Methods for Assessing the Effects of Chemicals on Reproductive Functions
Edited by V. B. Vouk and P. J. Sheehan
© 1983 SCOPE

Epidemiological Approach To Human Reproductive Failure Assessment

DOROTHY WARBURTON, ZENA STEIN and JENNIE KLINE

1 INTRODUCTION

The most direct method of assessing the effects of chemical exposures on human reproduction is to look in human populations for an association between the exposure and the frequency of end-points indicating reproductive failure or dysfunction. That such an epidemiological approach is not simple, quick, inexpensive or without pitfalls is attested to by the widespread attempts to find laboratory methods which can accurately predict human reproductive performance. These laboratory methods have the advantages, not immediately available from an epidemiological approach, of providing clues to the mechanisms involved and demonstrating potential hazards before actual human populations are allowed to be exposed. However, the validity of any laboratory approach must ultimately depend on a proven relationship between the laboratory end-point and human population risks. The uncertainties which arise when this kind of evidence is not available are well illustrated by:

(1) the problem in extrapolating from animal data on teratogenesis to the human situation; and
(2) the difficulties in interpreting the significance of induced somatic chromosomal aberrations in man.

The existence of great interspecific and intraspecific variability in susceptibility to teratogens (Cahen, 1966) makes the prediction of human risks from animal experiments very unreliable, as was well illustrated by thalidomide, which was not teratogenic in rats and mice. While chromosomal damage can clearly be induced both *in vitro* and *in vivo* by chemical exposures in man, the usefulness of such damage as a predictor of reproductive or other problems in the individual or in the population has yet to be established. These problems will not be solved until epidemiological studies confirm an association between the chromosomal effects observed in the laboratory and reproductive outcome.

This paper will attempt to describe the epidemiological approaches currently available for the investigation of human reproductive failure related to chemical

exposures. The special problems of experimental design and statistical analysis in this area will be pointed out, and suggestions made as to how improved methods of data recording and gathering could improve such studies in the future. This paper is based largely on methods available in the United States, with the understanding that our country often lags behind in systems of health care and vital statistics which facilitate epidemiological research. We will draw heavily on a document which was prepared for a Conference on the Evaluation of Human Populations Exposed to Potential Mutagenic and Reproductive Hazards, which was held in Washington, DC in January 1981, under the sponsorship of the

Table 1 Reproductive end-points to indicate reproductive dysfunction

Sexual dysfunction: decreased libido; impotence	
Sperm abnormalities: decreased number; decreased motility; abnormal morphology	
Subfecundity: abnormal gonads, ducts and external genitalia; abnormal pubertal development; infertility of male or female origin; amenorrhoea; anovulatory cycles; delay in conception.	
Illness during pregnancy and parturition: toxaemia; haemorrhage	
Early fetal loss (to 28 weeks)	
Late fetal loss (after 28 weeks) and stillbirth	Perinatal death
Intrapartum death	
Death in first week	
Decreased birth weight	
Change in gestational age at delivery: prematurity; postmaturity	
Altered sex ratio	
Multiple births	
Birth defects, major and minor	
Chromosome abnormalities in fetal deaths, at amniocentesis, in perinatal deaths, in livebirths	
Infant death	
Childhood morbidity	
Childhood malignancies	
Age at menopause	

March of Dimes Birth Defects Foundation, the Center for Disease Control, the National Institute for Occupational Safety and Health, the National Institute for Environmental Health Sciences, and the Environmental Protection Agency. The document was prepared by a committee, chaired by Dorothy Warburton and Zena Stein, and consisting of L. Edmonds (CDC); M. Hatch (Columbia); J. Kline (Columbia); L. Holmes (Harvard); P. Shrout (Columbia); M. Weinstock (Columbia); D. Whorton (Berkeley); A. Wyrobek (Livermore).

2 REPRODUCTIVE END-POINTS

Table 1 lists end-points which might be useful for assessing human reproductive failure or dysfunction. Some of these were discussed in detail from the laboratory point of view in other papers at the present meeting. However, the questions of study design and analysis are equally relevant to the evaluation of these traits in a population.

The inter-relationships among these end-points are complex, and largely unexplored in man. In a few circumstances the type of adverse outcome which is most likely can be predicted from what is known about the mode of action of the chemical exposure; in other cases finding a particular reproductive outcome will rule out certain kinds of mechanisms of action, e.g., constitutional chromosome abnormalities cannot be the result of tubal disease. Table 2

Table 2 Relation of reproductive outcomes to environmental insults

Environmental insult	Mechanism of action of toxic substance	Observable outcome
Paternal toxin	Defective spermatogenesis	SP, I, nSA, aSA, BD
	Defective Wolffian duct system	SP, I
	Defective sexual performance	I
Maternal toxin (preconception)	Defective oogenesis	I, nSA, aSA, BD
	Defective Mullerian duct system	I, nSA, P
	Defective hormonal regulation	I, nSA, P
Maternal toxin (postconception)	Defective gestation–nutrition, hormones, placental function	I, nSA, SGA
	Defective parturition	P, PM
Mutagen (maternal or paternal exposure)	Defective gametes	SP, I, nSA, aSA, BD
Teratogen	Defective fetal anatomical development	I, nSA, BD
Fetal toxin	Defective fetal development	SGA, SD
Carcinogen	Transplacental carcinogenesis	CA

