



Analytic Epidemiology and Multivariable Methods

by Deborah Rosenberg, PhD and Arden Handler, DrPH

Multivariable methods are a tool for characterizing the complex relationships among individual, medical and social factors which define the context of maternal and child health. When using these methods, a clear conceptual framework should drive both the study design and the data analysis. Deciding which study type is appropriate, which, if any, multivariable model is appropriate, which variables will be included, the definitions of these variables, and the coding for each, is a prerequisite for meaningful epidemiologic analysis.

RESEARCH DESIGNS IN ANALYTIC EPIDEMIOLOGY

Up to this point, all of the examples of measures of association and hypothesis testing have assumed that complete population data, or data from a random sample are available. Sometimes, however, these types of data are not available, or it is not feasible or efficient to use them in this form. Sometimes, in order to investigate a particular question, more complex research designs are required. Sometimes, although the data you are using are from a random sample or are population based, it is important to impose a design structure as a way to help with decisions about the appropriate measures of association to investigate and report.

The research designs that generate the types of data that can be categorized in 2×2 tables are derived from analytic epidemiology (the study of the determinants and risk of disease) as opposed to descriptive epidemiology, which is the study of the patterns and frequency of disease. These analytic study designs are often called *observational* because they were developed to examine associations between risk factors and outcomes, in contrast to *intervention* studies that explore the associations between interventions and outcomes, and in epidemiology are called experimental studies or clinical trials.

The two designs that are considered the workhorse designs of analytic epidemiology are the *cohort study* (*exposure* → *disease*) and the *case-control study* (*disease* → *exposure*). We will discuss these as well as those designs that are based on data from the entire population (*disease* ↔ *exposure*) and those based on ecologic comparisons. These latter two study types are typically not considered as strong as either the

cohort or case-control designs, but they are commonly used in the public health setting. Our goal is not to provide an in-depth review of epidemiologic study designs but rather to provide some relevant basic points about each design type. We refer the reader to Hennekens and Buring, *Epidemiology in Medicine* and Rothman, *Modern Epidemiology* for more details on these study designs.

Cohort Study

This study design is considered the gold standard when exploring the association between a risk factor or participation in a health program, and health status outcomes or health events. When data are collected on a population that is free of disease, it is possible to follow the cohorts of exposed and unexposed individuals from exposure to outcome, and the incidence of the outcome in both the exposed and unexposed groups can be generated. The ratio of these two risks is the relative risk as described in the previous module.

Cohort Design I: Sampling from a Disease-Free Population Prior to Knowing Exposure Status

		Disease		
		Y	N	
Exposure	Y			?
	N			?
		?	?	N disease-free individuals

This cohort design is based on sampling only individuals who do not have the outcome of interest at the beginning of the study. The number of individuals in the four cells of the table, therefore, is unknown until the end of the study at which time the calculation of disease incidence as well as of exposure prevalence is possible. This study design is, by definition, prospective.

Sometimes, if an exposure of interest is quite rare, sampling is carried out separately among those with and without the exposure in order to insure adequate numbers of exposed individuals.

Cohort Design II: Sampling from a Disease-Free Population According to Exposure Status (n_1 and n_2)

		Disease		
		Y	N	
Exposure	Y			n_1
	N			n_2
		?	?	N disease-free individuals

Although the numbers of exposed and unexposed are fixed, the number of individuals in the four cells of the table is still unknown until the end of the study at which time calculation of disease incidence is possible. The prevalence of exposure, however, cannot be estimated since n_1 and n_2 have been fixed by the sampling design. This design is also, by definition, prospective.

While the cohort study is considered optimal, it is expensive and time consuming to follow a cohort of individuals forward in time to examine differential experiences or outcomes. For instance, to assess the effect of smoking on tubal pregnancy, large numbers of women would need to be followed for multiple years, a task which is usually beyond the resource capabilities of professionals in state and local health agencies.

Case-Control Study

The case-control study is more commonly used than the cohort study because it is less expensive to accumulate cases of an outcome of interest and subsequently gather controls who are similar enough to the cases to allow for a comparison of differential exposure. Case-control studies are mounted when the outcome of interest is very rare, and drawing a random sample of cases from the population would therefore yield too few cases over too long a time period and would involve too much expense. While we often use the terms cases and controls, a true case control study involves a retrospective assessment of differential exposure between the two comparison groups.

The Case Control Design: Sampling from m_1 and m_2

		Disease		
		Yes	No	
Exposure	Yes	a	b	n_1
	No	c	d	n_2
		m_1	m_2	N individuals

In a case control study, sampling is carried out separately among those with and without the outcome in order to insure adequate numbers of diseased individuals. The numbers in all four cells of the table can be filled in at the time of data collection. Here, calculation of disease incidence is not possible since m_1 and m_2 have been fixed by the sampling design. The prevalence of exposure, however, can be estimated.

Because disease incidence cannot be calculated in a case control study, the relative risk cannot be used as a measure of association. Instead, the measure of association between exposure and outcome, the odds ratio (ad/bc), as described in the previous module, is used as an approximation of the relative risk. For instance, if 100 children with asthma and 100 children without asthma were sampled in order to examine the relationship between exposure to cockroaches and occurrence of asthma, the incidence of asthma among the exposed and unexposed could not be generated since the sampling design has created a distorted distribution of individuals such that the overall incidence of disease falsely appears to be 50% .

Sometimes, case-control studies are mounted within a cohort study in which individuals have been followed over time and for whom information has been collected on a variety of outcomes. One can use all of the cases of the outcome of interest found in this cohort and take a sample from the remaining individuals to serve as the controls and then examine the exposure-outcome relationship in the combined group. Because the cases are generated from within a cohort study, this is called a *nested case-control study*.

Data from the Entire Population

In state and local health agencies data are typically available on the entire population, those with and without disease and those with and without exposure. These data are usually thought of as cross-sectional since collection or sampling of the exposure and outcome information is at the same point in time.

Usually, data collected or sampled in this fashion are considered prevalence data and the relationship between an exposure and an outcome can be considered the ratio of two prevalence rates, or relative prevalence.

Cross-sectional Design: Sampling from the Entire Population

		Disease		
		Yes	No	
Exposure	Yes	a	b	n ₁
	No	c	d	n ₂
		m ₁	m ₂	N individuals

In a cross-sectional design, individuals from the entire relevant population are sampled, including those with and without the outcome of interest. The numbers in all four cells of the table can be filled in at the time of data collection and the distribution of individuals is not distorted by the sampling design.

With cross-sectional data, when the outcome is clearly new cases (e.g., low birthweight or asthma deaths), occurrence of the outcome can actually be considered incidence and the ratio of these two incidence rates is the relative risk. Moreover, when information about the timing or biological sequence of events (e.g., smoking and low birthweight) is known, it is possible to make inferences about causality. Sometimes, cross-sectional data collected on a group of individuals over time, are used to retrospectively construct a cohort of exposed and unexposed individuals. This retrospective cohort study is manageable and efficient, but depends upon reliable documentation of both exposure and outcome in the data-set of interest.

Ecologic Designs

We can also generate 2×2 tables from data in which our exposure and outcome cannot be linked at the individual level. For example, the smoking and low birthweight data for 50 counties described in Module 2 is an example of data drawn from an ecologic study. While the association between two data elements that are not linked at the individual level is considered a cause for concern with respect to making assumptions about individuals (*ecologic fallacy*), more current thinking holds that the ecologic level of analysis may be an appropriate and valid tool in its own right. While we usually focus on individual risk factors and outcomes, examination of the relationship between population risk factors or exposure to a program on a population level, and population outcomes is also of value as we develop population based intervention strategies.

Ecologic Design: Data are not Linked at the Individual Level

		Disease %		
		High	Low	
Exposure %	High	a	b	n ₁
	Low	c	d	n ₂
		m ₁	m ₂	N groups

In an ecologic design, the data for exposure and disease are used in aggregated form and cannot be linked at the individual level. The four cells of the table, therefore, do not contain the number of individuals with a given joint exposure-outcome status, but rather contain the number of groups (usually geographic areas or health facilities) with a given joint exposure-outcome status.

DESIGNS FOR PROGRAM EVALUATION

While analytic epidemiologic designs have been the focus of this module, when an MCH professional in a state or local health agency confronts the questions that need to be answered to improve programming on behalf of women, children and families, there are actually a multitude of study designs from which to choose. In particular, one needs to consider whether the designs which are drawn from the social sciences or clinical research, called experimental and quasi-experimental designs, are appropriate, or whether the epidemiologic analytic designs described above are more relevant. To make this decision, one needs to understand the basic distinction between these different types of designs and to examine the appropriateness of the design type for the question at hand. Clearly, we can approach study design from many vantage points but our aim, no matter what our strategy, is to develop a design which is as free from bias as feasible given the constraints of resources and time, as well as ethical and political realities.

