Epidemiologic Measures of the Course and Outcome of Pregnancy

David A. Savitz1, Irva Hertz-Picciotto1,2, Charles Poole1, and Andrew F. Olshan1

1 Department of Epidemiology, University of North Carolina School of Public Health, Chapel Hill, NC.
2 Department of Epidemiology and Preventive Medicine, University of California, Davis, Davis, CA.

Received for publication June 17, 2002; accepted for publication November 21, 2002.

INTRODUCTION

Successful reproduction, the spectrum of events leading from conception to birth of a healthy infant, is both biologically and epidemiologically complex. Problems that arise during the course of the reproductive process define the adverse outcomes in epidemiologic studies of pregnancy. A simplified time line for the process leading from conception to birth is shown in figure 1, along with an indication of approximately when critical events occur. With the focus first on the desired or “normal” outcomes, conception results from a viable sperm’s reaching the ovum and progressing to implantation. Normal development over the first weeks of life depends on differentiation and migration of cells, events that must follow precise timing, leading to formation of organ systems and subsequent fetal growth and development. Reproductive epidemiology encompasses the entire scope of these events, often extending backward to the determinants of conception (e.g., semen quality, menstrual cycles) and forward to postnatal health and development, often through infancy, childhood, and puberty, and occasionally to adulthood. For the purposes of this paper, our focus is on the narrower time frame of conception to birth, with full recognition that the boundaries are arbitrary and sometimes ambiguous.

With this sketch of normal events, we can consider deviations resulting in adverse outcomes, addressed in detail in the following sections. Some general methodological issues in studying reproductive health are first noted. Because the very decision to reproduce is discretionary, some outcomes, for example, conception or failure to conceive, are observable only in couples at risk of conception. Some downstream events can be observed only after essential prerequisites have been met: Successful conception is required to have a fetus that is at risk of spontaneous abortion or being born preterm. When a couple experiences infertility, is this a characteristic of one or the other member of the unit or of the combination? In order to accurately characterize the frequency of events, there is a need for clear specification of the units (persons or fetuses or time periods) actually at risk, which can be challenging.

The temporal nature of the reproductive process is a recurrent theme. The timing of events is often ambiguous in that some outcomes cannot be observed at the time they are occurring. For example, etiologic events leading to infertility may be influenced when the potential parent was in utero and his or her own germ cells were being formed. Congenital anomalies often have their etiology in the first several weeks of pregnancy, but the fetus must survive to some point in gestation to be diagnosed as having the anomaly. The formation of the placenta early in pregnancy may be a key factor influencing risk of preterm birth and fetal growth measured at or near the end of pregnancy. Some outcomes are defined as the failure of something to occur at the right time; for example, when the neural tube does not fully close, spina bifida results. Some outcomes are defined by normal events occurring at the wrong time: Preterm delivery is formally defined solely by the gestational age at which delivery of a livebirth occurs. Finally, manifestation of an event may not be close in time to its recognition: Several weeks may pass between the death of a fetus and the recognition of a miscarriage.

Many of the logistic problems concern the disparity between measures of prevalence (the proportion affected at a given point in time) and incidence (the occurrence of the condition in a previously unaffected set of individuals). The incidence of infertility is difficult to conceptualize and often impossible to measure, for example, the moment when the sperm’s motility falls below the level that would be necessary for conception or when the occlusion of the fallopian tubes makes fertilization impossible. We instead must rely on prevalence at the time of an attempt to conceive, often remote in time from its origins and susceptible to distortion relative to the true incidence of the underlying biologic condition resulting in infertility. Major structural birth defects, often identified through a combination of prenatally diagnosed and terminated pregnancies plus prevalence among livebirths and stillbirths, are far removed in time and
comprehensiveness from the underlying cohort of fetuses in whom incidence would ideally have been measured.

Because some of the issues for statistical analysis are similar across outcomes, a few introductory remarks are in order. The need for careful attention to the timing of events in reproductive epidemiology applies to both study design and data analysis. Explicit recognition is required for the time of the outcome (e.g., gestational age at time of delivery), exposure (whenever it is time varying), and the time that a pregnancy comes under observation, either when the mother becomes aware of her pregnancy (and hence, potentially aware of its termination) or when the pregnancy comes to the attention of the health care system (and hence, under observation by the investigator studying women enrolled in prenatal care). In many cases, the appropriate analyses require use of statistical models that take person-time at risk into account, for instance, Cox proportional hazards models. Hertz-Picciotto et al. (1) introduced the application of Cox models for studies of spontaneous abortion and demonstrated their value in accounting for pregnancies that entered prenatal care at different gestational ages. Such methods also apply to studies of events that occur during the course of pregnancy, including loss, early delivery, or onset of complications.

Although pregnancy is a relatively short time period, exposures can and often do vary considerably from month to month (2). The biologic consequences of those exposures are also expected to vary because the timing of critical embryonic and fetal events is so specific. Exposure at one time in pregnancy may have quite different consequences from exposures several weeks earlier or later, most widely recognized for congenital anomalies, but quite possibly also true for other outcomes. Therefore, the failure to use time-dependent exposure measures can lead to biased results. For example, in the study of pesticide exposures and fetal death from congenital anomalies, the association increased as the time window for exposure was narrowed to the specific period of organogenesis (3). When the time window of relevant exposure overlaps with the time period for the outcome, Cox models are needed to eliminate bias, as shown by O’Neill et al. (4) in an investigation of urinary tract infection and preterm births.