* I, infertility; nSA, karyotypically normal spontaneous abortion; aSA, aneuploid spontaneous abortion; P, prematurity; BD, birth defects (malformations); CA, cancer; SD, slow development; SGA, low birth weight for gestational age; PM, postmaturity; SP, abnormal semen.

attempts to relate the observable reproductive outcome to the mechanisms by which a chemical exposure might act. Many exposures might be expected to have multiple effects, e.g. infertility, fetal loss, birth defects, and decreased birth weight may all result from an agent which is toxic to embryos, but for which there is a difference in susceptibility of the mother and fetus, and in the timing of the exposures. Similarly an agent leading to meiotic abnormalities will result not only in offspring born at term with chromosome abnormalities, but also in infertility and fetal loss, depending upon the stage at which the chromosomal imbalance was lethal.

3 GENERAL PRINCIPLES OF EPIDEMIOLOGICAL INVESTIGATIONS

3.1 Definitions of Levels of Investigation

One may envision an epidemiological investigation of reproductive failure as arising from two kinds of situations, which we call a 'cluster' and an 'exposure'. A 'cluster' arises when a population appears to experience a rise in the frequency of some adverse reproductive outcome. Health authorities may be alerted by the community itself (e.g., a labour union), by alert physicians in the area, or by a monitoring system if such is in place. Examples would be the increase in phocomelia observed in Europe during the 1960s (Taussig, 1962), and the increase in spontaneous abortion reported in Oregon populations exposed to 2, 4, 5-T (USEPA, 1979). Here the supposed outcome is known; the task is to establish whether there are real grounds for considering that an increase has occurred, and if so, to attempt to find the exposure which might be involved.

An 'exposure' arises when a population is known or suspected to be exposed to a potentially hazardous agent. Examples would be the Seveso incident in Italy, and the Three Mile Island nuclear accident in the US. In this case the exposure is known, but the range of reproductive end-points which might be affected is not known.

In practice, the distinction between a cluster and an exposure may not always be possible, since the population experiencing a cluster may have already attributed this to a particular exposure. The response to such a situation may include several stages involving varying amounts of time, money and community involvement, which can be classified as follows. In level I studies the investigator looks only at readily accessible sources of data, there is little community contact, and the investigation can be carried out quickly. In level II studies full scale epidemiological investigations (case-control or cohort) are put in place, involving individual contact with the population, and the gathering of new data on exposures and outcomes. This may involve acquiring laboratory or interview data, or undertaking more extensive record searching than in level I. In level III

studies longer term follow-up studies are undertaken of a population at risk, requiring the greatest investment of time and money.

3.2 Statistical Considerations

3.2.1 *The Concept of Statistical Power*

A consideration of statistical power is crucial to any investigation of the hazardous effects of chemical exposures. Power is a measure of the ability of a study to detect an effect, and is defined as $(1-\beta)$ where β is the probability of falsely accepting the null hypothesis. Power depends upon the sample size, the size of the effect which is present, and the level of significance which is chosen, i.e. the probability of falsely rejecting the null hypothesis (Cohen, 1977). For a given study, then, one can calculate for a chosen value of power the magnitude of the effect which is likely to be detected. The negative results of a study which has demonstrated no statistically significant effect are best expressed in a 'power statement' of the form: 'This study has 95% power to detect a 3.5-fold increase in the rate of congenital anomalies', i.e., only 5% of the time a true effect of the given size would be missed by chance.

Experimentalists are more accustomed to thinking of minimizing their chances of falsely *rejecting* the null hypothesis, and may be rigorous in requiring a significance level of 0.01 before considering the experiment to have disproved the null hypothesis. However, when the investigation attempts to establish whether there is a harmful effect of an environmental hazard on human reproduction, it may be more important to minimize the chances of missing a significant effect. An exposed population may not be content to remain in an area where an investigation showed no significant effect, if there is still a one in five chance that a three-fold increase in reproductive problems has occurred. In the following sections, methods of experimental design and analysis will be considered largely from the stand point of maximizing the power of the investigation.

3.2.2 *Sample Size*

Increasing the sample size is the simplest way to increase the power of an investigation. However, studies of environmental exposures are often limited in this respect by the small size of the exposed population. This is particularly true when considering reproductive effects, which will affect only a part of the group at any one time. For example, although there may be 1000 exposed workers at a chemical plant, only 100 pregnancies may occur during the relevant period of study. The power of the study can be somewhat increased by increasing the control group. The effective size of a comparison with unequal groups is equal to the harmonic mean of the two groups, i.e., $n = 2n_1n_2/(n_1+n_2)$. Increases

beyond three-fold in the size of the control population compared to the exposed population are not worth the additional effort.

