Research Methods in Epidemiology

David S. Younger, MD, MPH, MS[a,b,*], Xiaoling Chen, MPH[a]

KEYWORDS

• Research • Epidemiology • Statistics

KEY POINTS

- Measures of disease frequency include incidence, prevalence, odds and mortality rates.
- Summary measures of disease and disability have relevance to global health and have been used successfully by the Global Burden of Disease Study.
- Measures of effect and association include risk difference, risk ratio, relative excess, null state and counterfactual, and prevalence ratios.
- Epidemiologic estimation examines validity and error through estimation of *P* values, hypothesis testing, confidence intervals, and confidence limits.
- The major types of epidemiology study designs are randomized controlled trials, and nonexperimental study types including cohort, case control, cross-sectional and ecological study types.

INTRODUCTION

Neuroepidemiology is a branch of epidemiology involving the study of neurologic disease distribution and determinants of frequency in human populations. It includes the science of epidemiologic measures, estimation, hypothesis testing, and design of experimental and nonexperimental studies. This article reviews basic aspects of epidemiology that will be useful for articles relating to the neuroepidemiology of diverse neurologic disorders. Those interested in reading more deeply into the area of research methods of epidemiology are directed to general texts and monographs on the subject.[1–8]

MEASURES OF DISEASE FREQUENCY

Greenland and Rothman[9] review measures of occurrence elaborated further below.

The authors have nothing to disclose.

[a] Division of Neuroepidemiology, Department of Neurology, New York University School of Medicine, New York, NY, USA; [b] College of Global Public Health, New York University, New York, NY, USA

* Corresponding author. 333 East 34th Street, 1J, New York, NY 10016.

E-mail address: David.younger@nyumc.org

Neurol Clin 34 (2016) 815–835

http://dx.doi.org/10.1016/j.ncl.2016.05.003 **neurologic.theclinics.com**

0733-8619/16/$ – see front matter

Incidence

Incidence is defined as the occurrence of new cases of disease that develop in a candidate population over a specified time period. Cumulative incidence is the proportion of candidate population that becomes diseased over a specific time period mathematically expressed as follows:

$$\text{Incidence (I)} = \frac{\text{number of new cases of a disease}}{\text{number of candidate population}} \text{ over a specified time period.}$$

Note that the numerator is a subset of the denominator and thus the possible value of cumulative incidence ranges from 0 to 1, or if expressed as a percentage, from 0% to 100%. Time is not an integral part of this proportion, but rather is expressed by the words that accompany the numbers. Thought of as an average risk of getting a disease over a certain period of time, cumulative incidence is a commonly cited measure such as the "lifetime risk" currently estimated at 1 in 8, among United States men for the occurrence of stroke. Cumulative incidence is mainly used in fixed populations when there are no or only small losses to follow-up.

Epidemiologists have recognized that outcome events are not inevitable and may not occur during the period of observation; hence, the set of incidence times for a specific event in a population may not be precisely timed or observed. One way to deal with this complication has been to develop a measure to account for the length of time each individual contributes to the population at risk for the event during the period of time during which the event was a possibility and would have been counted in the population had it occurred. The sum of the person-times over all population members, termed the *total person-time at risk* or the *population-time at risk*, is the total of time during which disease onsets could occur in the population of interest. The incidence rate (IR), person-time rate, or incidence density of the population is defined as the number of new cases of disease (incident number) divided by the person-time over the period:

$$\text{Incidence rate (IR)} = \frac{\text{number of disease onsets}}{\sum_{\text{persons}} \text{time spent in population}}$$

When the risk period is of a fixed length equal to Δt, the proportion of the period that a person spends in the population at risk is their amount of person-time divided by Δt such that the average size of the population over the period of time is represented by:

$$\overline{\text{N}} = \sum_{\text{persons}} \frac{\text{time spent in population}}{\Delta \text{t}}$$

The total person-time at risk over the period is equal to the product of the average size of the population over the period $\overline{N}$, and the fixed length of the risk period Δt. If the incident number is denoted by A, it follows that the IR equals $A/(\overline{N} \times \Delta t)$. This formulation shows that the IR has units of inverse time that can be rewritten as year^{-1}, month^{-1}, or day^{-1}. The outcome events that can be counted in the numerator of an IR are those that occur to persons who are contributing to the denominator of the IR at the time that the disease onset occurs. Likewise, only time contributed by persons eligible to be counted in the numerator if they suffer such an event should be included in the denominator. An alternative way of expressing a population IR is as a time-weighted average of individual rates. An individual rate is either 0/(time spent in population) = 0 if the individual does not experience the event, or 1/(time spent in

population) if the individual does experience the event. One then has the number of disease onsets A shown as follows:

$$A = \sum_{\text{persons}} (\text{time spent in population})\ (\text{individual rate})$$

and so

$$IR = \frac{\sum_{\text{persons}} (\text{time spent in population})\ (\text{individual rate})}{\sum_{\text{persons}} (\text{time spent in population})}$$

Although a central one in epidemiology, the IR has certain limitations. First, it does not capture all aspects of disease occurrence as illustrated by the analogy of noting that a rate of 1 case/(100 years) = 0.01 year^{-1} could be obtained by following 100 people for an average of 1 year and observing 1 case, but could also be obtained by following 2 people for 50 years and observing 1 case. To distinguish these situations, more detailed measures are needed, such as incidence time. Second, the numeric value lacks interpretability because the IR ultimately depends on the selection of time in which it is presented to give it significance. This is illustrated by the fact that an IR of 100 cases per 1 person-year might be expressed as: $100\frac{\text{cases}}{\text{person}-\text{year}}$, $8.33\frac{\text{cases}}{\text{person}-\text{month}}$, $1.92\frac{\text{cases}}{\text{person}-\text{week}}$, or $0.27\frac{\text{cases}}{\text{person}-\text{day}}$. Likewise, the IRs of 0.15, 0.04, and 0.009 cases per person-year could be multiplied by 1000 to be displayed as 150, 40, and 9 cases per 1000 person-years, regardless of whether the observations are made over 1 year of time, 1 week of time, or over a decade, just as one could measure the speed of a car in miles per hour even regardless of whether it is measured for only a few seconds.

Although the IR often includes the first occurrence of disease onset as an eligible event for the numerator of the equation, in many diseases there may be repeated events particularly in neurologic disorders such as multiple sclerosis and chronic inflammatory demyelinating polyneuropathy, both of which are characterized by relapses and remissions in which there may even be a disease-free period between recurrences. When the events tallied in the numerator of an IR are the first occurrence of disease, then the time contributed by each person in whom the disease develops should terminate with the onset of disease, meaning that further information would be obtained from further observation. Thus, each person who experiences the outcome should contribute time to the denominator until the occurrence of the event, but not afterward. In studies of both first and subsequent occurrences of a disease in which it is not important to distinguish between the first and subsequent occurrences, then the time accumulated in the denominator of the rate would not cease with the occurrence of the outcome event. A useful approach is to define the "population at risk" differently for each occurrence of the event, such as studies of individuals restricted to the population of those who have survived the first event of a given disease such as chronic inflammatory demyelinating polyneuropathy. The distinguished populations may be closed or open, with the former adding no new members over time and lost only to attrition. The term *cohort* is sometimes used to describe a study population, although typically it is reserved for a narrower concept as that of a group of persons for whom the membership is defined by a single event. If the number of people entering a population is balanced by the number exiting, the population is said to be stationary or in a steady state. In a stationary population with no migration, the crude IR of an inevitable outcome such as death will equal the reciprocal of the

average time spent in the population until the outcome occurs such that a death rate of 0.04 year^{-1} would translate into an average time from entry until death of 25 years.

