique is weak for showing the effectiveness of an intervention because risk factors and outcomes may change over time independent of the intervention.

Conclusions

Several research designs are available as tools for prenatal diagnostic imaging research. The choice of research design should be made carefully on the basis of the research question, the exposure and outcome, ethics, and the availability of resources. It must be noted that each design has unique advantages and disadvantages as well as critical elements, which must be carefully considered. It is important that a plan for statistical analysis be considered at the design stage of a study because flaws in design and data collection cannot be remedied after the fact with statistical analysis.

References

1. Schulz KF, Grimes DA. *The Lancet Handbook of Essential Concepts in Clinical Research*. Philadelphia, PA: Elsevier; 2006.
2. Fletcher RH, Fletcher SW. *Clinical Epidemiology*. 4th ed. Baltimore, MD: Lippincott Williams & Wilkins; 2005.
3. Porta M (ed). *A Dictionary of Epidemiology*. 5th ed. New York, NY: Oxford University Press; 2008.
4. Hulley SB, Cummings SR, Browner WS, Grady DG, Newman TB (eds). *Designing Clinical Research*. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2007.
5. Grimes DA, Schulz KF. Descriptive studies: what they can and cannot do. *Lancet* 2002; 359:145–149.
6. Grimes DA, Schulz KF. Clinical research in obstetrics and gynecology: a Baedeker for busy clinicians. *Obstet Gynecol Surv* 2002; 57(suppl 3):S35–S53.
7. Alexander GR, Himes JH, Kaufman RB, Mor J, Kogan M. A United States national reference for fetal growth. *Obstet Gynecol* 1996; 87:163–168.
8. US Preventive Services Task Force. *Guide to Clinical Preventive Services*. 2nd ed. Baltimore, MD: Williams & Wilkins, 1996.
9. Grimes DA, Schulz KF. Cohort studies: marching towards outcomes. *Lancet* 2002; 359:341–345.
10. Kirkwood B, Sterne J. *Essential Medical Statistics*. 2nd ed. Oxford, England: Blackwell Publishing; 2006.
11. Hua M, Odibo AO, Macones GA, Roehl KA, Crane JP, Cahill AG. Single umbilical artery and its associated findings. *Obstet Gynecol* 2010; 115:930–934.
12. Vikhareva Osser O, Valentin L. Clinical importance of appearance of cesarean hysterotomy scar at transvaginal ultrasound in nonpregnant women. *Obstet Gynecol* 2011; 117:525–532.
13. Peipert JF, Grimes DA. The case-control study: a primer for the obstetrician-gynecologist. *Obstet Gynecol* 1994; 84:140–145.
14. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol* 2004; 159:882–890.
15. Ernster VL. Nested case-control studies. *Prev Med* 1994; 23:587–590.
16. Odibo AO, Zhong Y, Longtine M et al. First-trimester serum analytes, biophysical tests and the association with pathological morphometry in the placenta of pregnancies with preeclampsia and fetal growth restriction. *Placenta* 2011; 32:333–338.
17. Lilienfeld AM, Lilienfeld DE. *Foundations of Epidemiology*. 2nd ed. New York, NY: Oxford University Press; 1980.
18. Adzick NS, Thom EA, Spong CY, et al. A randomized trial of prenatal versus postnatal repair of myelomeningocele. *N Engl J Med* 2011; 364:993–1004.
19. Schulz, K, Chambers I, Hayes RJ, Altman DG. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995; 273:408–412.
20. Begg C, Cho M, Eastwood S, et al. Improving the quality of reporting of randomized controlled trials: the CONSORT statement. *JAMA* 1996; 276:637–639.
21. Schulz KF, Altman DG, Moher D; CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010; 340:c332.
22. Hollis S, Campbell F. What is meant by intention to treat analysis? Survey of published randomised controlled trials. *BMJ* 1999; 319:670–674.
23. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet* 2000; 355:1064–1069.
24. LeFevre ML, Bain RP, Ewigman BG, Frigoletto FD, Crane JP, McNellis D. A randomized trial of prenatal ultrasonographic screening: impact on maternal management and outcome. RADIUS (Routine Antenatal Diagnostic Imaging With Ultrasound) Study Group. *Am J Obstet Gynecol* 1993; 169:483–489.