3.2.3 *Selecting the End-point*

The higher the frequency of an outcome in the unexposed population, the greater will be the power of the study when sample size is constant. It will take a very much larger sample to detect a doubling of a rare event than a frequent event. Table 3 illustrates this, giving the sample size necessary to detect an effect of a particular size for various reproductive end-points.

It will therefore be an advantage for the investigator to choose those reproductive end-points which are not only biologically relevant to the exposure but are most common. The observed frequency of a particular outcome can be maximized also by increasing the intensity of case-finding, or by restricting the investigation to a sample where a particular outcome is more frequent. Thus, chromosomal abnormalities are 100 times as common in spontaneous abortions as in live births, so that a possible increase in rates of non-disjunction might be studied in aborted fetuses rather than in live births.

3.2.4 *Significance Level*

Significance levels of 0.05 or 0.01 are commonly used in experimental situations. Since we will only reject the hypothesis of no association if the frequency of adverse reproductive outcomes is greater in an exposed than in an unexposed

Table 3 Relationship of power to end-point frequency

Endpoints	Frequency per 100	Unit	Sample size of two equal comparison groups	Increase in frequency detectable with 95% confidence
Major congenital malformations	3	Live births	50	7.7-fold
			100	5.3-fold
			250	3.3-fold
Stillbirths		Full term births	300	3.2-fold
Birth weight ≤ 2500 grams	7	Live births	50	4.6-fold
			100	3.4-fold
			250	2.3-fold
			300	2.3-fold
Spontaneous abortion,	15	Pregnancies	50	3.0-fold
			100	2.4-fold
			250	1.8-fold
Infertility		Couples	300	1.7-fold

$\alpha = 0.10$, two-tailed.

population, we can validly use a one-tailed test. In some situations one might consider setting the significance level at 0.10, at least as a level to indicate a need for further investigation, since the undesirability of falsely concluding that no effect exists may outweigh the undesirability of falsely concluding than an effect does exist.

Whenever an investigation must explore a large number of possible associations between end-points and exposures, some of the hypotheses tested may show statistically significant results by chance alone. Thus if one tests for an increase in three adverse effects, with a 5% significance level, the chance that one of the three will have a statistically significant increase is greater than 5%. The appropriate adjustment is to use a significance level $= \alpha/k$ where α is the desired level of significance, and k is the number of hypotheses tested (Miller, 1966). The necessity for increasing the significance level in this kind of situation decreases the power of a study. It is always advisable to formulate specific hypotheses concerning the exposure to be examined or the type of reproductive outcome which can be reasonably expected to occur, since for these *a priori* hypotheses the significance level does not have to be adjusted.

3.2.5 *The Magnitude of the Effect*

The actual magnitude of the effect is of course not under the control of the investigator. However, the observed association will be greatest if all sources of variation irrelevant to the association are kept to a minimum. Misclassification of exposure or outcome adds noise to the system and thus decreases the observable association.

Precise data on exposures may be very difficult to ascertain. Real exposures are likely to involve multiple toxic agents such as those found at chemical dump sites, and the exposure may vary in time as well as in amount from individual to individual. Studies may be carried out some time after an actual exposure has occurred, so that current measures of exposure in air or water may be uninformative. Occupational exposures may provide better data, but must be subdivided according to which parent is exposed. Biological effects such as somatic chromosomal damage, increased sister chromatid exchanges, or other observables described at this meeting may provide measures of exposure, but this remains to be documented. A combination of laboratory measures of exposure and an epidemiological investigation of reproductive outcomes might provide the strongest approach to the problem of assessing reproductive hazards in human populations.

Misclassification of a reproductive outcome can be avoided by making the investigated outcome as specific as is possible from knowledge of the type and timing of the exposure. Thus if the frequency of conditions due to new gene mutations is to be investigated, one should exclude cases of single gene disorders born to known carrier parents, and cases of malformations largely believed to be

the result of *in utero* positioning. Similarly, a paternal exposure is unlikely to act during gestation itself, although one can postulate mechanisms for such an effect, e.g. substances carried in the semen at intercourse, or on work clothing.

The observed effect size can be altered by confounding variables, i.e. variables which relate both to the reproductive outcome and the chance of exposure. The bias introduced in this way may lead to faulty conclusions, either positive or negative, about the existence of an association. As an example consider febrile illness during pregnancy, which might be associated both with an increased rate of fetal loss and the ingestion of particular drugs. A false association between the drugs and the fetal loss might then be inferred.

Statistical methods such as 'log-linear' regression for discrete data and multiple linear regression for continuous data will control for confounding variables and permit an estimation of interactions. One may also simply stratify the data for analysis according to relevant variables (e.g. smoking women *vs* non-smoking women) and use statistical techniques such as the Mantel–Haenszel procedure or Fisher's method to combine the evidence from several strata and test the consistency of an effect across strata.