Choosing a Study Design

In general, study designs that focus on determining if there is a causal relationship between exposures and outcomes have traditionally been divided into two groups:

1. **Intervention studies:** designs in which the exposure is *manipulated* by the investigator/public health professional. These designs come from a variety of fields including clinical medicine, epidemiology and the social sciences. As mentioned above, intervention studies in the field of epidemiology and clinical medicine are often called clinical trials, while in the social sciences they are often called experimental designs (with random assignment to groups) or quasi-experimental designs (without random assignment to groups).
2. **Observational studies:** designs in which the exposure *is not manipulated* by the investigator/public health professional. These designs come from the field of *analytic epidemiology* and were initially developed for situations in which the exposures of interest are observed phenomena, as opposed to interventions or programs. The major designs in this category are the cohort and case-control study designs described above.

Historically, observational designs were reserved exclusively for examining risk factors and outcomes. However, as epidemiologists have become more active in health services research and in conducting program evaluations, they have begun to use observational designs to examine the relationship between participation in programs and health status outcomes. When evaluating programs, however, there has been de facto *manipulation* of the exposure, not just *observation* of the exposure. The manipulation may have been at the policy level, not directly controlled by the evaluator, but manipulation has occurred nonetheless. Therefore, manipulation of the exposure no longer defines the difference between intervention and observational studies. When the exposure is a health program or policy, then, what influences whether the study design is drawn from the realm of observational studies (cohort and case-control studies) or from the realm of intervention studies (quasi-experimental and experimental designs)?

When the exposure is a health program or policy, the choice of using observational designs instead of quasi-experimental/experimental designs is dependent on whether there is the theoretical possibility of collecting pretest or baseline data from the same units (e.g., individuals, clinics, provider practices) on which there is an outcome measurement:

- In observational studies there are two phenomena of interest, an exposure which can be a program or policy, and an outcome, such as health status or use of health services.

- In experimental and quasi-experimental designs there are also two phenomena of interest, an exposure and an outcome, but the outcome *can be measured at two points* in time, pretest/baseline data (whether one collects it or not) and posttest data (whether one collects it or not.)

In program evaluation, experimental designs and quasi-experimental designs are usually used when there is outcome information at baseline about one or more groups, an intervention takes place, and outcome information can also be collected on the same group or groups subsequent to the intervention. The research question can be stated as, "does the intervention *change* the outcome status of the individuals being studied?" Observational designs are used in program evaluation when an intervention takes place, outcome information is subsequently collected, but there was no theoretical possibility to collect outcome information at baseline on the same units or subjects because the outcome can only occur following the exposure (i.e., many health status outcomes). Now, the research question is stated as, "does the intervention have an impact on *whether* the outcome occurs?"

Observational designs can be thought of as analogous to post-test-only designs in the experimental/quasi-experimental design framework. Posttest only designs, however, particularly those without random assignment, are considered some of the weaker designs of this paradigm since they cannot control for differences in outcome status prior to the intervention. When a health outcome can occur only once, however, in effect there are no baseline differences to account for; all individuals in the study have the same status prior to the intervention. In this context, concern about the weakness of the posttest only, or observational design, is unwarranted.

Comment

Experimental designs involve a program or a treatment, an outcome measure, and a comparison from which change due to the program or treatment can be inferred. Experimental designs involve random assignment to a "treatment" or control group.

Random assignment is the use of a process of random selection as to who receives the treatment and who does not receive the treatment under study. Random assignment creates theoretically equivalent comparison groups.

Quasi-experimental designs involve a program or treatment and an outcome measure but the comparison groups from which change is to be inferred are not created through random assignment. The comparison groups are nonequivalent so attributing cause to the program or treatment is more difficult than in an experimental design.

For more information on experimental designs refer to Campbell and Stanley's *Experimental and Quasi-Experimental Designs for Research on Teaching*; for more information on quasi-experimental designs refer to Cook and Campbell's *Quasi-Experimentation: Design and Analysis Issues for Field Settings*.

Example One. WIC participants in one county are participating in a variety of interventions to change behaviors; two of interest to a local health agency are a household budgeting workshop and a breastfeeding workshop. The local health agency is interested in the effect of these workshops on client behavior.

To study the effect of a household budgeting workshop on the purchasing behavior of WIC participants, the program evaluator is likely to choose an experimental or quasi-experimental design because the

collection of pretest data (purchasing behavior) on the same subjects before and after the workshop is theoretically possible.

However, if the evaluator is interested in the impact of a WIC breastfeeding workshop provided during pregnancy on the breastfeeding behavior of first-time mothers, then the evaluator is likely to choose an observational study design. Since the population has yet to experience the behavior in the particular episode under study, there is no theoretical possibility of administering a pretest to the actual participants in the intervention. In other words, primiparous mothers do not have prior breastfeeding behavior, so it is not possible to obtain baseline data. Instead, an epidemiologic cohort study design, in which those exposed to the workshop and those not exposed to the workshop are followed to observe breastfeeding behavior, can be employed.

In both of the above scenarios, the exposure has been manipulated. However, in the first scenario there is the theoretical possibility of collecting pretest data on the individuals in the workshop and therefore an experimental or quasi-experimental design is selected. In the second scenario there is no possibility of collecting pretest data on these individuals so an observational design is chosen.

Example Two. The state health agency is interested in evaluating the impact of a teenage pregnancy prevention program on the *knowledge, attitudes and sexual practices* of not previously pregnant high-risk adolescent females engaged in sexual activity which places them at risk of STDS/HIV and/or pregnancy. There is also an interest in exploring the impact of this program on pregnancy rates. The program has been implemented in several counties and is limited to 50 female teens per site; those who are eligible for the program but are unable to participate due to space limitations are placed on a waiting list (there is no random assignment). In order to examine the impact of the program on participants' knowledge, attitudes and sexual behavior, information on these three domains can be collected before the teens begin the program and then again at one or more points after their participation in the program is ended. Information on pregnancy can be obtained through follow-up for several years after participation in the program is concluded.

If the program manager/evaluator chooses to collect pre-test and post-test information on knowledge, attitudes and behavior from program participants and the individuals on the waiting list, the design would be quasi-experimental. When considering the relationship between exposure to the intervention and the occurrence of pregnancy, on the other hand, the evaluator is likely to conduct an epidemiologic cohort study because the event of pregnancy can only occur after (during) the intervention.

Use of Historical or Community Controls

For many program evaluations in which the outcomes are such that baseline or pre-test data on the outcome of interest cannot be collected on the unique individuals who have participated in the intervention, the evaluator can often obtain an indirect pretest measure from historical or community controls. For example, in the study described above of breastfeeding behavior, historical data on the breastfeeding behavior of first time mothers in the WIC program might be compared to the breastfeeding behavior of workshop participants and non-participants.

Likewise, in the study described above of pregnancy prevention among high risk teens the evaluator might also examine the pregnancy rates of high risk teens in the communities in which the intervention took place before its implementation and compare these to the rates for the teen participants and non-participants in the program. In both of these cases, the evaluator is establishing a baseline from community or population data rather than from the identified participants or non-participants in the program.

Summary

In sum, when the goal is to answer questions about programs, the more useful distinction between observational studies and quasi-experimental and experimental designs is not manipulation of exposure but rather whether there is the theoretical possibility of obtaining pre-test or baseline measures on *the same units or subjects* for which there is outcome measurement. Selecting the type of study design from one paradigm or another is dependent on the question being asked, the kinds of data available, whether baseline information is theoretically possible, and the resources that can be brought to bear on the task.

While naming the study design is less important than developing a design that can answer the questions of interest, choosing nomenclature helps one to think through the appropriate measures of association, the relevant statistical tests, and the biases that may affect whether it will be possible to establish that an association between a program or policy and an outcome is in fact genuine.

Test Yourself

Question:

Which study design would you choose for the following? Which is the exposure and which is the outcome?

1. The state health agency is exploring a variety of strategies to increase immunization rates of pediatric providers. Pediatric providers are randomly assigned to one of three interventions. The outcome of interest is immunization rates of two-year-olds.
2. The state health agency is examining the impact of participation in primary care case-management for children with special health care needs on reduction in the use of hospital emergency rooms for primary care visits. Children are assigned to primary care case-management on a first-come first served basis; those on the waiting list constitute the comparison group.
3. The state health agency has designed an asthma prevention program for a cohort of low birthweight infants who had respiratory problems at birth; only LBW infants at three hospitals in the state have been selected for participation. The goal is to follow those exposed to the intervention and those not exposed to the intervention to determine the incidence rate of asthma through age five.

Answer:

1. The evaluator would logically choose an experimental design. Providers are the units that are randomly assigned to treatment, and immunization rates for two-year-olds in each provider's practice can be measured before and after the implementation of the intervention.
2. The evaluator would logically choose a quasi-experimental design as emergency room use before and after the intervention can be examined, and a non-equivalent comparison group is available.
3. The evaluator would logically choose a cohort study design following the exposed and unexposed infants to examine the difference in incidence rates of asthma between the two groups. There is no possibility of a pre-test or baseline measure for this group of infants.

ADJUSTMENT FOR CONFOUNDING

How do we know if an association between a risk factor / program / system change and a health event / program / system outcome is real?