Within a given time interval, the incident number of cases can also be expressed in relation to the size of the population at risk such that in the absence of immigration or emigration, such a rate becomes the proportion of people who become cases among those in the population at the start of the interval. The latter is defined as the proportion of a closed population at risk that becomes diseased within a given period of time. Thus, the number of disease onsets *A* is then the sum of the individual proportions:

$$A = \sum_{\text{persons}} \text{individual proportions}$$

And thus,

$$\text{Incidence proportion (IP)} = \frac{\sum_{\text{persons}} \text{individual proportions}}{\text{initial size of the population}} = A/N$$

So, defined as individual risks, this formulation illustrates that the IP is an average risk, ignoring the amount of person-time contributed by individuals but with a more intuitive interpretation than the IR. With a range from 0 to 1, it is dimensionless however an IP of disease of 3% means something very different when it refers to a 40-year period than a 40-day period.

With regard to their numerical value, a cumulative incidence and rate can only be compared if they are based on the same time unit, for example, cumulative incidence over a 1-year period and rate per person-year. Under this circumstance, the general rule is that in absolute value, the rate will always be larger than the cumulative incidence notably when *x* cases are lost to follow-up with or without censoring (*C*).

In the absence of censoring:

$$I = \frac{x}{N} \text{ and } IR = \frac{x}{N - \frac{1}{2}x}$$

While in the presence of censoring with I and IR can be, respectively, represented as:

$$\frac{x}{N - \frac{1}{2C}} \text{ and } \frac{x}{N - \frac{1}{2C} - \frac{1}{2x}}$$

As long as $x > 0$, the denominator of the rate will always be smaller than that of the cumulative incidence, explaining the greater absolute value of the rate. This occurs because the cumulative incidence is based on the number of individuals at risk at the beginning of the interval whereas the rate is based on person-time of observations over the follow-up period, subtracting person-time lost by cases.

The hazard rate (H) is an alternative definition for an instantaneous IR also called instantaneous conditional incidence. It is defined as each individual's instantaneous probability of the event at precisely time *t* or at a small interval $[t, t + \Delta t]$ given or conditioned on the fact that the individual was at risk at time *t*. Thus the hazard rate is defined for each particular point in time during the follow-up in mathematical terms for a small time interval, assuming Δt is close to zero as follows:

$$\frac{P\,(\text{event in interval between } t \text{ and } [t + \Delta t] \mid \text{alive at } t)}{\Delta t}$$

The H is analogous to the conditional probability of an event that is calculated at each event time using the Kaplan-Meier approach however because its denominator is "time at risk," it is instead a rate measured in unit of time^{-1}. Moreover, in contrast

with the Kaplan-Meier conditional probability, the H cannot be directly calculated as it is defined for an infinitely small time interval; however, the H function over time can be estimated using available parametric survival analysis techniques.

Prevalence

Unlike incidence measures, which focus on new events or changes, prevalence focuses on existing states. Although incidence measures the frequency with which new disease develops, prevalence measures the frequency of existing disease. It is simply defined as the proportion of the population with the disease. Point prevalence and period prevalence refer to 2 types of prevalence measures. The first refers to the proportion of the population that is diseased at a single point in time, thought of as a snapshot of the population, and the latter as the proportion of the population that is diseased during a specified duration of time. Mathematically, point prevalence and period prevalence are expressed respectively as follows:

$$\frac{\text{Number of existing cases of disease}}{\text{Number of total population}} \quad \text{At a point } \textit{or} \text{ in a period of time.}$$

Prevalence depends on the rate at which new cases of disease develop and the duration (*D*) or length of time that individuals have the disease. The duration of disease starts at the time of diagnosis and ends when the person either is cured or dies. Mathematically, the relationship between prevalence and incidence is as follows:

$$P/(1 - P) = IR \times D$$

where *P* is prevalence (the proportion of the total population with the disease), 1-*P* is the proportion of the total population without the disease, IR is the incidence rate, and *D* is the average duration (or length of time) that an individual has the disease. This equation assumes that the population is in steady state of inflow equals outflow, and that the IR and duration do not change over time.

If the population at risk and the prevalence pool are stationary and everyone is either at risk or has the disease, then the number of people entering the prevalence pool in any time period will be balanced by the number exiting from it. Supposing there is no immigration into or emigration from the prevalence so that no one enters or leaves the pool except by disease onset, death, or recovery, the size of the population at risk will be the size of the population (*N*), minus the size of the prevalence pool (*P*); and during any time interval of length Δt, the number who enter the prevalence pool would be $IR\ (N - P)\ \Delta t$ and the outflow of the prevalence pool would be $IR'P\ \Delta t$ where IR' is the IR of exiting the prevalence pool. In the absence of migration, the reciprocal of IR' will equal the mean duration of the disease, $\overline{D}$; and so it follows that:

$$\text{Inflow} = IR\ (N - P)\ \Delta t = \text{outflow} = (1/\overline{D})\ P\Delta t \text{ which yields } \frac{P}{N - P} = IR \times \overline{D} \text{ where } P/(N - P)$$

is the ratio of diseased to nondiseased people in the population or equivalently the ratio of the prevalence proportion to the nondiseased proportion, or the prevalence odds. If the prevalence is small (<0.1), then *P* approximates $IR \times \overline{D}$.

Although incidence is most useful for evaluating the effectiveness of a program that seeks to prevent diseases from occurring in the first place, researchers who study the cause of disease and prefer to examine new cases (incidence) over existing ones (prevalence) because they are interested more in exposures that lead to developing the disease in question.[2] Prevalence obscures the relationship because it combines incidence and survival. On the other hand, prevalence is most useful for estimating

the needs of medical facilities and allocating resources for treating individuals who already have a disease.

Odds

Odds are the ratio of the probability of an event of interest to that of the nonevent and can be defined both for incidence and for prevalence. When dealing with incidence probabilities, the odds are simply:

$$\frac{I}{1-I}$$

And while knowing the odds allow the calculation of probability:

$$\frac{\text{odds}}{1+\text{odds}}$$

The point prevalence odds are similarly expressed as:

$$\frac{\text{Point prevalence}}{1-\text{Point prevalence}}$$

Both odds and proportions can be used to express the frequency of a disease and can approximate a proportion when the disease is very small (eg, <0.1). If the proportion of a disease is known in a population to be 0.20, the odds of the disease can be expressed as:

$$\frac{\text{proportion of the population with disease}}{1-\text{proportion of population with disease}} = \frac{\text{proportion of the population with disease}}{\text{proportion of the population without disease}}$$

Although as an isolated measure epidemiologists rarely if ever use odds to express disease occurrence, the odds ratio (OR) is a useful measure of association because it estimates the relative risk (RR) in case-based studies.