Our goal is to examine the epidemiologic measures commonly applied to assess the course and outcome of pregnancy. For each of the measures, we discuss the operational definition, timing of etiology and recognition, challenges to accurate measurement, and special analytic issues. Even when there are unobserved and thus unmeasurable elements, clear conceptual definitions and terminology are needed to serve as the benchmark to which feasible measurements can be compared. We will also consider the implications of deviations between the ideal and practical measures for the interpretation of study results.

### INFERTILITY

Fertility is defined as “the actual reproductive performance of an individual, a couple, a group, or a population” (5, p. 58), and thus the failure to reproduce defines infertility. In contrast to this performance-based measure, fecundity refers to “the physiological capacity of a woman to produce a child” (5, p. 58), which would probably be more appropriately applied to a woman whose partner is optimally capable of doing so as well. Clinical attention generally but not always is restricted to couples who have experienced unprotected intercourse that does not result in a conception for at least 1 year. Sterility typically implies a more permanent condition, not just a failure for some defined period of time. In fact, the probability of conception is a continuous measure on a population level, with approximately 40 percent of couples conceiving in the first menstrual cycle, another 20 percent in the second, another 10 percent in the third cycle, and declining probabilities each month thereafter (6, 7). There is not a discontinuity in the probability of conception after 12 cycles, and in fact most couples that conceive after 12 months of unprotected intercourse still do so through natural means, not as a result of medical intervention (8). This reflects both the incomplete receipt and limited effectiveness of treatment, combined with the continued, low monthly probability of conception even for couples that have been trying for a full year to achieve a pregnancy.

Given the probabilistic nature of conception, it is generally more informative to examine the monthly probability of conception in a life-table approach or equivalently to evaluate time to pregnancy (6) as compared with dichotomizing couples based on the 12-month cutpoint or any other arbitrary cutpoint. The couple constitutes the unit of analysis, and for each month (actually each menstrual cycle, if this information is available), the probability of conception is
measured as the proportion of couples who did not conceive in the previous cycle but do conceive in the present one. This measure quantifies the couple’s fecundability, or the probability of conception, specific to the number of months the couple has been trying to conceive. To assess the effects of potential determinants of fecundability such as age, tobacco use, or pesticide exposure, researchers compare the monthly probability of conception across exposure levels by calculating fecundability ratios or differences. In principle, even couples that try to conceive and fail to do so within the observation period of the study can be included in a prospective study and contribute to such an analysis. In practice, many studies are retrospective in nature and begin with a pool of couples all of whom eventually conceived, thus excluding the least fecund entirely from the analysis (9).

An alternative approach is to treat fertility and infertility as a dichotomy, assessing the proportion of couples that fail to conceive within a 12-month period. This is referred to as primary infertility when there is no history of conception and as secondary infertility when a pregnancy has been conceived at some time in the past. Most physiologic processes that influence fertility would be likely to diminish but not eliminate the monthly probability of conception, for example, by damaging (but not destroying) sperm production and viability, ovulation, or tubal transport. Hormonally mediated effects are likely to be along a continuum. Other influences may in fact render a couple sterile through absence of sperm or complete ovulatory failure, yielding a monthly probability of conception of zero. Where the effect is truly all or none, that is, the ability to conceive is destroyed or unaffected, analysis of the dichotomous measure would be able to identify the adverse effect of the agent efficiently. However, even if some individuals are rendered sterile, for example, because all sperm production is eliminated, the same causative agent is likely to merely shift the probability of conception downward in others, for example, because sperm production is diminished, and thus continuous measures of fecundability would be more informative. The rationale for this assertion is that many of the common mechanisms for infertility would be expected to operate along a continuum.

The assessment of fecundability should start, ideally, the first month of trying to conceive in order to generate the complete chronology of fecundability. For this purpose, “trying to conceive” would include intercourse at the appropriate time in relation to ovulation in the absence of contraception. For the first month and subsequent months at risk of pregnancy, known influences on fecundability such as coital frequency, contraceptive history, maternal age, and maternal tobacco use (if not the exposure of interest) would be included for consideration as potentially relevant covariates. Couples who enter a study of fecundability at some other point in time, that is, those who have already been at risk for some period but failed to conceive, are an increasingly select subset of less fertile couples, since their more fertile counterparts have already conceived and are no longer eligible to continue in or join the study. Measurement of monthly probability of conception thus needs to specify the absolute number of months at risk. A measure such as “prevalence of subfecundity,” defined as the proportion of couples who report being unable to conceive either at a given point in time or over a defined period of time, will reflect a few couples who have been trying to conceive only for a short time and many couples who have been trying for an extended period, as well as loss of couples who have simply stopped trying. Longitudinal information is essential, although information can be used from couples that enroll beyond their first month of trying to conceive as long as the duration of being at risk is known so that they can be entered appropriately into the life table for analysis.