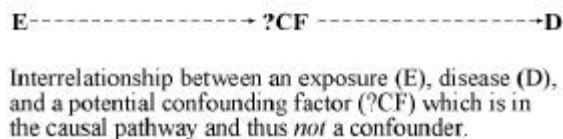
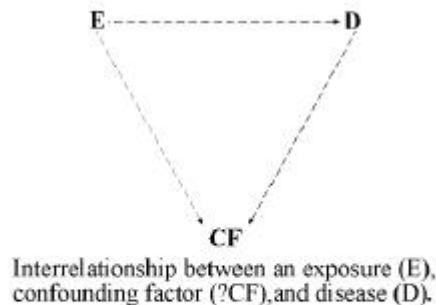
We have described the use of statistical testing to evaluate whether an association is due to chance or reflects a “true” relationship. While we may discover statistically significant results, statistical significance is only one part of the picture. The relationship between the two variables may actually be due to bias in the design of the study or to confounding of the relationship between the two variables by a third variable. We have discussed the two major types of bias, information bias and selection bias, previously. To discover bias that might distort the relationship between an exposure and outcome, it is essential to use critical thinking skills to judge whether differences between groups are the result of disparities in the way the groups were selected or are the result of differential ascertainment of information on both exposures and outcomes for the groups.

Sometimes even when the study design is free of bias, an apparent association may not be “real” but may instead be due to the mixing of a third factor with the exposure and the outcome. For example, if one finds a relationship between participation in the WIC program and increased birthweight of infants, it is necessary to consider whether a third factor such as age of the WIC participants, parity of the WIC participants, or health behaviors of the WIC participants may explain the relationship between WIC and increased infant birthweight. Confounding, unlike bias, cannot be eliminated by improving the study design; confounding is a reflection of the complex interrelationship between variables in the real world.

To be eligible as a confounder, a third factor or variable must meet the following criteria:

1. It must be a risk factor for the outcome, independent of its association with the risk factor or exposure under study. It does not have to be causally associated with the outcome, but simply correlated with it.
2. It must be correlated with the risk factor or exposure under study, independent of its association with the outcome.
3. It must not be in the hypothesized causal pathway *between* the risk factor and outcome.

EVALUATING THE ROLE OF CONFOUNDING



From *Epidemiology in Medicine*, Charles Hennekens and Julie Buring. Copyright © 1988 by Little, Brown & Company. Used by permission of Little, Brown & Company.

How do we assess whether a third factor is a confounder of the association of interest?

When evaluating confounding, person variables such as ethnicity/race, age, parity, education, health insurance status and socioeconomic status are typically the "usual suspects". It is also desirable to consider other variables as potential confounders as well; this decision should be based on both what is available in the data-set at hand and what is known about the relationship under consideration.

Steps in Evaluating Confounding Based on Analysis of 2×2 Tables

The following steps are those typically used when your measure of association is the odds ratio or relative risk. The aim is to determine if the crude OR or RR is different than the odds ratio or relative risk obtained after adjustment for a third variable. Determining whether a difference between the crude and adjusted measures is meaningful is a matter of judgement since there is no formal test for the presence of confounding.

First, explore whether the third factor is related to the two variables of interest. If it is not related to both variables, it cannot be a confounder.

Second, if the third factor is related to the two variables of interest, conduct stratified analysis. To conduct stratified analysis using 2×2 tables, the exposure and the outcome variables must be dichotomous, and the potential confounder must be categorical although not necessarily dichotomous. In stratified analysis, one examines several two by two tables; each table shows the relationship between the two variables of interest at each level of the third variable.

Third, compare the crude odds ratio or crude relative risk to the weighted average of the stratum specific odds ratios or relative risks generated from stratified analysis. These stratum specific measures are unconfounded estimates of the relationship between the two variables at each level of the third variable. If the weighted odds ratio or relative risk is different than the crude (unstratified) estimate of association, then confounding is present.

The formulas for the weighted (adjusted) estimates are extensions of the formulas for the crude relative risk and odds ratio. The numerator of the adjusted relative risk is a weighted sum of the numerators of the relative risk in each stratum; the denominator is a weighted sum of the denominators of the relative risk in each stratum. Likewise, the numerator of the adjusted odds ratio is a weighted sum of the numerators of the odds ratio in each stratum; the denominator is a weighted sum of the denominators of the odds ratio in each stratum. In both measures, the weights are based on the total sample size.

Rothman - Boice Summary Relative Risk :

$$= \frac{\sum_{i=1}^{\# \text{ strata}} \frac{a_i n_{2i}}{N_i}}{\sum_{i=1}^{\# \text{ strata}} \frac{c_i n_{1i}}{N_i}}$$

Mantel - Haenszel Summary Odds Ratio :

$$= \frac{\sum_{i=1}^{\# \text{ strata}} \frac{a_i d_i}{N_i}}{\sum_{i=1}^{\# \text{ strata}} \frac{b_i c_i}{N_i}}$$

The confidence limits for the adjusted relative risk and odds ratio are likewise extensions of the formulas for their unadjusted counterparts. They are based on a weighted sum of the standard errors of the stratum specific estimates. Statistical and epidemiologic software will calculate the appropriate confidence intervals for adjusted measures.

Evaluating Effect Modification Based on Analysis of 2×2 Tables

Stratified analysis can also be used to identify whether a third factor is an effect modifier of the relationship between a risk factor and an outcome. While a confounder uniformly (or close to uniformly) changes the association between the exposure and outcome across all strata of the third variable, an effect modifier differentially changes the association between the exposure and outcome across the strata. Also called interaction, effect modification is present if the relationship between the risk factor and the outcome is different, either in direction or magnitude, at each level of the third factor. The focus when assessing effect modification, then, is on comparing the relative risks or odds ratios from stratum to stratum rather than on comparing the crude and the weighted summary estimate which is appropriate for assessing confounding.

A test statistic, the *Breslow-Day Test for Homogeneity of the Odds Ratio* can be used to statistically determine if effect modification is present. However, it is best to use this test in conjunction with a non-statistical assessment of the stratum specific estimates incorporating what is known about the sample size in each stratum. If effect modification is present, the weighted summary measure should not be used; the stratum specific estimates should be used instead.

In the presence of confounding an adjusted measure appropriately controls for a third factor and it is unnecessary to focus on the third factor itself. In contrast, in the presence of effect modification, the third factor is critical to understanding the differential impact that the risk factor, program, or systems measure is having on the outcome of interest.

Suppose the following 2×2 table illustrates the crude association between a hypothetical "exposure" and outcome. The crude relative risk is 1.67, indicating a moderate association.

Crude Association

		Outcome		
		Y	N	
E X P	Y	500	4,000	4,500
	N	700	9,800	10,500
		1,200	13,800	15,000

$$\text{Crude RR} = \frac{11.11}{6.67} = 1.67$$

Next, to investigate possible confounding or effect modification, the data are stratified according to levels of a third factor. The association between the hypothetical exposure and outcome is presented below under three conditions: no confounding, confounding, and effect modification. In all three sets of 2 × 2 tables, the strata can be re-combined to yield the crude table as shown above. For this example, the third factor has two levels resulting in two strata, although a third factor may have any number of levels resulting in the equivalent number of strata.

No Confounding or Effect Modification

		Stratum 1 Outcome		
		Y	N	
E X P	Y	250	1,250	1,500
	N	350	3,150	3,500
		600	4,400	5,000

		Stratum 2 Outcome		
		Y	N	
E X P	Y	250	2,750	3,000
	N	350	6,650	7,000
		600	9,400	10,000

$$RR_1 = \frac{16.67}{10.00} = 1.67$$

$$RR_2 = \frac{8.33}{5.00} = 1.67$$

$$\text{Adj. RR} = 1.67$$

Confounding

		Stratum 1 Outcome		
		Y	N	
E X P	Y	350	2,150	2,500
	N	250	2,250	2,500
		600	4,400	5,000

		Stratum 2 Outcome		
		Y	N	
E X P	Y	150	1,850	2,000
	N	450	7,550	8,000
		600	9,400	10,000

$$RR_1 = \frac{14.0}{10.0} = 1.40$$

$$RR_2 = \frac{7.50}{5.63} = 1.33$$

$$\text{Adj. RR} = 1.37$$

Effect Modification

		Stratum 1 Outcome		
		Y	N	
E X P	Y	365	2,135	2,500
	N	175	2,325	2,500
		540	4,460	5,000

		Stratum 2 Outcome		
		Y	N	
E X P	Y	135	1,865	2,000
	N	525	7,475	8,000
		660	9,340	10,000

$$RR_1 = \frac{14.6}{7.00} = 2.09$$

$$RR_2 = \frac{6.75}{6.56} = 1.03$$

In the leftmost set of tables, there is no confounding by the third factor since the adjusted relative risk of 1.67 is exactly the same as the crude relative risk seen above (1.67). In the middle set of tables, the third factor would typically be considered a confounder since the adjusted relative risk of 1.37 is quite different from the crude relative risk, and this average mirrors fairly well the relative risks in each stratum. In the rightmost set of tables, the third factor would typically be considered an effect modifier since the relative risks of 2.1 and 1.03 in each stratum are very different from each other, and in fact the Breslow-Day Test for Homogeneity of the Odds Ratios shows them to be significantly different from each other ($p=0.001$). Although an adjusted relative risk could also be calculated here, resulting in a value of 1.51, this average does not mirror the relative risks in each stratum and should not be used.