3.2.6 *Use of Conceptions Rather than Parents as Unit of Study*

One way to increase the size of the sample is to consider individual conceptions as the unit of study rather than men or women, since each individual may have several conceptions in the time under study. The danger in this approach is that conceptions of the same individual are more similar than conceptions of different individuals, and this may result in undue weight being given to the characteristics of the individuals with the most conceptions, who may not be representative of the population. One must first estimate the degree of correlation among pregnancies of the same parent, and compare this with the correlation among pregnancies of different parents in the exposed and unexposed groups. If the differences are slight, than a considerable gain in power may be possible by using each conception as a different observation.

3.3 Level I Investigation: Use of Readily Accessible Data

A level I analysis will consist of a comparison of the observed frequencies of reproductive events with available baseline data, in order to decide whether a significant increase in an adverse outcome has occurred.

Only certain reproductive end-points routinely included in vital records or screening programmes will be available for this type of study. In most parts of the world these will consist of late fetal loss, birthrate, incidence of low birth weight and some congenital malformations. Newborn or prenatal screening programmes may also be used when available, the latter being especially valuable in

providing an early warning of increases in some kinds of reproductive problems such as neural tube defects. In a 'cluster' situation level I studies can only be used if the end-point under suspicion is available in vital records. In 'exposure' situations a range of possible outcomes must be tested, and these might begin with those available for level I analysis.

In order to calculate the expected frequencies of the outcomes under study, one must define the population at risk, and then determine the baseline frequencies. For an exposure situation, the population may be defined by choosing a given geographical distance from the exposure source, or a given time period of exposure. For a cluster situation the problem is more difficult, and a political subdivision such as a city or county, with common vital statistics records, may be the most convenient way to define the population at risk.

Baseline frequencies may be established using data collected from the same population before the suspected exposure or cluster, if they are available. In their absence, state or national statistics may have to be used. Birth rates, infant mortality and low birth weight are all accurately recorded in the United States by the National Center for Health Statistics, and available for states and metropolitan areas in some years. Malformations are much less reliably recorded on birth and death certificates, but good baseline data on malformations may sometimes be available from special registries such as those kept in the United States by the Center for Disease Control malformation surveillance programmes (DHEW, 1979). It is important to compare malformation data only with data collected from a similar source because of the serious under-reporting of malformations in most record systems.

A level I investigation of a cluster of a particular outcome will also require an attempt to verify the reported cases through medical records, and a search for artefactual causes of an increase, such as changes in referral patterns, the introduction of a new diagnostic technique, or a change in definition of a variable. For example, the increase in congenital hypothyroidism originally reported at the site of the Three Mile Island nuclear accident was shown to be the result of the recent introduction of a newborn screening programme for this disease.

3.4 Level II Investigation: Epidemiological Field Studies

If a level I investigation leads to a clear-cut negative result, proceeding with further investigation may be unwarranted. However, such an investigation often lacks power to detect changes which might still be considered significant by the community, and it might then be considered wise to proceed with more extensive studies which would have greater power to detect an effect. Even if level I studies provided evidence for a positive effect, fuller investigations may be advisable to search for other end-points which might be affected. The decision as to when to

proceed will be made only partly on scientific grounds; the degree of community concern and political considerations also play a large role.

The two basic experimental designs which can be used in an epidemiological investigation of an association between a hazardous exposure and reproductive failure are the *case-control* study and the *cohort* study. In a case-control study the frequency of an exposure is compared among those who experience the outcome in question (cases), and those who do not (controls). In a cohort study the frequency of a reproductive outcome is compared among those who are exposed to the potential hazard and those who are not exposed. A cohort study may be retrospective, i.e. consider those outcomes which have already taken place, or prospective, i.e. study future outcomes in an identified exposed and unexposed group. The advantages and disadvantages of case-control and cohort studies are discussed below with particular reference to studies of reproductive failure.

3.4.1 *Sample Size*

When the end-point to be studied is rare and the exposure relatively common, case-control studies will require a smaller sample size for the same power to detect a given effort than will cohort studies. The opposite is true when the exposure is rare and the end-point is relatively common. In general, cluster situations lend themselves to case-control studies, and exposure situations to cohort studies, but this is not necessarily true.