Mortality Ratios

Other measures of risk are generally expressed using mortality ratios (MR) to estimate the frequency of this occurrence of death in a defined population over a specified interval, whether expressed as crude mortality for all causes in a population or a single cause. MRs can be studied in reference to infant and maternal deaths, or adjusted for sex, age, race, and ethnicity, or by particular conditions or the proportion thereof to provide insight into public health responses to the leading causes of mortality and health disparities. Crude mortality generally refers to the total number of deaths from all causes based on raw data per 100,000 population per year. Cause-specific mortality is the number of deaths from a specific cause per 100,000 population per year. Age-specific mortality or the death rate is the total number of deaths from all causes among individuals in a specific age category per 100,000 population per year in the age category. Often divided into neonatal deaths occurring during the first 27 days after birth and postneonatal deaths occurring from 28 days to 12 months, the infant mortality rate is the number of number of deaths of infants less than 1 year of age per 1000 live births per year. The morbidity rate represents the number of existing or new cases of a particular disease or condition per 100 population. The time period and the size of the population concerned may vary. The attack rate is the number of new cases of disease that develops usually during a defined and short time period, per the

number in a healthy population at risk at the start of the period. This cumulative incidence measure is usually reserved for infectious outbreaks. The case fatality rate is the number of deaths per number of cases of disease. This is a type of cumulative incidence, so it is generally useful to specify the length of time to which it applies. Last, the survival rate is the number of living cases per number of cases of diseased. This rate is the complement of the case fatality rate, and is also a cumulative incidence measure.

Disability and Summary Measures

The term *premature mortality* was originally proposed to address the inadequacy of MR in measuring the burden of disease owing to tuberculosis and has since proved to be a particularly useful way to describe other diseases.[10] In choosing an arbitrary limit to life, the calculation of the difference between the age at death and an arbitrary designated limit measured in years of life lost (YLL) owing to premature mortality is a useful assessment of the impact of premature mortality in a given population. The YLL rate, which represents years of potential life lost per 1000 population below an arbitrary endpoint age such as 65 years, is more desirable in comparing premature mortality in different populations because the YLL does not take into account differences in population sizes.[11]

Another measure of the burden of disease in a population termed disability-adjusted life years captures in a single figure, health losses associated with mortality and different nonfatal outcomes of diseases and injuries.[12] Summary measures used by the Global Burden of Disease Studies[13,14] such as Healthy Adjusted Life Expectancy (HALE) are derived from YLLs and years lived with disability (YLDs) to compare assessments of broad epidemiologic patterns across countries and time, and to quantify the component of variation in epidemiology related to sociodemographic development. Calculated by adding YLLs and YLDs, disability-adjusted life years add disability to the measure of mortality, and based on the universal measure of time in life years, provides a common currency for health care resource allocation and the effectiveness of interventions assessed one relative to another across a wide range of health problems. YLDs, equal to the sum of prevalence multiplied by the general public's assessment of the severity of health loss, has been used as a primary metric to explore disease patterns over time, age, sex, and geography, and in recognizing that the aging of the world's population has led to substantial increases in the number of individuals with sequelae of diseases.[15] Because YLDs have been declining much more slowly than MRs, the nonfatal dimensions of disease require more and more attention from health care systems. Neurologic disorders accounted for 7.7% of all cause YLDs in 2013, a 5% increase in age-standardized YLDs from 1990 to 2013 (2.4%–7.9%) with the leading causes being Alzheimer disease, Parkinson disease, epilepsy, multiple sclerosis, migraine, tension and medication overuse headaches, and other neurologic disorders.[14]

MEASURES OF EFFECT AND ASSOCIATION

Measures of effect compare what would happen to 1 population under 2 possible but distinct life courses or conditions, of which at most only 1 can occur. In contrast, a measure of association compares what happens in 2 distinct populations, although the 2 distinct populations may correspond to one population in different time periods. Subject to physical and social limitations, one can observe both populations and so can directly observe an association. Greenland and

colleagues[16] review measures of effect and measures of association further detailed below.

Risk Difference

Consider a cohort followed over a specific time or between a given age interval under 2 different conditions, some of whom who are exposed to a potentially harmful factor, and others who are not and one asks the question of the alternative potential outcomes in each of the 2 cohort groups. The IR of each potential outcome could be expressed as a difference in IR or *causal rate difference*, alternatively as a difference in incidence proportions to derive an absolute effect per se of a treatment intervention to derive an absolute effect on the incidence proportion. Supposing that we have a cohort of size N defined at the start of a fixed time interval and that anyone alive without the disease is at risk of the disease, everyone is exposed throughout the time interval leading to A_1 cases over a time T_1. A_0 cases will also occur over a total time risk noted as T_0. Thus the causal rate difference will be written as:

$$\frac{A_1}{T_1} - \frac{A_0}{T_0}$$

whereas the casual risk difference will be expressed as:

$$\frac{A_1}{N} - \frac{A_0}{N}$$

And the causal difference in average disease-free time would be noted as:

$$\frac{T_1}{N} - \frac{T_0}{N} = \frac{T_1 - T_0}{N}$$

When the outcome is death, the negative of the average time difference ($T_0/N - T1/N$) is often called the YLL.

Risk Ratio

Effect measures are most often calculated by taking ratios notably the rate and risk ratios terms RRs. The casual rate ratio can be expressed as:

$\frac{\frac{A_1}{T_1}}{\frac{A_0}{T_0}} = \frac{I_1}{I_0}$ where $I_1 = A_1/T_1$ is the IR|1 = exposed and 0 = unexposed, and the caused risk ratio can be expressed as: $\frac{\frac{A_1}{N}}{\frac{A_0}{N}} = \frac{A_1}{A_0} = \frac{R_1}{R_0}$.

Relative Excess

A RR of greater than 1, reflecting an average effect that is causal, can be expressed as an excess RR. The excess casual rate ratio is written as:

$$IR - 1 = \frac{I_1}{I_0} - 1 = \frac{I_1 - I_0}{I_0}$$ where $IR = I_1/I_0$ is the causal rate ratio.

The excess causal risk ratio can be expressed as:

$$RR - 1 = \frac{R_1}{R_0} - 1 = \frac{R_1 - R_0}{R_0}$$ where $RR = R_1/R_0$ is the casual risk ratio.