Notice that lack of confounding does not necessarily mean that the two strata are identical. Looking again at the leftmost set of tables, the outcome rates of 16.67 and 10.00 among the exposed and unexposed respectively in the first stratum are each higher than the outcome rates of 8.33 and 5.00 among the analogous groups in the second stratum, even though they result in the same relative risks. This is because the third factor is related to the outcome, with overall rates of 12% (600/5,000) and 6% (600/10,000) in the two strata, compared to the crude rate of 8% (1,200/15,000). No confounding exists, however, because the third factor is not related to the exposure since 30% of the individuals in each stratum are exposed (1,500/5,000 and 3,000/10,000), exactly the same percentage as in the crude table (4,500/15,000). Remember that confounding does not exist if a third factor is unrelated to both exposure and outcome, or if it is related to one or the other, but not to both.

When confounding does exist, it may be because the third factor is positively correlated with both the exposure and outcome, because it is negatively correlated with both the exposure and outcome, or because it is positively correlated with either the exposure or outcome and negatively correlated with the other. Looking again at the middle set of tables, in addition to the third factor being related to the outcome with the overall rates of 12% and 6% in the two strata, it is now also related to the exposure with 50% (2,500/5,000) and 20% (2,000/10,000) being exposed in the two strata compared with the 30% (4,500/15,000) in the crude table. Here, the third factor is positively correlated with both exposure and outcome; both the prevalence of exposure and the incidence of the outcome are higher in the first stratum.

Test Yourself

Question:

Using the following information, determine whether there is confounding or effect modification present. Assume a narrow confidence band around each relative risk.

	Crude	Adjusted	Stratum 1	Stratum 2	Confounding?	Effect Modification?
1.	1.8	1.8	1.7	1.9		
2.	1.8	1.5	0.9	2.5		
3.	1.8	1.4	1.4	1.5		
4.	1.8	1.8	3.2	1.0		
5.	0.5	0.6	0.5	0.7		

Answer:

	Crude	Adjusted	Stratum 1	Stratum 2	Confounding?	Effect Modification?
1.	1.8	1.8	1.7	1.9	N	N
2.	1.8	1.5	0.9	2.5	NA	Y
3.	1.8	1.4	1.4	1.5	Y	N
4.	1.8	1.8	3.2	1.0	NA	Y
5.	0.5	0.6	0.5	0.6	Y	N

Note that once the presence of effect modification is identified, as in #2 and #4 above, the concept of confounding is not applicable. Also, if confidence limits had not been assumed to be narrow, but instead were wide due to small sample size, the apparent effect modification in #2 and #4 might not be real. The stability of estimates, in this case relative risks, must be considered when assessing both potential confounding and effect modification.

In addition, care must be taken when assessing differences in estimates with values < 1 . For example, the crude and adjusted estimates of 0.5 and 0.6 in #5 do not appear very different, but remember that the relative risk can range from 0 to infinity and that half of all possible values are < 1 . If the values of 0.5 and 0.6 are inverted ($1/0.5=2.0$ and $1/0.6=1.67$), then the difference between them is more readily seen.

Typically, in the analysis of the relationship between an exposure and an outcome, multiple confounders are considered. The analyst is encouraged to examine the relationship of each potential confounder with the association under interest separately before proceeding to consider the joint effects of many potential confounders using multivariable methods. Once multiple confounders or effect modifiers are identified, multivariable methods can be used to obtain the best estimate of the association between the exposure and the outcome of interest. Stratified analysis may be used to simultaneously adjust for two or more factors, but very quickly this type of analysis becomes inefficient. Regression approaches are better suited for this situation and will be discussed later in this module.

CONFOUNDING AND STANDARDIZATION OF RATES

In the previous section, we discussed the concepts of confounding and effect modification. In analytic epidemiology, when the purpose is to investigate a potentially etiologic, or causal, association between a risk factor and an outcome, "controlling for" confounding is an essential step in obtaining an unbiased estimate of the strength of the hypothesized relationship. In a public health context, when the purpose is to compare health status in different populations—when membership in a population is the "exposure"—standardization of rates is often carried out in an effort to insure that the comparison is made accounting for differences in fundamental, structural characteristics. In either case, the goal is to account for any mixing of a third factor (or multiple other factors) with the primary association of interest.

Several terms, then, are used to describe methods that address potential confounding. Each term is typically applied in a particular analytic context, but the goal of accounting for confounding factors is the same.

"Standardization"
"Adjustment"
"Controlling for"
"Stratified Analysis"

As we've already seen, the process of accounting for confounding involves separating the data into a series of strata and then applying a method that yields a summary or average estimate, weighted by the distribution of observations across the strata.

Standardization of rates has been most commonly used in epidemiology when comparing the mortality experience across populations. Typically, the comparisons are between large populations, often between nations, or sometimes between states. The objective is to adjust for societal level features that fundamentally distinguish the populations being compared. The age structure of a population, for

instance, is a marker of social, economic, and political development in the broadest sense. Following is a series of age categories that might be used to produce an age-standardized mortality rate:

- < 1
- 1-4
- 5-14
- 15-24
- 25-44
- 45-64
- 65-84
- > 85

Since chronological age is positively correlated with most chronic diseases and since populations often have differing age structures, age meets the definition of a confounder and age standardized (or age-adjusted) rates are generated in order to make comparisons. For example, rates of coronary heart disease are known to increase with age. If a comparison is to be made between two geographic areas, one with 12% of its population over the age of 65, the other with only 6% of its population over the age of 65, higher death rates from CHD would be expected in the first population on the basis of this age difference alone. From a public health perspective, it is much more relevant to know if there is a difference in death rates in the two areas due to factors beyond the aging process itself—factors that may be amenable to public health interventions.

What typically distinguishes standardization of rates from other stratified methods for control of confounding is its use of an external standard. In other words, in addition to having data on the populations of interest, data from another population is used as a common benchmark. For example, the World Population, or the U.S. population from the 1940 Census, or from the 1970 Census might be used to compare U.S. and Canadian mortality rates in 1990. Sometimes an external standard may include the populations being compared, as when state rates are standardized according to national data.

In order to illustrate the process of rate standardization with an external standard, let's consider the following example comparing neonatal mortality rates in two hypothetical hospital groups in a region: one group is comprised of tertiary care hospitals with neonatal intensive care units (NICUs), the other is comprised of community hospitals which appropriately transfer the majority of high risk pregnant women for delivery at a tertiary center. The birthweight distribution in the two groups is considered a potential confounder since the tertiary care hospitals by definition serve higher risk pregnant women than do the community hospitals and birthweight is also known to be the major predictor of neonatal mortality. All live births in the state will be used as the external standard.

The data are as follows:

1. Hospital Group A: Community Hospitals

Birthweight Strata	Deaths	Births	% of Total	Stratum Specific Rate per 1000	Crude Rate per 1000
< 1500	53	150	1	353.3	
1500 – 2499	12	750	5	16.0	
>= 2500	28	14,100	94	2.0	
	93	15,000	100		6.2

2. Hospital Group B: Hospitals with NICUs

Birthweight Strata	Deaths	Births	% of Total	Stratum Specific Rate per 1000	Crude Rate per 1000
< 1500	65	300	2	216.7	
1500 – 2499	14	1,200	8	11.7	
>= 2500	14	13,500	90	1.0	
	93	15,000	100		6.2

External Standard: All Live Births in the State

Birthweight Strata	Deaths	Births	% of Total	Stratum Specific Rate per 1000	Crude Rate per 1000
< 1500	1,375	5,000	1	275.0	
1500 – 2499	490	35,000	7	14.0	
>= 2500	460	460,000	92	1.0	
	2,325	500,000	100		4.65

The crude relative risk of neonatal death (not accounting for the birthweight distribution) when the two hospital groups are compared is 1 (6.2/6.2). Standardizing by birthweight will help determine if this relationship is a fair reflection of the neonatal mortality experience in the two hospital groups.

Standardization can be accomplished in two different ways:

Direct standardization applies the stratum specific rates of each population to the number of individuals in the corresponding stratum in the standard population. This method yields an adjusted relative risk. The method is called "direct" because it uses the actual morbidity or mortality rates of the populations being compared.

Indirect standardization, on the other hand, applies the stratum specific rates of the standard population to the number of individuals in the corresponding stratum in each of the populations being compared. This method is called "indirect" because nowhere does it use the actual morbidity or mortality rates of the populations being compared. Instead of an adjusted relative risk, indirect standardization yields standardized morbidity or mortality ratios (SMRs), one for each population being compared. Direct estimates are preferable to indirect ones, but the indirect method is used when the rates from the populations being compared are based on small numbers and therefore considered unreliable.