A number of distinctive challenges to the accurate assessment of fecundability have been brought to the attention of researchers (10, 11). The definition of “trying to conceive” is subjective in that it implies a conscious desire to become pregnant that may not correspond well with behavior. For example, couples may be having unprotected intercourse but not consider themselves as trying to conceive. To the extent that the relevant behaviors can be captured, such couples at risk of pregnancy can still be studied. Couples may choose to time intercourse to increase or decrease probability of conception relative to the menstrual cycle, depending on whether they are actively trying or passively allowing pregnancy to occur. The perception of “trying to conceive” may well differ even between the involved partners, and it is likely subject to cultural variation in the interpretation of the concept.

Studies of time to pregnancy often rely on couples that are consciously planning their pregnancies, depleted of those who conceived without having ever planned the pregnancy. In addition to those who simply did not undertake some form of family planning, some may have experienced contraceptive failure (ineffective use of contraception or high fecundity combined with incomplete contraception), and some knew they were incapable of conceiving and thus never tried to do so. Starting with the unmeasurable biologic capability of a couple to reproduce, fecundity and behavioral and logistic issues make measurement increasingly difficult. Coital frequency, actual and self-reported planning of pregnancies, use and misuse of contraception, and faulty and biased memory of key events all may distort the assessment of determinants of fecundability. Perhaps the most important among these challenges is the ability to extrapolate information from the ideal couples that plan their pregnancies carefully and provide the “cleanest” data to those who have less optimal life circumstances and are often of interest because of potentially hazardous exposures.

The widespread use of assisted reproductive technologies to help couples who would otherwise be unable to conceive has clearly changed reproductive health on the population level in many settings. Pregnancies conceived in this way should be considered separately from natural conceptions for epidemiologic purposes. Often, such pregnancies have higher risk for pregnancy loss, either due to the underlying problem for which intervention was needed or as a result of the manner in which the pregnancy was achieved. Such pregnancies also differ in providing information on the precise timing of conception and monitoring for loss from the earliest phases of development. Finally, assisted reproductive technologies are unevenly distributed in the population,
since couples must possess financial resources to pursue these approaches to overcoming subfertility.

PREGNANCY LOSS

The loss of a pregnancy following conception is a common event, occurring in over 20 percent of pregnancies, but is notoriously difficult to identify with accuracy in the early stages of gestation. Pregnancy can be detected by monitoring hormone levels (biochemical pregnancy) before it can reliably be identified through clinical means, such as missed menstrual period (clinical pregnancy). Prospective studies, whether assessing biochemical or clinical pregnancy, allow measurement of the probability of loss specific to the time in gestation. Given survival to a specific day or week of gestation, we can assess the probability of fetal survival to some later day or week of gestation. When pregnancy cannot be monitored from the point of conception, the life table approach becomes crucial but requires reasonably accurate dating of the conception, so that pregnancies can be entered into the life table at the correct point in gestation and followed forward (1, 12).

The ideal approach has been applied in a limited number of studies that begin prior to conception and carefully monitor the events of early pregnancy (13, 14). The costs are substantial relative to studies of clinical pregnancy, and limited populations are available for study, namely, those women who are planning their pregnancy and are willing to adhere to a rather demanding research protocol involving daily urine collection. These techniques provide accurate identification of pregnancy after implantation has successfully occurred (before the first missed menstrual period, within 28 days of the last menstrual period) (15). Studies of broader populations that are not enrolled prior to conception thus fail to capture the total incidence of pregnancy loss.

Studies of pregnancy loss are generally of three types, defined by the sampling frame. As described above, the ideal approach identifies a cohort of women prior to conception and follows them from the point of conception forward. A second sampling frame consists of a cohort of pregnant women who are monitored from the point of identification or enrollment into the study for follow-up. The third consists of women identified after the end of their pregnancy, whether it was unaware that she was pregnant. On a population level, groups that become aware of pregnancy at an early point in gestation will therefore also be aware of more pregnancy losses and have an artificially inflated measured risk of spontaneous abortion if the timing of recognition and loss is not taken into account. Women who recognize their pregnancy at an earlier point can be enrolled into studies earlier in gestation and contribute to the early pregnancy-weeks at risk, whereas those who do not recognize their pregnancies until later will be uninformative for the early weeks of gestation. Observed pregnancies in the earliest weeks of gestation may be sparse, because few women know that they are pregnant in the first 4–6 weeks, and measures of the probability of loss in that early period could be biased by enrolling an unusual group of women with a poor pregnancy history who are worried and, thus, self-monitoring intensively or those who are extremely health conscious for other reasons. Early pregnancy recognition may thus be a marker of risk, independent of other methodological issues.

Induced abortions pose a distinct challenge to the accurate quantification of risk of pregnancy loss (20). With detailed longitudinal information, induced abortions are often treated as censored observations; that is, in survival analysis they constitute observations for which the outcome cannot be observed. As in any survival analysis, all pregnancies would ideally be enrolled at the time of recognition and censored at the time of
induced abortion. However, the potential for baseline risk of subsequent spontaneous abortion to differ systematically for pregnancies that do and do not end in induced abortion must be considered. Pregnancies that are terminated may be in part a result of medical conditions identified in the mother or fetus (e.g., congenital anomaly), social circumstances making childbearing undesirable at that time, or high fecundability making the couple prone to contraceptive failure, any of which could alter their baseline risk of spontaneous pregnancy loss. However, pregnancies that result in induced abortion should be considered for inclusion even if the concerns with nonrandom censoring cannot be readily addressed.