The excess rate can be expressed relative to I_1 *or* R_1, respectively, as:

$$1 - \frac{1}{\text{IR}} \text{ or } 1 - \frac{1}{\text{RR}}$$

The measures that arise from interchanging I_1 with I_0 and R_1 with R_0 in attributable fractions, also termed *preventable fractions*, can be interpreted easily. The expression fraction of the risk under nonexposure that could be prevented by exposure can be written as:

$$(R_0 - R_1)/R_0 = 1 - R_1/R_0 = 1 - RR$$

In vaccine studies, this measure is also known as the *vaccine efficacy*.

Null State and Counterfactual in Relation to Effect Measurement

When the occurrence measures being compared do not vary with exposure, then the measures of effect will equal 0 if expressed as a difference, or 1 if expressed as a ratio. A null effect or null state does not depend on the way an occurrence measure is compared. Counterfactual refers, as its name implies, to the effect measure contrary to fact such that if a cohort is exposed or treated, then the untreated state will be the counterfactual. The important feature of counterfactually defined effect measures is that they involve 2 distinct conditions: an index status, which usually involves some exposure or treatment, and a reference condition, such as no treatment against which the exposure or treatment will be evaluated. Although unapproachable in practical terms, the counterfactual can be thought of as the outcome if the exposed or treated group had not been exposed.

Prevalence Ratios

It was shown previously that the crude prevalence odds (PO) equals the crude IR, I, multiplied by the average disease duration, $\overline{D}$ when both the population at risk and the prevalence pool are stationary and there is no migration in or out of the it. Restating this relation between a single population under exposure and another unexposed, or separate exposed and unexposed populations:

$$\text{PO}_1 = \text{I}_1\overline{\text{D}}_1 \text{ and } \text{PO}_0 = \text{I}_0\overline{\text{D}}_0,$$

where subscripts 1 and 0 refer to exposed and unexposed respectively.

If the average disease duration is the same regardless of exposure $(\overline{\text{D}}_1 = \overline{\text{D}}_0)$, the crude prevalence OR (POR), will equal the crude IR:

$$\text{POR} = \frac{\text{PO}_1}{\text{PO}_2} = \frac{\text{I}_1}{\text{I}_0} = \text{IR}$$

EPIDEMIOLOGY ESTIMATION

Rothman and colleagues[17,18] review validity, precision, and statistics in epidemiologic studies discussed further.

Validity and Error

The epidemiologic estimate is the end product of the study design, the conducted study, and the data analysis. The goal of an epidemiologic study is to obtain a valid and precise estimate of the frequency of a disease or of the effect of an exposure on the occurrence of a disease in the source population under investigation. This

entails consideration of all the possible threats to internal and external validity. Taken to an extreme, however, epidemiologic studies designed to sample subjects from a target population of particular interest such that the study population is a probability sample from that population through oversampling subgroups in an effort to enhance *internal validity*, may fail to identify causal relationships, thereby limiting the external validity or generalizability of study. Most violations of internal validity can be classified into 3 general categories: confounding, selection bias, and information bias. Accuracy in estimation necessitates that statistical measures are estimated with little error, failures of which can be either random or systematic. Although random errors in the sampling and measurement of subjects can lead to systematic errors in the final estimates, important principles of study design emerge from separate considerations of sources of random and systematic errors. Random error or variation can detract from accuracy and has many components, but the major contributor is the process of selecting study subjects. Referred to as sampling error or variation, it factors more heavily in case control studies that involve a physical sampling process than it does in cohort studies. At least conceptually, the subjects in the study population selected to represent the population of broader interest may not satisfy the definition of a random sample for which strict statistical tools will be used to measure random variation. Other sources of random error include unexplained variation in occurrence measures such as observed IRs or prevalence proportions, mismeasurement of key study variables, and variance of the measurement or estimation process. Common approaches to reduce random error and to increase precision in epidemiologic estimation include increasing the study size, modification of the study design to increase precision, stratification of data to examine effects in subcategories, and significance and hypothesis testing using confidence intervals (CI) and interval estimation.

Systematic errors in epidemiologic estimation commonly referred to as *biases*, threaten the internal validity of a study. They are classified primarily into 3 general categories: confounding, selection bias, and information bias. Confounding occurs when the apparent effect of the exposure of interest is distorted because the effect of extraneous factors and is mistaken for, or mixed with, the actual exposure effect. The distortion introduced by a confounder may lead to overestimation or underestimation of an effect, toward or away from the null, depending on the direction of the association that the confounder has with exposure and outcome; or even change the apparent direction of an effect. By definition, confounders are extraneous risk factors for the outcome, associated with but not affected by the exposure or disease in the source population under study, and not an intermediate in the causal path between the exposure and outcome.

By contrast, selection biases are distortions resulting from procedures used to select subjects and from factors that influence study participation. The common element is that the relationship between exposure and outcome is different for those who participate, and all those who should be theoretically eligible for study, including those who do not participate. Bias in estimating an effect can be caused by measurement errors termed *information bias*, the direction and magnitude of which depends on whether the distribution of errors effect discrete variables with a countable number of possible values (misclassification errors), the values of 1 or more variables (nondifferential or differential misclassification), and its impact on binary variables. Misclassification can lead to alterations in the sensitivity or specificity of the measurement method. Although correctly classifying someone who is truly exposed as exposed, enhancing sensitivity, will be offset by falsely categorizing another unexposed and who is truly exposed; conversely, categorizing someone correctly as unexposed, strengthening specificity, will be lessened by misclassifying another as exposed.

Predictive probability positive is the probability that someone who is classified as exposed is truly exposed, whereas predictive probability negative is the probability that someone who is classified as unexposed is truly unexposed.

P Values

There are 2 types of *P* values: upper and lower, and accurate definitions of the associated statistics are often not considered rigorously in epidemiologic literature. An upper 1-tailed *P* value is the probability that a corresponding quantity computed from the data known as the test statistic, such as a *t*-test or χ^2 statistic, will be greater than or equal to its observed value, assuming that the test hypothesis is correct and there is no source of bias in the data collection or analysis. Similarly, a lower 1-tailed *P* value is the probability that the corresponding test statistic will be less than or equal to its observed value, again assuming that the test hypothesis is correct and that the underlying statistical model is correct. The 2-tailed *P* value, however, is usually regarded as twice the smaller of the upper and lower *P* values; however, being a probability, a 1-tailed *P* value must fall between 0 and 1, whereas the 2-tailed *P* value as defined, may exceed 1. Although equally regarded as a level of significance,[19] this term is usually avoided because it may be used in reference to alpha levels. In significance testing, small *P* values are supposed to indicate that at least 1 of the assumptions used to derive it is incorrect, but all too often, the statistical model is taken as a given so that a small *P* value is taken as indicating a low degree of compatibility between the test hypothesis and the observed data.