3.4.2 *Definition of Exposure*

In a case-control study the presence of the exposure may be determined by interview data, by place of work or residence, or by objective measures of exposure such as chemical analysis of body tissues or bioassays such as the incidence of chromosome or sperm abnormalities. The major disadvantage of case-control studies is that recall bias must be suspected whenever the major source of information on exposures comes from interview data. Recall bias refers to the fact that the case responses may be influenced by knowledge of the presence of an adverse reproductive outcome. Cases may over-report exposures, especially if publicity about the suspected exposure has already occurred, or cases may simply remember exposures more accurately than controls. Cases may also under-report exposures if these are associated with feelings of guilt (e.g. alcohol consumption). Validation of exposures by other methods (medical records or biochemical tests) for at least a part of the sample may indicate the extent of the problem. Another approach is to use several control groups, including one which has experienced another adverse outcome not likely to be the result of the same exposure as the one under investigation (e.g. Down's syndrome as a control for neural tube defects). The problem of recall bias will not be present in prospective cohort studies, since the outcome is not known when exposure status is determined. However, prospective studies have the disadvan-

tage of being the most expensive and time consuming to carry out. Also bias may be introduced by the inevitable loss of patients to follow-up.

In a cohort study the presence of the exposure is used to define the comparison groups. This may be done by using groups with varying degrees of exposure, or by using only an unexposed comparison group. The relationship between exposures and reproduction is complex. To be effective an exposure may have to occur before fertilization (perhaps many years) or it may have to occur during gestation; it may act through the male parent or through the female parent, or both. Cohort studies have the disadvantage of requiring decisions about what range of exposure to include in the study before it begins: a case-control study can investigate a wider range of exposure situations.

3.4.3 *Definition of Reproductive Outcomes*

A cohort study will permit an investigation of a wider range of outcomes than a case-control study. Data on outcome may be obtained in a cohort study from interviews, from vital statistics or from hospital and physicians' records. Some data, such as that on infertility and early spontaneous abortions, will often not be available from any source except interviews. When medical records are used as sources of data on outcome, one must be careful to exclude biases caused by different patterns of hospital or physician use among the exposed and unexposed group.

In a retrospective cohort study, recall bias may lead to an overestimate of the effect of exposure, since the presence of the exposure may influence recall of adverse outcomes.

3.4.4 *Selection of Comparison Groups*

In a case-control study the case group and the control group must be chosen to have the same *a priori* probability of exposure. In a cohort study, the exposed and unexposed group must be chosen to have the same *a priori* probability of an adverse outcome. Thus the methods used to determine exposure status and reproductive outcome must be the same for the comparison groups. Also, variables known to be associated with both outcome and exposure must either be matched in the comparison groups or controlled in the analysis by stratification of the sample. Good data are often available concerning the important risk factors for the adverse reproductive outcomes under study, but the data are usually not so good concerning the association of these factors with particular exposures. These associations can be sought in the data, and then controlled in analysis. However, it may be advisable to match comparison groups for known risk factors for a particular outcome even if an association with exposure status is not known. For example, race, maternal age and social class are important risk factors for many reproductive outcomes which might best be controlled in the experimental design.

3.4.5 *Measures of Association*

In case-control studies the measurement of risk is the 'odds-ratio', i.e. the odds that an individual in the case group has been exposed, compared to a control. Thus an odds ratio of 2.1 for heavy drinking with spontaneous abortion means that in the population studied a woman experiencing a spontaneous abortion is 2.1 times more likely to drink heavily than is a woman experiencing a term delivery.

In a cohort study, the measurement of risk is the 'relative risk', i.e. the odds that an individual in the exposed group will experience a particular outcome, compared to an unexposed individual. Thus a relative risk of 2.1 for heavy drinking with spontaneous abortion means that in the population studied a woman who is a heavy drinker is 2.1 times more likely to experience a spontaneous abortion than is a woman who does not drink.

3.5 Level III: Long-Term Follow-Up Investigations

The reproductive effects of chemical hazards and the predictive value of laboratory and clinical observations could best be assessed by following up groups of individuals over their reproductive period. If carried out through continued personal contact with the subjects, such studies involve great amounts of time and effort from the investigators. The best example of such a study in the field of reproduction is the follow-up of the atomic bomb survivors and their offspring in Japan.

Any follow-up study is bound to include a considerable group of patients who are lost to follow-up, and it is important to consider biases which may be introduced if such a loss is not independent of outcome. For example, in a study of women ascertained in early pregnancy, those who have an early spontaneous abortion may not seek medical care of the abortion, and may not return to the prenatal clinic.

Long-term follow-up studies are greatly facilitated if record linking systems exist so that birth and death certificates, disease registries, etc. can be searched for a particular subject. Such facilities are rather poorly developed in the United States, but are much further advanced in Canada and in many European countries (Beebe, 1980). It is important that the essential information for record linking be recorded for subjects who have extensive laboratory studies of exposure, somatic damage or reproductive effects, so that the value of these tests in assessing long-term effects on reproduction can be evaluated in the future.

4 STUDIES OF PARTICULAR REPRODUCTIVE END-POINTS

4.1 Infertility

Infertility is a relatively difficult outcome to investigate epidemiologically. Nevertheless, this has been a reported outcome of 1, 2-dibromo-3-chloro-

propane (DBCP) exposures in men (Whorton *et al.*, 1977) and anaesthetic gases in women (Knill-Jones *et al.*, 1972). In come cases such as the exposure to DBCP the outcome was sufficiently dramatic to be recognized by the population itself.