A distinct logistic challenge in research on spontaneous abortion is that prenatal care is initiated at various times throughout the pregnancy, while recognized spontaneous abortions occur primarily in the second and third months. The timing of the initiation of prenatal care depends on a range of factors, including health status, cultural milieu, previous reproductive experience, insurance status and extent, provider accessibility, and so on. Although pregnancy is recognized before prenatal care onset, there is no centralized resource for identifying large numbers of pregnant women at such an early point in gestation. As noted above, self-selected women who identify pregnancy earlier may well differ in their subsequent risk of pregnancy loss from those who identify pregnancy later. As a result of this, the study of medically treated spontaneous abortion (21) is inherently limited to a potentially biased subset of all pregnancies and losses without clear information to construct a life table. More problematic still, pregnancies may first come to attention at the time of loss; that is, the symptoms indicative of having lost the pregnancy were the first indication of pregnancy and the motivation to initiate prenatal care. Although there is no doubt that a pregnancy has occurred and a loss has occurred, since there is no time under observation in the study prior to the occurrence of the loss, inclusion of such pregnancies and losses would bias the rates and risks of pregnancy loss (1). For this reason, they must be excluded from any estimates of pregnancy loss, whether using survival analysis or other analytic methods.

Statistical methods to address the problem of timing of pregnancy recognition or initiation of prenatal care in studies of spontaneous abortion were first addressed by Taylor (12) and then developed for multivariate analysis by Hertz-Picciotto et al. (1). Briefly, proportional hazards models are ideally suited for estimation of relative risks associated with time-constant or time-varying exposures when the timing of both the initiation of prenatal care and the pregnancy termination is known. Risk sets are constructed on the basis of the pregnancies under observation at each time point, with time defined as days or weeks since the last menstrual period (a surrogate for time since conception). The use of Cox proportional hazards models is a major improvement over simpler methods; generalized estimating equation methods could be used for the same purpose. The possibility of selective entry into prenatal care can still threaten the validity of hazard or rate ratios obtained from such models, although some approaches may reduce bias (1).

**BIRTH DEFECTS**

Collectively, major structural birth defects occur in approximately 3–5 percent of all livebirths. However, each individual type of malformation is rare, with the most common on the order of 1/1,000 livebirths. Even these summary measures, expressed in the form of “prevalence at birth,” suggest the fundamental problem in obtaining accurate epidemiologic measures (22–24). The etiologic events that generate major structural malformations typically occur within the first 3–8 weeks postconception, but the recognition of the malformation may not occur until later in pregnancy (through ultrasound evaluation), at the time of birth, in early childhood, or later in life, or the event may never come to be recognized. Even with prenatal diagnosis, there is always some passage of time after the event has occurred before it is recognized and, thus, the need for continued fetal survival for the malformation to be recognized.

The ideal study would monitor the development of malformations as they are occurring, some at or near the time of conception (e.g., chromosomal abnormalities) and others arising during the early weeks of pregnancy (e.g., neural tube defects, oral clefts). With such information, the true incidence could be calculated specific to the gestational age of the fetus. Because there is a critical period for the initiation of malformations based on the developmental events that occur at specific times in gestation, once the malformation is present or the risk period has passed, subsequent events, such as prenatal diagnosis and pregnancy termination or survival of the fetus to birth, are irrelevant to the etiology yet are crucial to the study of etiology. These intervening events will affect the degree to which prevalence at birth differs from incidence. A fetus that develops the malformation may be lost during gestation. Obviously, unaffected fetuses will also be lost, affecting the denominator as well, although for many malformations, the proportion of losses may be greater among the malformed than normal fetuses. Aside from any true differences in the rate of malformations that are caused by a given exposure, differences in prevalence at birth will arise (or be masked) by an influence of that exposure on survival of malformed fetuses relative to unaffected fetuses (25). The problem of differential survival is not of concern for those malformations that are uniformly fatal (e.g., certain forms of aneuploidy) or never fatal (e.g., oral clefts) but of substantial concern for congenital defects that are variably fatal (e.g., spina bifida, trisomy 21).

To bridge the gap between incidence and identification, Stein et al. (25) proposed examining spontaneous abortions as an alternative measure for birth defects due to aneuploidy. Although the logistic challenges of collecting and karyotyping material from spontaneous abortions are daunting, the goal of moving back from chromosomally abnormal births to pregnancy losses associated with chromosomal abnormalities is well justified. In fact, some forms of aneuploidy found in spontaneous abortions were never seen in livebirths, indicating their incompatibility with fetal development, whereas others were more common in spontaneous abortions but also occurred among livebirths, suggesting decreased survival to birth for affected fetuses.
Prenatal diagnosis, which is now common for certain structural malformations such as neural tube defects, moves the diagnosis closer to the etiologic period. However, the differential use of these diagnostic technologies and the decisions regarding induced abortion further complicate the calculation and comparison of measures of occurrence. Such screening has become so common that measures of prevalence reflect a mixture of true difference in incidence, differential access to or use of prenatal diagnosis, and differential access to or use of elective abortion. For this reason, the more recent studies of neural tube defects (26, 27) consider both the cases that were diagnosed prenatally, whether or not they were terminated, and those carried to delivery, including both livebirths and stillbirths.