Hypothesis Testing

The use of *P* values and references to statistically significant findings highlights the dominant role that statistical hypothesis testing has occupied. Based on a value less than or greater than an arbitrary cutoff value, usually .05, called the *alpha* (α) level of the test, statistical significance testing of associations focuses on the null hypothesis, which is usually formulated as a hypothesis of no association between 2 variables from a population in which the observed study groups have been sampled in a random fashion. One may test for example the hypothesis that the risk difference in the population is 0 or the risk ratio is 1.0; alternatively, that the former is 0.1 and the latter equal to 2.0. The use of a fixed α cutoff is the hallmark of statistical hypothesis testing with both the alpha level and *P* value called the significance level of the test. This usage has led to misinterpretation of the *P* value as the alpha level of a statistical hypothesis test. A common misinterpretation of significance testing is to assert that there is no difference between 2 observed groups because the null test is not statistically significant and because the *P* value is greater than the cutoff for declaring statistical significance. Another misinterpretation of *P* values is that they represent probabilities of test hypotheses. A *P* value for a simple test hypothesis that exposure and disease are unassociated is not a probability of that hypothesis. The *P* value includes not only the probability of the observed data under the test hypothesis, but also the probabilities for all other possible data configurations. Although the *P* value is a continuous measure of the compatibility between a hypothesis and data, the alpha level is used to classify an observation as either significant, as when the $P \leq \alpha$, such that the test hypothesis is rejected, or not significant at the level α if $P > \alpha$, in which case the test hypothesis is accepted, or at least not rejected.

To avoid confusion, one should recall that the *P* value is a quantity computed from the data, whereas the alpha level is a fixed cutoff, usually 0.05, that can be specified without even seeing the data. Formal hypothesis testing avoids use of the *P* value in the formulation of hypothesis testing, instead defining the test based on whether the

value of the test statistic falls into a rejection region for the test statistic. An incorrect rejection is called a type 1 error or alpha error. A hypothesis testing procedure is said to be valid if, whenever the test hypothesis is true, the probability of rejection ($P \leq \alpha$) does not exceed the alpha level, provided there is no bias and the statistical model is correct. Hence, a valid test with $\alpha = 0.01$ (a 1% alpha level) will lead to a type 1 error with no more than 1% probability, provided there is no bias or incorrect assumption. However, if the test hypothesis is false but is not rejected, the incorrect decision not to reject is called a type II or beta error. If the test hypothesis is false, so that rejection is the correct decision, the probability that the test hypothesis is rejected is called the *power* of the test. The probability of a type II error is related to the power (*Pr*) by the equation: *Pr* (type II error) = 1 - *Pr*. The trade-off between the probabilities of type I and II errors depends on the alpha level chosen in that reducing the type I error when the test hypothesis is true requires a smaller alpha level, because a smaller *P* value will be required to reject the test hypothesis. Unfortunately, a lower alpha level increases the probability of a type II error if the test hypothesis is false while, increasing the alpha level reduces the probability of type II error when the test hypothesis is false, thus increasing the probability of type I error if it is true.

Confidence Intervals and Limits

Estimation measurement may benefit from more detailed statistics performed on a continuous scale with a theoretically infinite number of possible values than the simple dichotomy produced by statistical hypothesis testing of a simple parameter such as a risk or rate ratio, IR or another epidemiologic measure. Although one way to account for random error in the estimation process is to compute *P* values for a broad range of possible parameters in addition to the null value, if the range is broad enough, it is possible to calculate a confidence interval (CI) for which the *P* value exceeds a specific alpha level typically 0.05 as an example of interval estimation. The endpoints of the CI are termed *confidence limits* and the width of the depends on both the amount of random variability inherent in the data collection process and an arbitrary selected alpha level that specifies the degree compatibility between the limits of the interval and the data wherein one minus the alpha level (0.95 if alpha is 0.05) is called the confidence level of the CI and expressed as a percentage. Considering the relation of the CI to significance and hypothesis testing, consider a test of the null hypothesis with an $\alpha = 0.10$ If the 90% CI does not include the null point, then the null hypothesis would be rejected for $\alpha = 0.10$. On the other hand, if included in a 95% CI, then the null hypothesis would not be rejected for $\alpha = 0.05$. Because the 95% CI incudes the null point and the 90% did not, it can be inferred that the 2-sided *P* value for the null hypothesis is greater than .05 but less than .10. Thus, although a 2-sided *P* value instead indicates only the degree of consistency between the data and a single hypothesis, confidence limits provide an idea of the direction and magnitude of the underlying association as well as the random variability of point estimation.[20]

Because a given CI is only one of an infinite number of ranges nested within one another, points nearer to the center of the ranges are more compatible with the data than points distant from the center; thus, to see the entire set of possible CI, one constructs a *P value function*, comprising all points for which the 2-sided *P* value exceeds the alpha level of the CI. It summarizes the 2 key components of the estimation process. The peak of the curve indicates the point estimate and the concentration of the curve around the point estimate indicates the precision of the estimate. A narrow *P* value function would result from a large study with high precision, whereas a broad function would result from a small study with low precision. A CI represents only 1

possible horizontal slide through the *P* value function. A *P* value function from which one can find confidence limits for a hypothetical study with a rate ratio estimate of 3.1, has a curve that reaches its peak corresponding to the point estimate for the rate ratio; while the 95% CI can be read directly from the graph as the function values where the right-hand ordinate is 0.95. The *P* value for any value of the parameter can be found from the left-hand ordinate corresponding to the height where the vertical line drawn at the hypothesized rate ratio that equals 1 intersects the *P* value function. Likelihood intervals, likelihood functions, and natural logarithms of the likelihood function or log-likelihood, which are beyond the scope of this article can be found in advanced epidemiology texts, have sought to replace CI.[21–23]

EPIDEMIOLOGIC STUDY DESIGN

Epidemiologic study designs comprise both experimental and nonexperimental types. The term *experimental* implies that the investigator manipulates the exposure assigned to participants in the study. In a randomized clinical trial (RCT), the gold standard for medical investigation, the investigator creates groups through random allocation of an exposure or treatment. However, when experiments are infeasible or unethical, epidemiologists design nonexperimental or observational studies that simulate what might have occurred had an experiment been conducted. In that regard, the researcher is an observer rather than an agent who assigned interventions. Rothman and colleagues[24–26] review the major types of epidemiologic studies further explained below.

There are 4 types of nonexperimental studies: cohort, case control, cross-sectional and ecological types. In cohort studies, the subjects of a source population are classified according to their exposure status and followed over time to ascertain disease incidence. In case control studies, cases arising from a source population and a sample of the source population are classified according to their exposure history. Cross-sectional studies include as subjects all persons in the population at the time of ascertainment or a representative sample, selected without regard to exposure or outcome status, usually to estimate prevalence. In ecological studies, the unit of observation is a group of people rather than an individual.

Cohort Studies

In principle, cohort studies can be used to estimate average risks, rates, and occurrence times, but to do so the entire cohort remains at risk and under observation for the entire follow-up period. Such measurements, however, are only feasible when there is little or no loss to follow-up. When losses and competing risks occur, IRs can be estimated directly, whereas the average risk and occurrence time can be estimated using basic survival analysis involving stratification on follow-up time using life-table analysis methods. The main guide to the classification of persons or person-time should be defined explicitly according to the study hypothesis and design to estimate appropriately the exposure effects and avoid implicit assumptions. Chronic exposures based on anticipated effects is more complicated than when exposure occurs only at a point in time, which can be conceptualized as a period during which the exposure accumulates to a sufficient extent to trigger a step in the causal process. The time at which an outcome event occurs can be a major determinant of the amount of person-time contributed by a subject to each exposure category. The method of calculation for cumulative incidence and IR in cohort studies is shown in the example below.