Fertility may be measured very crudely by the number of births which occur to a given population over a particular period of time. If interview data are available in a level II study, one can use time taken to conceive during a period of unprotected intercourse as a measure of fertility. The first measure does not distinguish between actual failure of conception and loss of conceptions as spontaneous abortions. The second measure avoids this problem, except for those pregnancies which terminate so early as to be unrecognized.

Parental age, race, social class, religion and previous reproductive history are all important determinants of fertility which have to be taken into account in designing or analysing a study using this end-point. A case-control study using infertility as the adverse reproductive outcome might collect cases in an infertility clinic, and compare exposure rates with a control group.

While there are strong reasons for assuming that abnormalities of sperm count, morphology and motility may correlate with male infertility, only azoospermia or oligospermia can at the moment be considered equivalent to infertility in men.

4.2 Spontaneous Abortion

Spontaneous abortion is the most common event signifying reproductive dysfunction. About 15% of recognized pregnancies end in spontaneous abortion, defined here as termination of pregnancy at less than 28 weeks gestation. The incidence of early loss (before pregnancy is recognized by the first missed menses) is only beginning to be explored using the human chorionic gonadotropin (hCG) assays which can diagnose very early pregnancy. A recent study estimated the rate of such early loss to be about the same as that of recognized abortion (Miller *et al.*, 1980).

Data on spontaneous abortions can be obtained from several sources: vital certificate data, hospital records, physicians' records and interview data from the woman or her mate. Vital certificate data are likely to be very poor for pregnancies terminating under 20 weeks, which is the time when reporting of fetal deaths is usually mandated. Hospital and physician records will not contain information on those spontaneous abortions for which medical attention was not sought, which will tend to be those earliest in gestation. The pattern of use of medical services is likely to vary greatly between communities (e.g. rural *vs* urban) and between socioeconomic groups. In 40% of first trimester spontaneous abortions reported in interviews in a New York City study, no medical care was sought (J. Kline, personal communication). The most complete information on spontaneous abortions is likely to come from interviews with the women involved, although this kind of information is of course subject to recall bias, and errors in self-diagnosis. There is some evidence to suggest that maternal reports

are reliable, in that they are generally compatible with data available on the obstetrical record (Kline *et al.*, 1978).

Known risk factors which should be considered in the analysis of spontaneous abortions are maternal age, previous spontaneous abortions, multiple previous induced abortions, social class, maternal smoking and maternal alcohol consumption.

An increase in the rate of spontaneous abortion has been suggested to occur after exposure to anaesthetic gases (Pharaoh *et al.*, 1977), after exposure to lead smelters (Nordström *et al.*, 1979a) and after paternal exposure to vinyl chloride (Infante *et al.*, 1976). None of these studies was rigorously controlled.

Spontaneous abortion will be most informative as an end-point for studies of environmental hazards when it is further classified according to gestation, pathological description and chromosomal status. Data on the latter two features will require special procedures for collection and examination of specimens, and for culturing fetal tissue for chromosomal analysis. About half of all first trimester abortions and about one-third of all spontaneous abortions up to 28 weeks are chromosomally abnormal (Warburton *et al.*, 1980). More than 90% of all chromosomally abnormal conceptions end in spontaneous abortion, so that if one is interested in identifying factors which increase the rate of chromosomal anomalies such as Down's syndrome at term, far fewer pregnancies need to be examined if spontaneous abortions rather than term births are used. In most exposed populations of moderate size there will be no chance of observing a significant increase in chromosomal anomalies at term, since it would require about 23,000 pregnancies to detect a doubling with 80% power. To detect the rise in the incidence of spontaneous abortion which would accompany a doubling in the frequency of chromosome anomalies, only about 900 pregnancies would have to be examined, and only about 200 abortuses karyo typed.

Of the chromosomally normal spontaneous abortions, about half are morphologically abnormal, either grossly so, suggesting a failure in very early development, or with focal abnormalities of the kind seen at term. Classification of specimens may thus provide information about the mechanisms involved in a rise in spontaneous abortion of chromosomally normal fetuses.

4.3 Congenital Malformations

Congenital malformations are much less common than low birth weight, infertility or spontaneous abortion. The usual estimates of the frequency of major malformations in term births is 2–3%. However, this is a reproductive outcome which is perceived by the families involved as very distressing, and it is the most costly in terms of health resources required to care for the children. It is also more easily visible to the population at large than the other end-points, and thus the one most likely to be reported as a cluster.

Since fetal malformations are about five times as common in spontaneous abortions as in live births, they are apparently selected against during gestation. Greater power to detect an increase in malformations would occur if aborted fetuses rather than live births were studied; however, the increased difficulties of ascertainment of material, and the lack of individuals trained in the examination of fetuses may outweigh this advantage.