Direct Standardization

Using the numbers from the entire state and the rates for the two hospital groups shown in the hypothetical data above, a directly standardized relative risk is calculated as follows:

Adjusted Rate for the Community Hospitals :

$$\begin{aligned} &= \frac{5,000 \times 353.3 + 35,000 \times 16.0 + 460,000 \times 2.0}{500,000} \\ &= 6.5 \end{aligned}$$

Adjusted Rate for Tertiary Care Hospitals :

$$\begin{aligned} &= \frac{5,000 \times 216.7 + 35,000 \times 11.7 + 460,000 \times 1.0}{500,000} \\ &= 3.9 \end{aligned}$$

$$\text{Crude Relative Risk} = \frac{6.2}{6.2} = 1.0$$

$$\text{Standardized Relative Risk} = \frac{6.5}{3.9} = 1.7$$

In the process of calculating an adjusted relative risk, you can see that it is necessary to calculate what appear to be adjusted rates for each population. These recalculated rates, however, reflect what would be expected in the standard population if it had the morbidity or mortality experience of the populations being compared. In this example, the adjusted rates of 6.5 and 3.9, therefore, do not reflect the real mortality risk in the two hospital groups. They are byproducts of the standardization procedure and should not be used as stand-alone measures. In practice, these rates are sometimes reported despite the fact that they can lead to misleading and inappropriate conclusions.

The adjusted relative risk of 1.7 shows that the community hospitals have elevated neonatal mortality compared to the tertiary care centers even though the unadjusted relative risk was 1. Without adjustment, the better survival of neonates born in the tertiary care centers was masked due to the disparity in the birthweight distribution of the infants served by the two hospital groups; While the tertiary care centers have a higher incidence of low birthweight births than the community hospitals (10% v. 6%), they have a lower neonatal mortality rate within each birthweight stratum (216.7 v. 353.0, 11.7 v. 16.0, and 1.0 v. 2.0).

Indirect Standardization

Using the rates for the entire state and the numbers for the two hospital groups shown in the hypothetical data above, two standardized mortality ratios are calculated as follows:

For the Community Hospitals

$$\begin{aligned} \text{SMR} &= \frac{\frac{93}{15,000} \times 1,000}{\frac{150 \times 275 + 750 \times 14.0 + 14,100 \times 1.0}{15,000}} \\ &= \frac{6.2}{4.4} \\ &= 1.4 \end{aligned}$$

For the Tertiary Care Centers

$$\begin{aligned} \text{SMR} &= \frac{\frac{93}{15,000} \times 1,000}{\frac{300 \times 275 + 1,200 \times 14.0 + 13,500 \times 1.0}{15,000}} \\ &= \frac{6.2}{7.5} \\ &= 0.8 \end{aligned}$$

With indirect standardization, each SMR is itself an adjusted relative risk; the numerator is the observed crude rate in a population and the denominator is its expected rate given the neonatal mortality experience in the standard population. An $\text{SMR} > 1$ indicates higher rates than expected and an $\text{SMR} < 1$ indicates lower rates than expected. The two SMRs in this example lead to the same conclusion as did the adjusted relative risk—after accounting for birthweight the community hospitals have higher neonatal mortality than do the tertiary care centers.

It is not technically correct to create a ratio of two SMRs and interpret it as a relative risk; each SMR is itself a relative risk, and as such its value is compared to an expected value of 1. In practice, however, SMRs are sometimes compared in this fashion. In this example, the ratio of the two SMRs is $1.4/0.8$ or 1.75, very similar to the value of 1.7 for the adjusted relative risk obtained from direct standardization.

Test Yourself

Question:

The following table includes data from a hypothetical state survey that asked parents to report whether their children have a medical home. In addition, the table includes data from a county in the same state. The data are stratified by income level. Use indirect standardization to calculate an SMR that compares the county to the state with respect to the percent of children who have a medical home. Interpret the SMR.

Income Strata	# of Children with a Medical Home in the County	# of Children in the County	% of Children in the State Survey with a Medical Home
< \$10,000	780	1,200	70.0
\$10,000-\$30,000	2,530	4,600	45.0
\$30,000-\$50,000	2,170	3,100	60.0
>\$50,000	990	1,100	80.0
	6,470	10,000	

Answer:

$$\begin{aligned}
 \text{SMR} &= \frac{\frac{6,470}{10,000} \times 100}{\frac{780 \times 70 + 2,530 \times 45 + 2,170 \times 60 + 990 \times 80}{10,000}} \\
 &= \frac{64.7}{56.5} \\
 &= 1.15
 \end{aligned}$$

In this case, having a medical home is a positive, rather than an adverse outcome and therefore an SMR > 1 indicates that the children in this particular county are somewhat more likely to have a medical home than are children statewide.

Now that the process of direct and indirect standardization has been illustrated using a particular external standard, it is important to recognize that equivalent results would be obtained with some other external standard or even with an internal standard. For example, a national hospital data-set could have been used as an external standard, or one of the two hospital groups or the two hospital groups combined could have been used as an internal standard. While the choice of different standards will impact intermediate calculations, the adjusted relative risks or SMRs will all lead to the same interpretation. In fact, the adjusted relative risks from direct standardization will be the same regardless of the standard used.

The advantage of using a common external standard, either with direct or indirect standardization, is that many geographic areas and many time periods can be compared. For example, if the neonatal mortality rates of hospital groups in another region had been compared by standardizing for birthweight with the same statewide data as was used above, it would also be possible to compare the community hospitals across regions or the tertiary care hospitals across regions. If internal standards were used, the results for the two regions would not be immediately comparable.

Interestingly, stratified analysis as described in our earlier discussion of confounding and effect modification is equivalent to direct standardization with all of the observed data combined being the internal standard. For instance, following are the same hospital group data organized into a set of 2×2 tables for stratified analysis. There is one table for each birthweight stratum, the "exposure" is delivery in a community hospital v. delivery in a tertiary care center, and the outcome is neonatal death v. neonatal survivor.

Birthweight < 1,500 grams

		Neonatal Death		
		Y	N	
H O S P	A	53	97	150
	B	65	235	300
		118	332	450

Birthweight 1,500-2,499 grams

		Neonatal Death		
		Y	N	
H O S P	A	12	738	750
	B	14	1,186	1,200
		26	1,924	1,950

Birthweight \geq 2,500 grams

		Neonatal Death		
		Y	N	
H O S P	A	28	14,072	14,100
	B	14	13,486	13,500
		42	27,558	27,600

$$RR_1 = \frac{353.3}{216.7} = 1.6$$

$$RR_2 = \frac{16.0}{11.7} = 1.4$$

$$RR_3 = \frac{2.0}{1.0} = 2.0$$

$$\text{Adj. RR} = 1.7$$

Notice that the estimate of the adjusted relative risk (1.7) is the same as that obtained from direct standardization using an external standard, underscoring the correspondence between the two methodologies. With the data organized in this fashion, however, the difference in the stratum specific estimates can also be seen. The process of rate standardization always assumes that a summary measure is appropriate to report, and while it is debatable as to whether effect modification exists in this example, the organization of the data for stratified analysis encourages closer examination of the stratum specific estimates. In this case, it is interesting to see that, the relative risk of neonatal death at the community hospitals compared to the tertiary care centers is actually highest in the normal birthweight category ($2.0/1.0=2$). While one would expect the community hospitals not to perform as well with high risk infants since they are not equipped to manage them, it is disturbing that their performance is also worse with low risk infants.

Too often, stratum specific information is ignored in favor of the adjusted summary measure. Sometimes this is deemed necessary if, for example, many indicators are being examined, and reporting all of the stratum specific rates may be providing an audience more information than can be easily absorbed. In this circumstance, reporting one summary measure for each indicator may do a better job of communicating a coherent picture of the health status of a population. On the other hand, when the public health focus is on more effective and efficient targeting of interventions and on prioritizing allocation of resources, stratum specific information may in fact be more useful than summary measures.

Moreover, it is important not to confuse the reporting of relative comparisons across populations with examining the true level of an outcome in each population. There is a temptation to try and do both simultaneously by reporting the adjusted rates obtained from the standardization process. A series of such rates implies comparisons across populations, but it also implies that each rate reflects the true level of the

outcome in a population which, as was pointed out earlier, it does not. It is better to report an adjusted relative risk or SMR for comparison purposes and in addition report the actual observed stratum specific rates to give a sense of the occurrence of the outcome in each population. In the hospital group example, the adjusted relative risk of 1.7 could be reported along with the stratum specific neonatal mortality rates of 353.3, 16.0, and 2 per 1,000 live births for the community hospitals and then 216.7, 11.7, and 1.0 per 1,000 live births for the tertiary care centers.

Finally, remember that adjustment for confounding is not appropriate when the third factor is in the causal pathway. If the association of interest were smoking status and neonatal mortality instead of hospital group and neonatal mortality, it would be inappropriate to standardize for or stratify by birthweight since birthweight is in the causal path between smoking and neonatal mortality. This illustrates the importance of basing analytic choices on substantive and not mechanical grounds.