In a study, a researcher observed 2000 people in total with 1000 in each exposure status for 3 years to see whether they develop a disease. The result is shown in Table 1.

$$I = \frac{\text{number of new cases of a disease}}{\text{number of candidate population}}$$

$I_{exp} = 36/1000 = 0.036$, $I_{nonexp} = 10/1000 = 0.010$,
risk ratio $= I_{exp}/I_{nonexp} = 0.036/0.010 = 3.6$.

$$\text{Incidence rate} = \frac{\text{number of disease onsets}}{\sum_{persons}\text{time spent in population}}$$

$$IR_{exp} = \frac{\text{number of disease onsets}}{\sum_{persons}\text{time spent in population}} = \frac{36}{0.5\times 4+1\times 5+1.5\times 8+2\times 6+2.5\times 4+3\times 973} = \frac{36}{2960\ \text{person}-\text{year}}$$

$$IR_{nonexp} = \frac{\text{number of disease onsets}}{\sum_{persons}\text{time spent in population}} = \frac{10}{0.5\times 2+1\times 1+1.5\times 1+2\times 3+2.5\times 1+3\times 992} = \frac{10}{2988\ \text{person}-\text{year}}$$

Rate ratio $= IR_{exp}/IR_{nonexp} = 3.63$.

It is important to define and determine the time of the event in cohort studies as unambiguously and precisely as possible, incorporating the details of available data and the current state of knowledge about the study outcome. Cohort studies are expensive to conduct because stable estimates of incidence require a substantial number of cases of disease, and therefore person-time giving rise to the cases will be substantial. When studying a rare disease or one that has a long latency for development, especially when cost is a factor, a case control study design is preferable.

Table 1
Incidence and incidence rates

Years of Observation	No. of Cases in Exposed (1000)	No. of Cases in Nonexposed (1000)
0.5	4	2
1.0	5	1
1.5	8	1
2.0	6	3
2.5	4	1
3.0	9	2

Case Control Studies

The use and understanding of case control studies was an important methodologic advance in modern epidemiology as the field advanced from randomized to nonrandomized cohort studies to case control studies. Although conventional wisdom holds that cohort studies are useful for evaluating the range of effects related to a single exposure, case control studies nested within a single population using several disease outcomes as the case series is possible with a case cohort study design. Recognizing that case control studies are most practical for rare diseases when exposure is rare, ordinary case control studies may likewise be inefficient unless selective recruitment of additional exposed subjects is performed. Ideally, a case control study should be conceptualized as a more efficient version of a corresponding cohort study. Rather than including all of the experiences of the source population that gave rise to cases, as would be the practice in a cohort study design, controls are selected from the source population. Therein lies the challenge of achieving random sampling of controls from the source population, and when it is not possible to identify explicitly the source population, and simple random sampling is not possible, secondary source populations become an option using neighborhood controls, random digit dialing, hospital- or clinic-based controls, and even friends with close attention to representativeness, comparability of information accuracy, and the number of control groups needed.

Recognizing that the primary goal for control selection is that the exposure distribution among controls is such that it is the same as in the source population of cases, OR calculations achieve this goal by using control cases in place of the denominator in measures of disease frequency to determine the ratio of the disease frequency in exposed relative to unexposed people. Using person-time to illustrate, the goal requires that exposed controls (B_1) has the same ratio to the amount of exposed person-time (T_1) as unexposed controls (B_0) have to the amount of unexposed person-time (T_0), apart from sampling error to compute control sampling rates:

$$\frac{B_1}{T_1} = \frac{B_0}{T_0}$$

If A_1 is the exposed cases and A_0 unexposed cases over a study period, the exposed and unexposed IR are computed as:

$$I_1 = \frac{A_1}{T_1} \text{ and } I_0 = \frac{A_0}{T_0}$$

Using the denominators of the frequencies of the exposed and unexposed controls as substitutes for the actual denominators of the rates to obtain exposure-specific case control rates or pseudo-rates, those rates can be rewritten as:

$$\text{Pseudo-rate}_1 = \frac{A_1}{B_1} \text{ and Pseudo-rate}_0 = \frac{A_0}{B_0}$$

By dividing the pseudorate for exposed by the pseudo-rate for unexposed, we obtain an estimate of the ratio of the IRs in the source population provided that the control sampling is independent of exposure.

The ratio of the 2 pseudo-rates in a case control study also known as the cross-product ratio or OR is written as:

$$OR = A_1B_0/A_0B_1$$

It can be viewed as the ratio of cases to controls among the exposed subjects (A_1/B_1) divided by the ratio of cases to controls among the unexposed subjects (A_0/B_0) or as the odds of being exposed among cases (A_1/A_0) divided by the odds of being exposed among controls (B_1/B_0), in which case it is termed the *exposure OR*. Although either interpretation gives the same result, viewing this OR as the ratio of case control ratios shows more directly how the control group substitutes for the denominator information in a cohort study and how the ratio of pseudo-frequencies gives the same result as the ratio of IR, incidence proportion, or incidence odds in the source population, if sampling is independent of exposure. Further, it is not necessary to assume that the disease under study is rare.

Cross-Sectional Studies

Cross-sectional studies investigate the association between prevalence of diseases or mortality and prevalence of risk factors in a defined population at a certain time point. All the information is collected at the same time. Cross-sectional studies conducted to estimate prevalence are termed a *prevalence study*. Exposure is ascertained simultaneously with the disease and different exposure subpopulations may be compared with respect to their disease prevalence. Two potential limitations in cross-sectional studies are determining the time order of events and overrepresentation of cases with long duration and underrepresentation with short durations of illness, termed *length-biased sampling*. Because prevalence depends on incidence and duration, a high prevalence of disease or mortality may result from long duration of time despite a low incidence. Because cross-sectional studies investigate the status at one specific point in time, it may not be possible to determine whether the risk factors happened before disease. Cross-sectional studies may involve sampling subjects differentially with respect to disease status to increase the number of cases in the sample. Such studies termed *prevalent case control studies* are designed similar to incident case control studies except that the case series comprises prevalent rather than incident cases. Although significant associations may be found in a given cross-sectional study, it may not reflect the true causation.