The rarity of individual malformations usually makes it necessary to classify them in some way into meaningful groups. Classifying by the organ system involved is convenient, but does not generally achieve any aetiological separation. Classification into malformations due to chromosomal imbalance, malformations due to single mutant genes, malformations due to known environmental factors and those of mixed or unknown origin may be the most helpful. This will generally require more detailed information than is available in a level I analysis. Other useful methods of classification might involve grouping the malformations by the developmental stage when the anomaly originates, or by the embryological pathway which is disturbed. Information about the pathogenesis of most human malformation is too limited as yet to make this possible in most cases.

Increased rates of specific malformations have been definitively shown to result from maternal drug ingestions during pregnancy (e.g. thalidomide, anticonvulsants, warfarin), and some studies have suggested increases due to an environmental exposure such as anaesthetic gases or lead (Pharaoh *et al.*, 1977; Nordström *et al.*, 1979a, b). In general one might expect specific patterns of malformations, forming a syndrome, to occur as a result of *maternal exposure during gestation*, while rather diverse effects should occur from *genetic changes* induced by exposure of either parent before conception.

Malformations are generally recorded on birth certificates, in hospital records and, in some areas, in registries. However, there is usually a serious underreporting, at least in the United States. It is therefore important that the comparison groups in an epidemiological study of malformations derive their information by the same method of ascertainment. In a level II study examination of offspring by a trained dysmorphologist, preferably blind to exposure status of the patient, will provide the greatest ability to classify these by aetiology. Diagnostic laboratory studies such as chromosome studies may also be built into such a study. Registries of chromosomal abnormalities, at amniocentesis and in live birth, may provide good information on this class of birth defects in some areas (Hook, 1978; Prescott *et al.*, 1978).

The frequency and type of malformations are related to parental race, age and social class, as well as the sex and age of the offspring at examination. The frequency of detection of some anomalies will increase with the age of the subject (e.g. congenital heart disease) but the frequency of other anomalies which are lethal early in life (e.g. neural tube defects) will decrease. All these factors must be considered in the analysis as possibly confounding variables.

4.4 Birth Weight

Birth weight has many advantages for use as a measure of reproductive outcome. It is routinely recorded on birth certificates and hospital records in most parts of the world, and is thus amenable to level I analysis. When combined with an estimate of gestation, one can distinguish the babies small only because of prematurity from those small for gestational age. About 7% of American infants are less than 2500 grams at birth, and low birth weight is strongly associated with neonatal death rates, congenital malformations, developmental delay and chromosomal anomalies. It is also known to be a sensitive indicator of maternal exposures such as cigarette smoking. An increase in low birth weight infants is the best documented effect found at the Love Canal dump site (Vianna, 1980), and in areas close to lead smelters in Sweden (Nordström *et al.*, 1979a).

Maternal age, parity, and race are other known risk factors, which must be considered as possible confounding variables. Although birth weight may be classified into weight categories, or simply as low, normal and high for gestational age, the analysis will have more power if it is used as a continuous variable.

4.5 Childhood Cancer

Transplacental carcinogenesis, i.e. cancer caused by an environmental agent which crosses the placenta during pregnancy, may be considered an end-point indicating reproductive dysfunction. Childhood cancers are rare, but such effects may not be limited to cancer occurring in childhood. Theoretically any agent that can cause cancer in adults and is known to cross the placenta may cause cancer in the offspring. Since the latent period for carcinogenesis is often long, this may be a very difficult effect to detect, especially if the cancers which are induced are among the most common varieties. To date there are two known transplacental carcinogens in man, diethylstilboesterol (DES) (associated with vaginal cancer, Herbst *et al.*, 1977) and hydantoin (associated with neuroblastoma, Allen *et al.*, 1980). Both these are uncommon effects of exposure, but the rarity of the cancer in the case of DES, and the association with a syndrome of fetal defects known to be the result of a specific environmental agent in the case of hydantoin, enabled the detection of the effect.

Transplacental carcinogenesis can be investigated by case-control studies when a cluster of rare tumours is observed (as with DES). It could also be investigated prospectively in a cohort study when a substantial proportion of the population is exposed, by using identifying information from the children to match the data in tumour registries, or on death certificates.

5 CONCLUSIONS

The epidemiological approach to the study of environmental hazards is not less sophisticated than the laboratory approach. It requires careful planning,

attention to statistical and methodological problems, trained personnel, and a clear understanding of the biology of the situation. Studies of reproductive dysfunction in exposed populations have special problems:

(1) there is a great variety of outcomes which might be affected;
(2) there are three possibly exposed persons in the reproductive unit—mother, father and offspring;
(3) populations involved are likely to be small;
(4) exposure status is difficult to define; and
(5) publicity surrounding the investigation may lead to biased reporting.

In this report we have tried to describe the nature of these problems and indicate ways in which they can be circumvented.