Preimplantation genetic diagnosis is becoming a feasible alternative to the common diagnostic methods of amniocentesis and chorionic villus sampling for identifying certain genetic disorders among high-risk couples in the in vitro fertilization setting (28, 29). In preimplantation genetic diagnosis, molecular analysis is conducted on single cell biopsies from embryos prior to implantation. Selection of unaffected embryos can then be made prior to establishment of pregnancy. Advances in this method will result in the diagnosis of aneuploidy and other genetic abnormalities with greater sensitivity and thus greater impact on induced abortion and alteration of birth defect prevalence measures among nonterminated pregnancies.

PREGNANCY COMPLICATIONS

A variety of conditions that are detrimental to the mother’s and fetus’s health can arise during the course of pregnancy, including pregnancy-induced hypertension, gestational diabetes, placenta previa, and placental abruption. Pre-existing chronic diseases such as hypertension or diabetes are also harmful to pregnancy, often more severe than the pregnancy-induced counterparts. Pregnancy complications can be quite common, the most frequent affecting about 5 percent of all pregnancies. The ideal approach to assessment would be to monitor pregnancies prospectively in order to identify the time of onset of the event during the course of gestation. With such an approach, the week-specific incidence could be quantified, as well as cumulative incidence for longer intervals of interest, including the entire pregnancy. Instead of explicitly accounting for time of onset, it is much more common to quantify the proportion of women affected during the pregnancy, for example, the proportion of women who develop gestational diabetes at some time during the course of pregnancy. Assessment of the gestational week at identification would be preferred, recognizing that the precision of this measure will depend in part on the frequency of prenatal care visits or the woman’s initiative in seeking medical care for the complication.

The loss of information that results from failure to identify time of onset is typical of measures of risk versus rates or of risks over long intervals versus risk over short intervals. However, not all events are necessarily equal in regard to both etiology and consequences. Pregnancy-induced hypertension in week 35 of gestation may be different in its origins than the same event in week 20, and it clearly differs in its clinical impact. The treatment for a number of pregnancy complications, including pregnancy-induced hypertension, is delivery, and thus occurrence late in pregnancy poses much less of a threat to the health of the infant than the same condition early in pregnancy. By quantifying the gestational age-specific incidence of complications, allowing for the examination of risks in defined pregnancy intervals, the different clinical consequences related to time of onset can be taken into account.

A further complication arises from conditions that wax and wane over the course of pregnancy. Placenta previa, in which the placenta partially occludes the cervical os, may be present at one point in gestation but, through growth of the uterus and migration of the placenta, may be absent at a later time. Similarly, metabolic disorders such as diabetes may come and go during the course of gestation. For such conditions, the most fully informative measure would be prevalence at specific intervals of gestation, for example, occurrence of placenta previa during weeks 30–32.

The period of risk for pregnancy complications ends when the pregnancy ends, so that the total time at risk for a complication will vary as a function of the total duration of pregnancy. For complications that often arise very late in gestation, such as hypertension, the cumulative risk will be lower, all other things equal, for pregnancies that end earlier. Efforts to study influences on pregnancy-induced hypertension thus need to consider the timing of the onset of hypertension, because a measure of “hypertension at any time in pregnancy” would be affected by the duration of the pregnancy (2, 4). Pregnancies that continue longer will have an increased opportunity to develop hypertension, so that a risk factor for preterm birth could mistakenly be viewed as being inversely related to the risk of pregnancy-induced hypertension.

FETAL GROWTH

Fetal growth provides evidence of healthy development in utero and predicts postnatal health and development. Longitudinal information on fetal growth per se is rarely available, but what is known instead is attained size at birth, most often reflected in birth weight. On a population level, the distribution of birth weight can be considered to consist of a dominant distribution, largely Gaussian, and a residual tail to the left of very small infants who fall outside the dominant distribution (30, 31). Those infants whose weight places them in the residual tail of the distribution are at high risk of mortality and increased morbidity, and interventions that reduce the number of infants in that residual tail will be beneficial. On a population level, however, a shift in the main portion of the distribution as reflected by higher mean birth weight sometimes has no effect on infant mortality. Mortality for individuals is predicted by the deviation from the population’s mean within the dominant distribution. This pattern holds for male versus female infants, with males larger but not at lower risk for adverse outcome as a group, and infants born at high versus low altitude, with higher altitude associated with lower birth weight but not with higher mortality.

From this perspective, “low birth weight” is not as useful as once thought, because babies in this group represent a mix
of those whose growth is suboptimal, those whose growth trajectory was normal but who were delivered early, and those who are small for genetic reasons unrelated to viability. The challenge is to identify which infants fall into the three groups. As noted some time ago (32), small babies of smokers born at term fare better than small babies of nonsmokers born at term, as do small babies of other groups that have lower birth weight distributions (females, Blacks, Asians). This apparent paradox is eliminated when birth weight is standardized within each subgroup before making comparisons across subgroups. Comparing infants of smokers and nonsmokers who are at a given percentile of their distribution as opposed to those having identical birth weights reveals increased mortality for the infants of smokers across the entire size distribution. It appears that deviations from the subgroup norm are most pertinent to future risk. Various approaches to calculating these within-population deviations have been used (30, 33).