Ecological Studies

Ecological studies focus on the association between the summary measure of disease or mortality and risk factors, and the unit of ecological studies is a group not an individual. The groups can be countries, states, regions, schools, or zip codes, among others. Ecological measures can be classified into aggregate, environmental, and global measures. *Aggregate measures* reflect characteristics of individuals within a group and *environmental measures* represent physical characteristics of the geographic location; *global measures* may be characteristics of the group or place without analogy to the individual. Such studies are able to examine a broad range of diseases and risk factors using demographic and consumption data, and are very useful in generating hypotheses on association in advance of epidemiologic studies on individual observations. In addition, ecological studies have other advantages, such as low cost and the convenience link of aggregate data, freedom from the measurement and design limitations of individual-level studies, and simplicity of analysis and presentation. For some risk factors, aggregate measurements may be more accurate than individual measurements. Data collection, disease definition, and treatment may vary across units, and this can introduce bias. A major limitation of ecological studies is bias, which can be interpreted as the failure of associations seen at one level of grouping to correspond with effect measures at the grouping level

of interest. In other words, the association between risk factor and disease found on the group level may not hold true within individuals.

REGRESSION TECHNIQUES

Szklo and Nieto[6] review regression techniques.

Simple Linear Regression

The process of determining whether the relationship between the 2 variables is compatible with straight line begins with the visual inspection of a scatter plot followed by the calculation of the liner correlation coefficient, *r*. With a range of −1.0 to 1.0, inscribing perfectly straight lines of negative and positive 1.0 slopes, a value of 0 indicates instead no linear correlation. The correlation coefficient value contains no information about the strength of the association between the 2 variables that is represented by the slope of the theoretic line it inscribes that is further defined by 2 other parameters, β_0 and β_1, respectively, the *y* intercept when $x = 0$, and the regression coefficient as shown in the formula below:

$$y = \beta_0 + \beta_1 x$$

In linear regression, the method used to estimate the values of the regression coefficients is the least-squares method, which consists in finding the parameter values that minimize the sum of the squares of the vertical distance between each of the observed points and the line.[27,28] The notation traditionally used to represent the estimated linear regression line is as follows:

$$y = b_0 + b_1 x$$

The regression coefficient (b_1) estimates the average increase in the dependent variable per unit increase in the independent variable and like any other statistical estimate, it is subject to uncertainty and random error. Thus, it is important to estimate the standard error of the regression coefficient to evaluate its statistical significance and to calculate the confidence limits around its point estimate. The standard error estimate is provided readily by most statistical packages performing linear regression.

Multiple Linear Regression

Simple linear regression can be extended to multivariate regression using the same model adjusted when the outcome is a continuous variable as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \text{ or } y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k$$

The postulated risk factors (*x* or independent variables) can be continuous or categorical (dichotomous) with multiple levels that can be treated as ordinal or transformed in a set of binary variables. The estimated values of the regression coefficients are obtained by the least-squares method. An important assumption in the model is that there is no interaction between the variables in the model such that the change in *y* associated with a unit change in *x* for the entire range of *x* and vice versa. The regression coefficient (β_k) represents the average increase in outcome per unit increase in x_k, adjusted for all the other variables in the model. If interaction is present, stratified models can be used for each variable. Alternatively,

interaction terms, known as *product terms*, can be used in the regression equation. The multivariate model enables one to study the effect of main exposure while adjusting confounders, mediators, and interaction at the same time. The traditional way, such as restriction or stratification, is limited if there are many confounders. The model can adjust both categorical and continuous variable as demonstrated in the study by Weinstein and colleagues[29] of the association between clinical stroke and subsequent cognitive function in initially nondemented individuals. Outcome can be expressed in natural log-transformed cognitive scores, as necessary, to reduce skewness. The primary independent variable was stroke status. The unadjusted model was expressed as: log-transformed cognitive scores = $b_0 + b_1 \times stroke$. Model 1 was adjusted for age, sex, education level, cohort (original or offspring), and the Mini-Mental State Examination (MMSE) score. Model 2 was further adjusted for systolic blood pressure (SBP), diabetes, prevalent cardiovascular disease (CVD), prevalent atrial fibrillation and current smoking.

Model 1: log-transformed cognitive scores = $b_0 + b_1 \times stroke + b_2 \times age + b_3 \times sex + b_4 \times education + b_5 \times cohort + b_6 \times MMSE\ score$; for model 2, add $b_7 \times SBP + b_8 \times diabetes + b_9 \times CVD + b_{10} \times atrial\ fibrillation + b_{11} \times smoke$. There is a list of 1 domain of cognitive scores as an example in **Table 2**

Multiple Logistic Regression

For binary outcome variables, the logistic regression model offers a more robust alternative to binary multiple linear regression. The logistic regression model assumes that the relationship between a given value of a variable *x* and the probability of a binary outcome follows the logistic function:

$$P(y|x) = \frac{1}{1+e^{-(b_0+b_1x)}}$$

where $P(y|x)$ denotes the probability (P) of the binary outcome (y) for a given value of x. The outcome of this equation, a probability, is constrained to values within the range of 0 to 1. By translating instead into OR estimation, the probability equation can be expressed in the equivalent equation:

$$\text{Log}\left(\frac{P}{1-P}\right) = \log(\text{odds}) = b_0 + b_1x, \text{ where } P \text{ is the short notation for } P(y|x).$$

This expression is analogous to the simple linear regression function, except that the ordinate in now the logarithm of the odds or log odds, also known as *logit*, rather than the usual mean value of a continuous variable. Thus, if the relationship between exposure (x) and the occurrence of an outcome is assumed to fit the logistic regression model, that implies that the log odds of the outcome increases linearly with x. The multiple logistic regression model is shown as:

Table 2
Association of clinical stroke with cognitive performance

	Model 1		Model 2	
Outcome	**b_1 ± SE**	***P* Value**	**b_1 ± SE**	***P* Value**
LMi	−1.35 ± 0.52	.010	−1.27 ± 0.60	.035

Abbreviations: LMi, logical memory-immediate recall; SE, standard error.

$$\text{Log}\left(\frac{P}{1-P}\right) = \log(\text{odds}) = b_0 + b_1x_1 + b_2x_2 + \ldots b_kx_k,$$ [where the regression coefficient (b_k) is the average increase in the log odds of the outcome per unit increase in x_k, adjusted for all other variables in the model].

The OR corresponding to a unit increase in the independent variable is the antilogarithm or exponential function of the regression coefficient b_1 as follows:

$$\text{OR} = e^{b1}$$

Thus, the logistic regression model is a linear model in the log odds scale. What this means in practical terms is that, when a continuous variable is entered as such, the resulting coefficient and corresponding OR is assumed to represent the linear increase in log odds or the exponential increase in odds, per unit increase in the independent variable across the entire range of x values. Davydow and colleagues[30] studied whether depression, cognitive impairment without dementia (CIND), and/or dementia were each independently associated with risk of ischemic stroke. The outcome of interest was ischemic stroke. The primary independent variable was depression, CIND or dementia status at baseline defined categorically as no disorder, depression alone, CIND alone, dementia alone, cooccurring depression and CIND, or cooccurring depression and dementia. In unadjusted analysis, the model could be written as:

$$\text{Log}\left(\frac{P(\text{stroke})}{1-P(\text{stroke})}\right) = b_0 + b_1 \times \text{baseline status}.$$

Demographics, medical comorbidities, and health risk behaviors were treated as the possible characteristics to modify the association. In adjusted analysis the multiple logistic model is:

$$\text{Log}\left(\frac{P(\text{stroke})}{1-P(\text{stroke})}\right) = b_0 + b_1 \times \text{baseline status} + b_2 \times \text{demographics} + b_3 \times \text{comorbidities} + b_4 \times \text{health} - \text{risk behaviors}.$$

The OR are shown in **Table 3**.