SYNTHETIC ESTIMATION

Simple synthetic estimation is part of the process of indirect standardization. Stratum specific rates from either population data, or from a large sample survey are applied to observed numbers in the population of interest. In other words, the denominator of an SMR is a synthetic estimate. This estimate is used when data for an indicator are not collected for the local area and therefore there is no observed value from the local area to use as a numerator for an SMR. A synthetic estimate may also be used when the direct data for the area are available, but known to be under or over-reported, or very unreliable due to small sample size.

Indirect estimates, including SMRs as well as synthetic estimates, are probably biased (in a statistical sense), but they are usually quite reliable since they are derived from very stable rates from large populations or surveys. Recalling that the accuracy of an estimate is dependent on both bias and reliability, using a somewhat biased synthetic estimate may be preferable to using very unstable direct data. And when no direct data are available, the choice is between using a synthetic estimate or none at all.

Suppose we did not have access to direct data for the two hospital groups used in the previous example of standardization, but a comparison of neonatal mortality in the two groups was still desired. If the same hypothetical standard—based on all live births in the State—were used to calculate separate synthetic estimates for the two hospital groups, the results would be:

$$\text{Synthetic Estimate}_A = 4.4$$

$$\text{Synthetic Estimate}_B = 7.5$$

These are the denominators of the SMRs calculated earlier or the expected rates in the two hospital groups (see page 96). By using the neonatal mortality experience in the State as a whole as the standard, the assumption is that it is a reasonable reflection of the neonatal mortality experience in each hospital group. Notice, however, that the synthetic estimate for Hospital Group A is lower than that for Hospital Group B, implying that the community hospitals have a better neonatal mortality rate than do the tertiary care centers. This contradicts the results of both direct and indirect standardization as well as the stratified analysis. This kind of result illustrates the danger of using synthetic estimates.

For synthetic estimation to be credible, then, it is critical to use a standard with strata that account for important characteristics of the population for which the estimate is being calculated. With no direct data as a basis of comparison, we must be confident that the stratum specific rates applied to the numbers in

the population of interest are close to those that would be observed in that population. To calculate reasonable synthetic estimates for the tertiary care centers and community hospitals, for example, the standard would have to be stratified by hospital type as well as by birthweight—in effect, two distinct standards should be used. The data table for calculating the synthetic estimates would then be organized as follows, with 6 rather than 3 strata for the standard as well as for the hospital groups. Now, the synthetic estimates for the community hospitals can be calculated using only the rates from community hospitals statewide, and the synthetic estimate for the tertiary care centers can be calculated using only the rates from tertiary care centers statewide.

	# of Deaths in the Hospitals of Interest	# of Births in the Hospitals of Interest	Stratum Specific Rates from the Standard	Synthetic Estimates or Expected # of Deaths
Community Hospitals				
< 1500				
1500 – 2499				
>= 2500				
Tertiary Care Centers				
< 1500				
1500 – 2499				
>= 2500				

If an overall synthetic estimate for all hospitals in the region had been desired rather than separate estimates by hospital type, then using all live births in the state as the standard would have been appropriate. Combining the data for Hospital Group A and Hospital Group B into regional totals, we get:

Hospital Groups A and B Combined:

Birthweight	Deaths	Births	%	Stratum Specific Rates from the Standard per 1000	Crude Rate per 1000
< 1500	118	450	1.5	262.0	
1500 – 2499	26	1,950	6.5	13.0	
>= 2500	42	27,600	92.0	1.5	
	186	30,000	100.0		6.2

Remember that the crude rate in each of the hospital groups was 6.2 and therefore the combined crude rate is also 6.2. Multiplying the number of births in the birthweight strata in the above table by the neonatal mortality rates from the hypothetical standard, we get the following synthetic estimate:

$$\frac{450 \times 275 + 1,950 \times 14.0 + 27,600 \times 1.0}{30,000} = 6.0$$

This synthetic estimate of 6.0 is a reasonable reflection of the actual crude rate in the two hospital groups combined, 6.2.

Following is another example of synthetic estimation. The standard is national survey data stratified on multiple variables to calculate an overall state synthetic estimate of the percent of women who drink alcohol during pregnancy. The strata chosen to calculate the estimate were African-American/Non African-American, married/not married, and age < 20, 20-34, and >= 35. Separate estimates for African-American and Non African-American women, or for women in specific age groups could also be calculated analogous to the separate estimates for the community and tertiary care hospitals.

Although direct data have been available on the birth certificate for alcohol use during pregnancy since 1989, it is likely that in the first few years of data collection, the error rate was high. These data are for 1989.

Strata	# of Drinkers Reported on the Birth Certificate	# of Live Births	Stratum Specific Percents of Alcohol Use During Pregnancy National Survey Data	Synthetic Estimates or Expected # of Drinkers
African-American				
< 20, Married	0	281	0.06	17
< 20, Not Married	99	10,396	0.07	728
20-34, Married	183	7,531	0.11	828
20-34, Not Married	972	18,003	0.16	2,880
> = 35, Married	33	1,002	0.11	110
> = 35, Not Married	77	896	0.13	116
Non African-American				
< 20, Married	38	4,264	0.14	597
< 20, Unmarried	134	8,873	0.09	799
20-34, Married	2,761	98,712	0.24	23,691
20-34, Not Married	701	16,083	0.24	3,860
> = 35, Married	584	12,401	0.24	2,976
> = 35, Not Married	67	942	0.25	236
Total	5,649	179,384		36,838

$$\text{Direct Estimate from Birth Certificate} = (5,649/179,384)*100 = 3.1$$

$$\text{Synthetic Estimate} = (36,838/179,384)*100 = 20.5$$

In this example, the synthetic estimate of 20.5 % is much higher than the 3.1 % reported in the birth certificate data. We hypothesize that there is underreporting in the birth certificate data, yet we may also not feel confident in the synthetic estimate. Other stratification schemes might be tried to assess any change in the resulting synthetic estimate. In addition, a small local survey may be undertaken in an attempt to get another estimate as a point of comparison.

Should we report either of these less than ideal estimates of alcohol use during pregnancy? Often, it is a matter of public health discretion whether to report estimates in which we do not have full confidence, either because they contain statistical or epidemiologic bias or because they are unreliable. The consequences for public health programs of reporting or not reporting such estimates must be considered. Ironically, there is often more comfort in reporting estimates that are known or suspected to be inaccurate if they are direct estimates, such as poorly reported indicators from vital records, than in reporting indirect (synthetic) estimates that in fact may be more accurate. MCH professionals need to bring their clinical and programmatic knowledge to bear when deciding which indicators are important and which estimates of those indicators to report.

Test Yourself

Question:

Again using the hypothetical statewide data for the percent of children who have a medical home (See page 97 from Test Yourself), calculate a county synthetic estimate. Is it a reasonable estimate? Could its accuracy be increased?

Answer:

$$\begin{aligned}\text{Synthetic Estimate} &= \frac{780 \times 70 + 2,530 \times 45 + 2,170 \times 60 + 990 \times 80}{10,000} \\ &= 56.5\end{aligned}$$

The synthetic estimate of 56.5 % is fairly close to the actual observed county percent of 64.7. A variable such as insurance status, (or other variables related to both county of residence and having a medical home) could be used to create additional strata in an effort to increase the accuracy of the synthetic estimate.

OVERVIEW OF MULTIVARIABLE REGRESSION METHODS

Regression analysis can be viewed as an alternative to as well as an extension of stratified analysis. Like stratified analysis, regression approaches allow examination of multiple factors (independent variables) simultaneously in relation to an outcome (dependent variable) and provide a means of controlling confounding and examining effect modification. Unlike stratified analysis, regression approaches can more efficiently handle many variables, and continuous as well as discrete variables can be analyzed.

The most common regression models used to analyze health data express the hypothesized association between risk or other factors and an outcome as a linear relationship. Ordinary Least Squares (OLS) regression is used when the outcome of interest is a continuous variable (or is close to continuous). Logistic regression is used most often when the outcome of interest is dichotomous.

A linear model in its simplest form is as follows:

$$E(\text{Outcome}) = \text{Intercept} + (\text{Slope} \times \text{risk factor})$$

where E stands for "Expected Value of"

When the outcome is a continuous, normally distributed variable, the values of the outcome variable are assumed to be linearly related to the values of a risk factor or other variable of interest. The model is written as follows:

$$E(Y) = a + (b \times \text{risk factor})$$

When the outcome is a dichotomous, binomially distributed variable, it is the *natural logarithm of the odds of the outcome* that is assumed to be linearly related to the values of a risk factor or other variable of interest. The model is written:

$$E\left(\ln \frac{p}{1-p}\right) = a + (b \times \text{risk factor})$$

When the outcome is a very small binomial variable, or Poisson distributed, it is the *natural logarithm of the rate of the outcome* that is assumed to be linearly related to the values of a risk factor or other variable of interest. The model is written:

$$E(\ln r) = a + (b \times \text{risk factor})$$

Regression models provide a more comprehensive approach to testing hypotheses about associations based on means and proportions. Recall that hypothesis tests are carried out in the following way:

$$\text{Test Statistic} = \frac{\text{Observed Association} - \text{Expected Association}}{\text{Standard Error of the Association}}$$

Previously, we expressed an observed association in terms of differences between means and proportions or as odds ratios and relative risks. These same measures of association are relevant in regression analysis, but now they are embodied in the slope of the regression line, or what is called the beta coefficient. The test statistic from regression analysis, then, is written:

$$\text{Test Statistic} = \frac{\text{Observed Slope(beta)} - \text{Expected Slope(beta)}}{\text{Standard Error of the Slope(beta)}}$$

The observed beta coefficient can theoretically range from negative infinity to positive infinity ($-\infty$ to ∞), and its expected value under the assumption of no association (the null hypothesis) is 0. In OLS regression, with an outcome that is a continuous, normally distributed variable, the slope or beta coefficient is a measure of differences between means. In logistic regression, with an outcome that is a binomial proportion, the slope or beta coefficient is the odds ratio.