Historically, the most commonly used indicator of “prematurity” (often without regard to whether it meant small or early or both) was birth weight below 2,500 g (low birth weight), purely a measure of size that does not take into account how long the pregnancy lasted or the infant’s inherent growth potential. Without further subdivision of infants weighing less than 2,500 g, the pool reflects many infants who are part of the dominant birth weight distribution but simply fall toward the lower end, along with an unknown fraction who are truly in the residual tail of the distribution due to shortened gestation or some growth-retarding experience and at increased risk for adverse outcomes. The duration of gestation can and should first be isolated as a cause for small size at birth, as discussed in the following section. A baby of normal size born at 36 weeks would correspond to an unusually small 39-week newborn, for example, although both are of identical weight, thus making weight for gestational age more informative than birth weight alone. Although preterm infants can be isolated from other small infants, what cannot be known on an individual level is whether the infant is a healthy, small baby or has truly suffered from growth retardation relative to its growth potential. More restrictive cutoffs, such as 2,000 or 1,500 g, are increasingly effective in isolating infants in the residual distribution from those in the dominant distribution and enriching the case group with infants who suffered from growth restriction.

The ideal assessment of the determinants of fetal growth would start from the expected growth trajectory for each infant through the course of pregnancy and quantify the deviation in actual growth relative to the expected pattern. Measures of deviation from the subgroup norms, based on percentiles or standard deviation units, are intended to serve this goal. For example, applying standards for defining small-for-gestational-age (31, 34, 35) asks whether a given infant is unusually small in a statistical sense (below the 10th percentile) relative to other infants of the same duration of gestation, gender, parity, and race. By defining “abnormal” as statistically unusual, for example, below the 10th percentile of weight for gestational age or more than 2 standard deviations below the mean, a set fraction of any population of infants must be classified as growth retarded. The presumption is that those who are at the extreme low end of that distribution of birth weight are likely to have had their growth restricted in some manner. The same rationale underlies the simple measure of “term low birth weight” in that the combination of those attributes suggests adequate time for growth to have occurred but the baby remains relatively small.

There are a number of problems in these efforts to use measures of size to assess suboptimal growth. The distribution of births at a given gestational age is not a random sample of all pregnancies that have attained that gestational age but restricted to those that ended at that point. Particularly for early births, the infants who are born are notably smaller on average than those who continue to develop in utero as assessed using ultrasound (36, 37), suggesting that failure to grow normally may signal early delivery or that growth restriction and preterm delivery share an etiology. A more optimal approach would be to have longitudinal information on a large population of unselected pregnancies measured in utero by ultrasound to estimate weight, allowing fetuses or births at a given point in gestation to be compared with the appropriate referent group of all fetuses of comparable gestational age. Finally, as discussed in detail below, accurate identification of duration of gestation is challenging, and errors will result in shifts in the percentile of weight for age: An infant of a given weight could represent a normal 36-week birth, a large 34-week birth, or a small 38-week birth.

The timing of growth and of when deviations from expected growth occur is not reflected in size at birth, only the cumulative growth to the time of delivery. The trajectories of fetal growth are known to differ by race, gender, and plurality (31, 38) and may well be affected by exogenous factors. Environmental insults may have a transient effect on growth, for example, slowing it for a time and allowing the fetus to catch up later. With longitudinal information on fetal size through the course of pregnancy, attained weight at specified points in pregnancy could be studied as well as growth during specified intervals.

Many investigators have chosen to focus on birth weight as a continuous measure, avoiding an arbitrary dichotomy and potentially enhancing statistical power. If in fact some condition or exposure shifts the entire birth weight distribution, the identification of that shift will be enhanced when examining mean birth weight as compared with the proportion below some cutpoint, such as 2,500 g. The rationale is that, although the growth-retarded infants cannot be identified individually, shifting all birth weights downward will increase the number and proportion of births that are abnormal. However, shifts in mean birth weight will be driven by shifts within the dominant distribution and will not be sensitive to changes in the tail of the birth weight distribution. As shown by Wilcox (30), changes in the overall mean of the dominant distribution, without changes in the residual, do not seem to be of importance for infant mortality and likely not for other forms of morbidity either.
PRETERM BIRTH

Preterm birth is defined solely by the duration of completed gestation at the time of birth, by convention expressed in completed weeks or days. The most commonly used cutpoint to distinguish preterm from term is 37 weeks, so that an infant born before the completion of the 37th week of gestation (259 days) is referred to as preterm and an infant born after that time is referred to as term. Other more stringent cutpoints can be applied to define more extreme deviations from the normal 39- to 40-week pregnancy and thereby isolate more severely affected infants, for example, 34, 32, or even 30 weeks’ gestation.

There are substantial challenges to accurately assessing the time of conception, and thus the duration of gestation, except under such extraordinary circumstances as in vitro fertilization. Not surprisingly, all routinely available clinical indicators are demonstrably fallible (39–42). The date of the last normal menstrual period is most often used as the marker of the beginning of pregnancy, although the actual ovulation that led to conception is believed to occur, on average, approximately 2 weeks after the menstrual period itself. Nevertheless, by convention, the obstetric dating based on last menstrual period is most commonly used. Inaccuracies in last menstrual period dating arise because of erroneous recall of dates and errors for certain individuals in the assumption that ovulation always follows menstruation by two weeks (41, 42). In particular, delayed ovulation appears to account for much of what is interpreted as postterm deliveries (43).