The ideal study would include laboratory methods for refining the information on outcome (e.g. sperm analysis, chromosomal analysis of aborted fetuses) and on exposures (e.g. biochemical tests, bioassays in man or other organisms). However, there is no substitute for the epidemiological approach to validate the predictive value of laboratory tests or results in experimental animals for human reproductive performance.

6 REFERENCES

Allen, R. W., Jr, Ogdon, B., Bentley, F. L. and Jung, A. L. (1980). Fetal hydantoin syndrome, neuroblastoma and hemorrhagic disease in a neonate. *J. Am. med. Ass.* **244**, 1464–1465.

Beebe, G. W. (1980). Record linkage systems—Canada vs. the United States. *Am. J. Publ. Health* **70**, 1246–1248.

Cahen, R. L. (1966). Experimental and clinical chemoteratogenesis. *Adv. Pharmacol.* **4**, 263–349.

Cohen, J. (1977). *Statistical Power Analysis for Behavioural Sciences*. Revised Edition, p. 188. Academic Press, New York, San Francisco, London.

DHEW (1979). *Congenital Malformations Surveillance Annual Summary 1978*. Center for Disease Control, Bureau of Epidemiology. U. S. Department of Health, Education and Welfare, Public Health Service, Atlanta.

Herbst, A. L., Cole, P., Colton, T., Robboy, S. J. and Scully, R. E. (1977). Age-incidence and risk of diethylstilbesterol-related clear cell adenocarcinoma of the vagina and cervix. *Am. J. Obstet. Gynec.* **128**, 43–50.

Hook, E. B. (1978). Monitoring human mutations and consideration of a dilemma posed by an apparent increase in one type of mutation rate. In Chung, C. (Ed.). *Genetic Epidemiology*, pp. 483–528. Academic Press, New York, San Francisco, London.

Infante, P. E., Wagoner, J. K., McMichael, A. J., Wexweiler, R. J. and Falk, H. (1976). Genetic risks of vinyl chloride. *Lancet* **i**, 734–735.

Kline, J., Shrout, P. E., Susser, M., Stein, Z. and Weiss, M. (1978). II. An epidemiological study of the role of gravidity in spontaneous abortion. *Early hum. Dev.* **1**, 345–356.

Knill-Jones, R. R., Newman, B. J. and Spence, A. A. (1972). Anaesthetic practice and pregnancy; controlled survey of women anesthetists in the United Kingdom. *Lancet* **i**, 1326–1328.

Miller, J. I., Williamson, E., Glue, J., Gordon, Y. B., Grudzinskes, J. B., and Sykes, A. (1980). Fetal loss after implantation. A prospective study. *Lancet* **2**, No. 8194, 554–556.

Miller, R. G. (1966). *Simultaneous Statistical Inference*. McGraw-Hill, New York.

Nordström, S., Beckman, L. and Nordenson, I. (1979a). Occupational and environmental risks in and around a smelter in northern Sweden. V. Spontaneous abortion among female employees and decreased birthweight in their offspring. *Hereditas* **90**, 291–296.

Nordström, S., Beckman, L., and Nordenson, I. (1979b) Occupational and environmental risks in and around a smelter in northern Sweden. VI. Congenital malformations. *Hereditas* **90**, 297–302.

Pharaoh, P. D., Alberman, E. and Doyle, P. (1977). Outcome of pregnancy among women working in anaesthetic practice. *Lancet* **i**, 34–36.

Prescott, G., Rivas, M., Shanbeck, L., Macfarlane, D., Wuandt, H., Breg, W., Lubs, H., Magenis, E., Summitt, R., Palmer, C., Hecht, F., Kimberling, W. and Clow, D. (1978). The interregional cytogenetic register system. *Birth Defects Original Article Series* **14(6C)**, 269–279.

Taussig, H. B. (1962). A study of the German outbreak of phocomelia. The thalidomide syndrome. *J. Am. med. Ass.* **180**, 1106–1114.

USEPA (1979). *Report of Assessment of a Field Investigation of Six-Year Spontaneous Abortion Rates in Three Oregon Areas in Relation to Forest 2, 4, 5-T Practices.* Environmental Protection Agency, Environmental Studies Program, Washington, DC.

Vianna, N. (1980). Adverse pregnancy outcomes—potential endpoints of human toxicity in the Love Canal. Preliminary results. In Porter, I. H. and Hook, E. B. (Eds.). *Human Embryonic and Fetal Death*, pp. 165–168. Academic Press, New York, San Francisco, London.

Warburton, D., Stein, Z., Kline, J. and Susser, M. (1980). Chromosome abnormalities in spontaneous abortion: data from the New York City study. In Porter, I. H. and Hook, E. B. (Eds.). *Human Embryonic and Fetal Death*, pp. 261–287. Academic Press, New York, San Francisco, London.

Whorton, D., Krauss, R. M., Marshall, S. and Milby, T. (1977). Infertility in male pesticide workers. *Lancet* **ii**, 1259–1261.