Regression models can be quite complicated, including many independent variables, both continuous and discrete. In this way, regression models attempt to capture the complex context in which health events occur. A more complicated linear regression model looks like:

$$E(\text{Outcome}) = \text{Intercept} + (\text{beta}_1 \times \text{factor}_1) + (\text{beta}_2 \times \text{factor}_2) + \dots + (\text{beta}_n \times \text{factor}_n)$$

Test statistics for each of the many beta coefficients may be of interest, or more typically, most of the variables are included as potential confounders or effect modifiers in order to insure appropriate adjustment of the beta coefficient for the variable of primary interest. Statistics that assess the strength of the entire model, or the association between all of the variables in the model jointly with the outcome may also be examined.

In addition to statistical testing, regression models can help us better describe the occurrence of a health outcome in a population in a multivariable context which, of course, better mirrors the multivariable world. Following are examples of different regression models with brief explanations of how the results might be interpreted. The examples illustrate inclusion of different types of variables as well as how to substitute values into the regression equation to obtain useful, reportable estimates. In the examples of OLS regression, estimates of means and then mean differences are shown; in the examples of logistic regression, estimates of odds and then odds ratios are shown. The numbers in these examples are hypothetical, based only loosely on what might be found in the MCH literature.

First, here are some examples using ordinary least squares regression:

Example 1: Birthweight is the outcome variable measured continuously in grams, adequacy of prenatal care (PNC) is the independent variable measured dichotomously with women who received no or inadequate PNC coded 1 and those who received adequate PNC coded 0.

$$E(\text{birthweight in grams}) = 3,150 + (-250 \times \text{PNC})$$

On average, the birthweight of infants born to women who received no or inadequate prenatal care is estimated to be:

$$3,150 - 250(1) = 2,900 \text{ grams}$$

On average, the birthweight of infants born to women who received adequate prenatal care is estimated to be:

$$3,150 - 250(0) = 3,150 \text{ grams}$$

In this case, the beta coefficient is a measure of the difference in mean birthweight among women with varying patterns of prenatal care utilization. Infant birthweight among women who received no or inadequate prenatal care is 250 grams less, on average, than infant birthweight among women who received adequate prenatal care. The standard error of the beta coefficient can be used to calculate a confidence interval around this value of 250 grams, and depending on whether this confidence interval includes 0, the associated p value will indicate whether this mean difference is statistically significant.

Example 2: Birthweight is the outcome variable measured continuously in grams, adequacy of prenatal care is the independent variable measured continuously in number of visits.

$$E(\text{birthweight in grams}) = 3,150 + (12 \times \# \text{ of prenatal care visits})$$

On average, the birthweight of infants born to women who received 10 prenatal care visits is estimated to be:

$$3,150 + 12(10) = 3,270 \text{ grams}$$

On average, the birthweight of infants born to women who received 5 prenatal care visits is estimated to be:

$$3,150 + 12(5) = 3,210 \text{ grams}$$

On average, the birthweight of infants born to women who received no prenatal care is estimated to be:

$$3,150 + 12(0) = 3,150 \text{ grams}$$

As before, the beta coefficient is a measure of the difference in mean birthweight, but now the difference can be assessed for each single additional prenatal care visit. For each visit, infant birthweight increases, on average, by an estimated 12 grams. Also as before, a standard error of the beta coefficient can be used to calculate a confidence interval around this value of 12 grams, and depending on whether this confidence interval includes 0, the associated p value will indicate whether this mean difference is statistically significant.

When continuous variables such as number of prenatal care visits are included in a linear regression model, the assumption is that the association between the independent and dependent variable is the same across all of the values of the independent variable. In the above example, it is assumed that each additional prenatal care visit adds equivalent benefit; the change in infant birthweight if a woman receives 10 versus 9 prenatal care visits is the same as the change if a woman receives 2 versus 1 prenatal care visit. For many variables used in maternal and child health, this may not be a valid assumption.

Consider maternal AGE as a variable of interest in studying perinatal outcomes. Women who deliver at a very young age and women who deliver at ages over 35 are usually considered at higher risk of adverse outcomes than other women. In order to capture this non-linear relationship in a regression model, "dummy" variables are coded. These are a set of dichotomous variables, coded "1" and "0", that jointly represent the effect of age.

Suppose 3 categories of maternal AGE are to be examined:

< 18
18-34
35 +

In order to examine these categories in a regression model, the following 2 "dummy" variables would be created:

AGE	LESS18	PLUS35
< 18	1	0
18-34	0	0
35+	0	1

Notice that the number of "dummy" variables coded is equal to one less than the number of categories of interest. Here, the 3 categories of AGE are fully defined with 2 "dummy" variables. The variable value which is coded "0" on each "dummy" is called the reference group; here, it is the women who are 18-34. This is the comparison group for all odds ratios pertaining to age.

A similar categorization scheme may be called for when considering prenatal care visits. Women with very few and very many prenatal care visits may be at higher risk for delivering a low birthweight infant. The following dummy variables might be used in a regression model to address this situation:

# OF PRENATAL CARE VISITS	LESS5	PLUS12
< 5	1	0
5-12	0	0
> 12	0	1

Let's examine what might occur in a regression model employing these dummy variables for prenatal care visits.

Example 3: Birthweight is the outcome variable measured continuously in grams, prenatal care is the independent variable measured with two dummy variables coded as in the above table.

$$E(\text{birthweight in grams}) = 3,150 + (-240 \times \text{LESS5}) + (-70 \times \text{PLUS12})$$

On average, the birthweight of infants born to women who received fewer than 5 prenatal care visits is estimated to be:

$$3,150 - 240(1) - 70(0) = 2,910 \text{ grams}$$

On average, the birthweight of infants born to women who received from 5-12 prenatal care visits is estimated to be:

$$3,150 - 240(0) - 70(0) = 3,150 \text{ grams}$$

On average, the birthweight of infants born to women who received more than 12 prenatal care visits is estimated to be:

$$3,150 - 240(0) - 70(1) = 3,080 \text{ grams}$$

The first beta coefficient (LESS5) is a measure of the difference in mean birthweight between women with fewer than 5 prenatal care visits compared to women with 5-12 visits; the second beta coefficient (PLUS12) is a measure of the difference in mean birthweight between women with more than 12 prenatal care visits compared to women with 5-12 visits. Infant birthweight among women who received fewer than 5 visits is 240 grams less, on average, than infant birthweight among women who received 5-12 visits. Infant birthweight among women who received more than 12 visits is 70 grams less, on average, than infant birthweight among women who received 5-12 visits. The standard errors of the beta

coefficients can be used to calculate confidence intervals around the values of 240 and 70. Depending on whether these confidence intervals include 0, the associated p values will indicate whether these mean differences are statistically significant.

Comment and Example

Use of "dummy" variables is important because it permits inclusion of many categorical variables in a regression model without assuming a linear relationship. Suppose we want to assess the differences in mean number of well child visits received by two-year-olds in a three county region. The counties could be coded as follows:

County	VARXZ	VARYZ
X	1	0
Y	0	1
Z	0	0

The first dummy variable is named "VARXZ" to indicate that County X is coded 1 and County Z is the reference group; likewise, the second dummy variable is named "VARYZ" to indicate that County Y is coded 1 and again County Z is the reference group. A regression model of interest might be:

Dependent Variable: # well child visits

Independent Variables:

1. VARXZ
2. VARYZ
3. # pediatricians per child pop. in the county
4. Insurance Status (yes/no)

The results of this model would help answer questions about whether there are any differences in the amount of well child care being received by children in the three counties after accounting for differences in the supply of pediatricians and the insurance status of the children.