The accuracy of estimating gestational age has been enhanced markedly with the widespread use of ultrasound to visualize the fetus and assess development. Early in gestation, prior to the completion of 20 weeks, there is thought to be limited variability in growth, so that fetal size accurately reflects duration of gestation. Nevertheless, all methods of dating the pregnancy that rely on stage of development are susceptible to circular reasoning, analogous to the once-popular pediatric assessment in which the infant’s physical maturity is used to infer the duration of its gestation. There is no way to distinguish between a fetus or infant of a given gestational age who is unusually large or otherwise physically advanced and a normal fetus of greater gestational age.

Many researchers use more stringent cutpoints for preterm birth, motivated by the advances in the quality and efficacy of neonatal care that have markedly lowered the risk of marginally preterm infants. Infants of 35 and 36 weeks’ gestation pose little problem in clinical management, and increasingly, infants of even 33 and 34 weeks are not the primary focus of obstetricians. Cutpoints as low as 32 or even 30 weeks are commonly considered because those define severe, obvious adverse outcomes. However, the risks for survival and many forms of morbidity do rise as the duration of gestation falls below 37 weeks, although the risk rises more and more steeply as the duration becomes progressively shorter (44). Considering the number of affected births, with many more marginally as opposed to severely preterm infants, a public health perspective argues for a focus on the full spectrum of preterm infants and not just those who are severe enough to demand immediate clinical attention.

The markedly different consequences of very early preterm births as compared with marginally preterm births call into question the appropriateness of treating preterm birth as a dichotomy. The implicit assumptions are that all births before 37 weeks’ gestation are equal in risk as well as all births after 37 weeks, both of which are erroneous. In response to this concern, some investigators consider gestational age as a continuous measure across the entire spectrum of ages, without recognizing that the impact of a shift of a given magnitude (e.g., 1 week) depends markedly on the absolute gestational age. When the analytic approach considers mean differences in gestational age across exposure groups, most births will be in the 38- to 41-week interval, and the analysis will thus be sensitive primarily to shifts within the normal range of gestational ages. However, it is the shifts in births within the <37-week interval that are of profound health consequence. An analogous problem of lesser importance is the assumption that all births after 37 weeks are equal. Although postterm delivery is potentially detrimental, clinical intervention to precipitate delivery has reduced the importance of this concern.

The timing of delivery depends on both the natural course of the pregnancy and the clinical management of that pregnancy. Early signs of possible preterm labor, for example, result in attempts to prevent progression to preterm birth. To the extent that those efforts are successful, a postponed, possibly term delivery may result. On the other hand, when there are signs of fetal distress or pregnancy complications that threaten the health of the pregnancy, intervention and early delivery may be viewed as the favored approach and result in a medically indicated preterm birth. Such decisions reflect obstetric practice, as reflected in the marked social and demographic variation in the use of labor induction (45). Medically induced early delivery may well prevent some intrapartum deaths and delay some losses until the neonatal period. Given this mixture of spontaneous events and medical interventions, the outcome of preterm birth itself is heterogeneous (46). In particular, the informal assumption that preterm delivery and preterm labor are more or less synonymous is incorrect: Many instances of preterm labor do not result in preterm delivery, and many preterm deliveries are not preceded by preterm labor. For understanding the human biology of pregnancy, the ideal approach would be to allow pregnancy to follow its natural course, yet clinical interventions are common and thus an important aspect of epidemiologic research on preterm birth.

The ideal approach to measuring preterm birth would start with an accurately measured date of conception and follow the pregnancies up to the point of delivery. To enhance the information on date of delivery alone, we would want information on the timing of onset of labor, timing of rupture of the chorioamniotic membranes, and clinical interventions that either postpone or accelerate the timing of delivery. With information on the exact timing of delivery, the analytic options would allow for any number of different dichotomies but also for an analysis using time-to-event models. One could study time to onset of labor, for example, censoring those pregnancies that are interrupted by obstetric intervention. Similarly, the time to spontaneous rupture of membranes could be examined. The precise timing of
biologic and clinical events would open up a range of analytic approaches to addressing more refined questions about the course of late pregnancy and events surrounding delivery. Flexible analytic approaches that can consider the conditional risk of delivery (or related events such as labor onset or membrane rupture) by week of gestation would be optimal, perhaps combined with some weighting methods that consider the varying implications of delivery as a function of gestational age.

One common error that arises in studies of preterm birth that rely on vital records is a mismatch of the timing of the births that occur in the numerator and denominator, as noted earlier with regard to studies of pregnancy loss. When the timing of the births rather than conceptions is the basis for cohort definition, for example, all births in a given calendar year, the preterm births ending for instance in December of the preceding year are missed, and term births conceived at the same time might end in January or February and thus be included. At the end of the interval, preterm births in December would correspond to conceptions that were due to end in January or later of the subsequent year. In a steady state or for long calendar periods, this may have little consequence, but where there are strong seasonal or secular trends in influences, there could be biases that would arise from this mismatch of preterm births and the pregnancies at risk.