Next, here are a few examples using a logistic regression model:

Example 4: Birthweight is the outcome variable measured dichotomously with low birthweight (LBW) coded 1 and normal birthweight coded 0, prenatal care is the independent variable measured dichotomously with women who received no or inadequate PNC coded 1 and those who received adequate PNC coded 0.

$$E(\ln \text{ odds of LBW}) = -2.80 + (0.47 \times \text{PNC})$$

On average, the odds of delivering a low birthweight infant among women whom received no or inadequate prenatal care is estimated to be:

$$e^{-2.80 + 0.47(1)} = 0.097$$

On average, the odds of delivering a low birthweight infant among women who received adequate prenatal care is estimated to be:

$$e^{-2.80 + 0.47(0)} = 0.061$$

The odds ratio for the association between prenatal care and low birthweight is estimated to be:

$$\frac{e^{-2.80 + 0.47(1)}}{e^{-2.80 + 0.47(0)}} = e^{0.47(1-0)} = 1.6$$

Here, you can see that when variables are coded "1" and "0", the beta coefficient itself is a measure of the natural logarithm of the odds ratio. On average, women who received no or inadequate prenatal care are 1.6 times more likely to deliver a low birthweight infant than are women who received adequate prenatal care. The standard error of the beta coefficient can be used to calculate a confidence interval around the value of 0.47 which when exponentiated yields a confidence interval around the value of 1.6. Depending on whether this confidence interval includes 1, the associated p value will indicate whether this odds ratio is statistically significant.

Example 5: Birthweight is the outcome variable measured dichotomously with low birthweight coded 1 and normal birthweight coded 0, prenatal care is the independent variable measured with two dummy variables coded as in the above table.

$$E(\ln \text{ odds of LBW}) = -2.80 + (0.53 \times \text{LESS5}) + (0.28 \times \text{PLUS12})$$

On average, the odds of delivering a lbw infant among women who received fewer than 5 prenatal care visits is estimated to be:

$$e^{-2.80 + 0.53(1) + 0.28(0)} = 0.10$$

On average, the odds of delivering a lbw infant among women who received 5-12 prenatal care visits is estimated to be:

$$e^{-2.80 + 0.53(0) + 0.28(0)} = 0.06$$

On average, the odds of delivering a lbw infant among women who received more than 12 prenatal care visits is estimated to be:

$$e^{-2.80 + 0.53(0) + 0.28(1)} = 0.08$$

The odds ratio for the association between receiving fewer than 5 visits versus 5-12 visits and low birthweight is estimated to be:

$$\frac{e^{-2.80 + 0.53(1) + 0.28(0)}}{e^{-2.80 + 0.53(0) + 0.28(0)}} = e^{0.53(1-0)} = 1.7$$

The odds ratio for the association between receiving more than 12 visits versus 5-12 visits and low birthweight is estimated to be:

$$\frac{e^{-2.80 + 0.53(0) + 0.28(1)}}{e^{-2.80 + 0.53(0) + 0.28(0)}} = e^{0.28(1-0)} = 1.3$$

The first beta coefficient (LESS5) is a measure of the natural logarithm of the odds ratio for women with fewer than 5 prenatal care visits compared to women with 5-12 visits; the second beta coefficient (PLUS12) is a measure of the natural logarithm of the odds ratio for women with more than 12 prenatal care visits compared to women with 5-12 visits. On average, women who received fewer than 5 visits are 1.7 times more likely to deliver a low birthweight infant than are women who received 5-12 visits; on

average, women who received more than 12 visits are 1.3 times more likely to deliver a low birthweight infant than are women who received 5-12 visits. The standard errors of the beta coefficients can be used to calculate confidence intervals around the values of 0.53 and 0.28 which when exponentiated yield confidence intervals around the values of 1.7 and 1.3. Depending on whether these confidence intervals include 1, the associated p values will indicate whether these odds ratios are statistically significant.

If a continuous variable is to be used in a logistic regression model, an increase or decrease in the odds ratio for each single unit change in such a variable is usually not meaningful. For example, suppose we felt comfortable using number of prenatal care visits in its continuous form. The resulting beta coefficient from a logistic regression model might be 0.049, yielding an odds ratio for each additional prenatal visit of e^b or $e^{0.049} = 1.05$. This is not a readily interpretable result. Typically with a continuous variable, odds ratios are calculated for what is considered to be clinically or programmatically meaningful. An odds ratio might be calculated for a 5 visit difference, for instance, $e^{b(5)}$ or $e^{0.049(5)} = 1.28$.

Example 6: Birthweight is the outcome variable measured dichotomously with low birthweight coded 1 and normal birthweight coded 0, adequacy of prenatal care is the first independent variable measured dichotomously with women who received no or inadequate prenatal care coded 1 and women who received adequate care coded 0, and smoking status during pregnancy (SMOKE) is a second independent variable measured dichotomously with smokers coded 1 and nonsmokers coded 0.

$$E(\ln \text{ odds of LBW}) = -2.80 + (0.35 \times \text{PNC}) + (0.60 \times \text{SMOKE}) + (0.04 \times \text{PNC} \times \text{SMOKE})$$

In addition to the two independent variables for adequacy of prenatal care and smoking status, the model also includes a third variable that is the multiplication of these other two. This variable is called an interaction term and permits assessment of whether there is effect modification. Below are the possible combinations of values on the three variables in the model:

Coding of Variables

	PNC	SMOKE	PNC×SMOKE
Women with no/inadequate pnc who smoke	1	1	1
Women with adequate pnc who smoke	0	1	0
Women with no/inadequate pnc who do not smoke	1	0	0
Women with adequate pnc who do not smoke	0	0	0

On average, the odds of delivering a lbw infant among women who received no or inadequate prenatal care and smoked during pregnancy is estimated to be:

$$e^{-2.80 + 0.35(1) + 0.60(1) + 0.04(1)} = 0.16$$

On average, the odds of delivering a lbw infant among women who received adequate prenatal care and smoked during pregnancy is estimated to be:

$$e^{-2.80 + 0.35(0) + 0.60(1) + 0.04(0)} = 0.11$$

On average, the odds of delivering a lbw infant among women who received no or inadequate prenatal care and did not smoke during pregnancy is estimated to be:

$$e^{-2.80 + 0.35(1) + 0.60(0) + 0.04(0)} = 0.086$$

On average, the odds of delivering a lbw infant among women who received adequate prenatal care and did not smoke during pregnancy is estimated to be:

$$e^{-2.80 + 0.35(0) + 0.60(0) + 0.04(0)} = 0.06$$

On average, the odds ratio for the association between no or inadequate prenatal care and lbw among women who smoked during pregnancy is estimated to be:

$$\frac{e^{-2.80 + 0.35(1) + 0.60(1) + 0.04(1)}}{e^{-2.80 + 0.35(0) + 0.60(1) + 0.04(0)}} = e^{0.35(1-0) + 0.04(1-0)} = 1.48$$

On average, the odds ratio for the association between no or inadequate prenatal care and lbw among women who did not smoke during pregnancy is estimated to be:

$$\frac{e^{-2.80 + 0.35(1) + 0.60(0) + 0.04(0)}}{e^{-2.80 + 0.35(0) + 0.60(0) + 0.04(0)}} = e^{0.35(1-0)} = 1.42$$

When an interaction term (the variable assessing effect modification) is in a model, the meaning of each beta coefficient is not so straightforward. The first odds ratio above, for example, is determined by both the value of the beta coefficient for the prenatal care variable as well as the beta coefficient for the prenatal care by smoking variable. Likewise, the standard error and the test of statistical significance for this association is based on information from both the prenatal care variable and the prenatal care by smoking variable. These more complicated calculations are required because, by definition, the presence of effect modification requires the use of stratum specific estimates rather than a simpler summary measure.

To make the regression results easier to interpret when effect modification is present, often the analysis is stratified. In this example, two separate regression models could be run, one including only women who smoked during pregnancy, the other including only women who did not smoke during pregnancy. Now, prenatal care would be the only independent variable in each model, and its beta coefficient could be interpreted in the straightforward way it was in the earlier examples.

If there is no effect modification, as in this example with the equivalent stratum specific odds ratios of 1.48 and 1.42, stratified analysis is unnecessary and the interaction term can be removed from the model. If one regression model is used including both prenatal care and smoking, then the beta coefficient for prenatal care (after exponentiation) is the odds ratio for the association between adequacy of prenatal care and low birthweight adjusted for the effect of smoking. The model might look like:

$$E(\ln \text{ odds of LBW}) = -2.80 + (0.37 \times \text{PNC}) + (0.62 \times \text{SMOKE})$$

The association of adequacy of prenatal care and low birthweight adjusted for the effect of smoking is $e^{0.37} = 1.44$. For these data, then, smoking is a confounder since the crude odds ratio was 1.6 as seen earlier (page 109).

Comment

Stratified analysis is analogous to the logistic regression model. For this example, the 2×2 tables would be:

Smokers

Low Birthweight

Y N

No or Inadequate PNC	Y		
	N		

Non-Smokers

Low Birthweight

Y N

No or Inadequate PNC	Y		
	N		

Test Yourself

Question:

Following are hypothetical results of a logistic regression model for examining adolescent suicide in relation to age and gender. Suicide is a dichotomous variable, 1=yes, 0=no. Age is a dichotomous variable, 1=20-21 years old, 0=15-19 years old. Gender is a dichotomous variable, 1=male, 0=female. What is the odds ratio for the association between gender and suicide after adjusting for the effect of age? Interpret this result.

$$E(\ln \text{ odds of committing suicide}) = -9.7 + 0.1 \times \text{age} + 1.3 \times \text{gender}$$

Answer:

OR= $e^{1.3}$ =3.7. On average male adolescents are 3.7 times more likely to commit suicide than female adolescents.