**DISCUSSION**

Certain recurrent, methodological issues pose challenges and force compromises in the ascertainment of pregnancy outcomes. The timing of events from conception to the end of pregnancy is crucial to epidemiologic study, yet many of the events of interest are difficult to pinpoint at the time of occurrence. The timing of conception is rarely known with precision, the loss of viability of a fetus is usually recognizable some time after it has occurred, birth defect onset is rarely recognized, and fetal growth patterns are usually not measured through the course of pregnancy. When the information does become accessible, it is often incomplete with regard to the timing or comprehensiveness of ascertainment. Researchers often use traditional measures that do not consider the timing of onset (e.g., prevalence of birth defects among livebirths), or we rely on fallible estimates of the timing of events (e.g., assigning fetal demise as the time of clinical recognition of pregnancy loss). At a minimum, we need to be explicit in our assumptions and make full use of the information that is known, for example, assigning onset of events to intervals when we cannot be more precise or imputing the time of occurrence based on available information.

Because the measures of occurrence are susceptible to inaccuracy, the potential for differential accuracy as a function of exposure must be considered. If exposure is associated with recognition of an event, whether or not it is also associated with the occurrence, bias in the measure of association will result. For example, the timing of pregnancy recognition greatly affects the cumulative risk of pregnancy loss, and some exposures may be associated with more intensive monitoring and awareness of pregnancy. An exposure that is associated (not necessarily causally) with greater desire to conceive, with more regular menstrual cycles, or with greater concern about the health of the pregnancy could result in earlier recognition of pregnancy and increased risk of recognized loss of pregnancy. Again, explicitly taking into account the time when pregnancy is recognized will help to identify such differences among exposure groups and provide an analytic framework to compensate for those differences.

Where there is a substantial gap between the causal event resulting in the adverse outcome and its measurement, as is the case for many birth defects, the exposure may influence the probability that the affected fetus will survive to the point of recognition. A given exposure may have no effect on the etiology of a malformation but may change the probability of survival of the malformed fetuses to birth, thus resulting in either increased or decreased birth prevalence. Whenever the measure is “prevalence at birth” or “proportion of pregnancies affected,” questions about the deviations from true incidence should be asked.

In order to fully appreciate the compromises that are necessary and the consequences of that imperfection, it is useful to specify the ideal measure. Suspending all practical and ethical constraints, what would we really like to know? How would the population at risk be identified and followed through time and how would cases be identified? With such a sketch in hand, decisions can be made about where to compromise along with an assessment of the implications of that compromise. Increasingly, advanced approaches in molecular biology (e.g., monitoring of human chorionic gonadotropin to identify timing of conception and subclinical pregnancy loss) and diagnostic technology (e.g., ultrasound to assess fetal weight in utero) allow noninvasive but direct assessment of the events and conditions of interest. The tools developed for clinical medicine often provide windows into otherwise unobservable processes, allowing noninvasive assessment of fetal development, growth, placental vasculature, genetics, and so on. Although not all such methods are feasible or useful in large epidemiologic studies, they may provide an opportunity to compare a more accurate measure with a routine measure and thereby inform decisions regarding study methods.

Not all studies would be expected to apply such technologies, but integrating information from more ideal and less optimal approaches will enable us to quantify the magnitude of deviation and to better interpret studies that rely on the more fallible measure. Tradeoffs between smaller, more precise studies and larger, cruder ones often must be considered (47). The populations that can provide the most detailed information are smaller and more highly selective than would be desired, but such studies can tell us what is lost in more routine approaches to the same question. Validation substudies, in which more intensive, demanding, expensive approaches are applied to a subset of participants, provide information for making comparisons and yield information that can be incorporated into refined statistical estimates that take account of the quality of the routine measurements (48).

At a minimum, the precise operational measures used and their implications need to be fully described. In reproductive epidemiology, there are few routinely applied measures that are so simple as to not warrant detailed description. Inevitably, investigators respond to the opportunities available in
their study setting, and the result is tremendous variability in the quality of assessment. As technology and clinical practice change, the methods of ascertainment evolve as well, raising concerns about the integration of research findings across studies. With increased popularity of assisted reproductive technologies, for example, the ascertainment of infertility has been altered. Screening and selective termination of pregnancies affected by certain developmental abnormalities change the case-finding mechanisms and completeness of ascertainment for certain congenital anomalies. Fetal ultrasound has greatly affected the assessment of pregnancies affected by certain developmental abnormalities. As the menu of approaches to assessment expands, the need for definitions of outcomes to be clearly articulated is greater than ever. The full array of requirements to be at risk of becoming an identified case must be specified and the operational case definition described in detail. Opportunities for false positive and false negative assessment should be considered, along with the likelihood that such error would differ in relation to exposure status. Finally, with a clear recognition of where the major compromises have been made, opportunities to do better should be sought through identifying favorable study populations, incorporating new tools into epidemiologic protocols, and refining study design, conduct, and analysis to reduce the impact of the recognized limitations.

ACKNOWLEDGMENTS

This work was supported in part by a grant from the National Institute of Environmental Health Sciences (P30ES10126).

REFERENCES