The adjusted OR resulting from the exponentiation of the logistic regression coefficient obtained is often used as a surrogate of the RR or prevalence rate ratio, respectively. This interpretation is only justified for the analyses of rare outcomes, but when the frequency of the outcome of interest in high, the OR is a biased estimate of the RR

Table 3
Multiple logistic model

	Unadjusted	Adjusted
Baseline Status	**OR (95% CI)**	**OR (95% CI)**
Depression alone	1.11 (0.88, 1.40)	1.09 (0.85, 1.38)
CIND alone	1.55 (1.27, 1.90)	1.37 (1.11, 1.69)
Dementia alone	1.36 (1.04, 1.77)	1.08 (0.81, 1.44)
Cooccurring depression and CIND	1.95 (1.48, 2.56)	1.65 (1.24, 2.18)
Cooccurring depression and dementia	1.51 (1.09, 2.10)	1.16 (0.82, 1.65)

Abbreviations: CI, confidence interval; CIND, cognitive impairment without dementia; OR, odds ratio.

or the prevalence rate ratio, because it tends to exaggerate the magnitude of the association. Thus, it is important to keep in mind the built-in bias associated with the OR as an estimate of the incidence or prevalence rate ratio when the outcome is common. An alternatives regression procedures to consider the log-binomial regression model, which results in direct estimates of the incidence or prevalence rate ratio.[31]

ACKNOWLEDGMENTS

The author (D.S. Younger) wishes to acknowledge the faculty of the Columbia University Mailman School of Public Health, New York, NY, who provided him the expertise in biostatistics and epidemiology to write this article and to provide it as a resource to neurology trainees and colleagues.

REFERENCES

1. Agresti A. An introduction to categorical data analysis. 2nd edition. Hoboken (NJ): John Wiley & Sons; 2007.
2. Aschergrau A, Seage GR III, editors. Essentials of epidemiology in public health. 2nd edition. Boston: Jones and Bartlett Publishers; 2008.
3. Hoffmann JP. Generalized linear models. An applied approach. New York: Pearson; 2004.
4. Hosmer DW Jr, Lemeshow S, Sturdivant RX. Applied logistic regression. 3rd edition. Hoboken (NJ): John Wiley & Sons; 2013.
5. Kleinbaum DG, Kupper LL, Nizam A, et al. Applied regression analysis and other multivariable methods. 4th edition. Belmont (CA): Thomson Brooks/Cole; 2008.
6. Szklo M, Nieto FJ. Epidemiology. Beyond the basics. 3rd edition. Burlington (MA): Jones & Bartlett Learning; 2014.
7. Rosner B. Fundamentals of biostatistics. 7th edition. Boston (MA): Brooks/Cole Cengage Learning; 2006.
8. Rothman KJ, Greenland S, Lash TL, editors. Modern epidemiology. 3rd edition. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2008.
9. Greenland S, Rothman KJ. Measures of occurrence. Chapter 3. In: Rothman KJ, Greenland S, Lash TL, editors. Modern epidemiology. 3rd edition. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2008. p. 32–50.
10. Dempsey M. Decline in tuberculosis. The death rate fails to tell the entire story. Am Rev Tuberc 1947;56:157–64.
11. National Center for Health Statistics. Health, United States, 2004. Hyattsville (MD): Department of Health and Human Services, National Center for Health Statistics; 2004. Available at: http://www.cdc.gov/nchs/hus.htm.
12. Murray CJ, Acharya AK. Understanding DALYs (disability-adjusted life years). J Health Econ 1997;16:703–30.
13. GBD 2013 DALYs and HALE Collaborators, Murray CJ, Barber RM, Foreman KJ, et al. Global, regional, and national disability-adjusted life years (DALYs) for 306 diseases and injuries and healthy life expectancy (HALE) for 188 countries, 1990-2013: quantifying the epidemiological transition. Lancet 2015;386:2145–91.
14. Global Burden of Disease Study 2013 Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990-2013: a systematic analysis for the Global Burden of Disease Study. Lancet 2013;2015(386):743–800.
15. Vos T, Flaxman AD, Naghavi M, et al. Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet 2012;380:2163–96.

16. Greenland S, Rothman KJ, Lash TL. Measures of effect and measures of association. Chapter 4. In: Rothman KJ, Greenland S, Lash TL, editors. Modern epidemiology. 3rd edition. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2008. p. 51–70.
17. Rothman KJ, Greenland S, Lash TL. Validity in epidemiologic studies. In: Rothman KJ, Greenland S, Lash TL, editors. Modern epidemiology. 3rd edition. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2008. p. 128–47.
18. Rothman KJ, Greenland S, Lash TL. Precision and statistics in epidemiologic studies. In: Rothman KJ, Greenland S, Lash TL, editors. Modern epidemiology. 3rd edition. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2008. p. 148–67.
19. Cox DR, Hinkley DV. Theoretical statistics. New York: Chapman and Hall; 1974.
20. Bandt CL, Boen JR. A prevalent misconception about sample size, statistical significance, and clinical importance. J Periodontol 1972;43:181–3.
21. Goodman SN, Royall R. Evidence and scientific research. Am J Public Health 1988;78:1568–74.
22. Edwards AWF. Likelihood. 2nd edition. Baltimore (MD): Johns Hopkins University Press; 1992.
23. Royall R. Statistical inference: a likelihood paradigm. New York: Chapman and Hall; 1997.
24. Rothman KJ, Greenland S, Lash TL. Types of epidemiologic studies. In: Rothman KJ, Greenland S, Lash TL, editors. Modern epidemiology. 3rd edition. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2008. p. 87–99.
25. Rothman KJ, Greenland S. Cohort studies. In: Rothman KJ, Greenland S, Lash TL, editors. Modern epidemiology. 3rd edition. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2008. p. 100–10.
26. Rothman KJ, Greenland S, Lash TL. Case-control studies. In: Rothman KJ, Greenland S, Lash TL, editors. Modern epidemiology. 3rd edition. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2008. p. 111–27.
27. Armitage P, Berry G, Matthews JNS. Statistical methods in medical research. 4th edition. Oxford (United Kingdom): Blackwell Publishing; 2002.
28. Draper N, Smith H. Applied regression analysis. 3rd edition. New York: John Wiley & Sons; 1998.
29. Weinstein G, Preis SR, Beiser AS, et al. Cognitive performance after stroke-The Framingham Heart Study. Int J Stroke 2014;9:48–54.
30. Davydow DS, Levine DA, Zivin K, et al. The association of depression, cognitive impairment without dementia and dementia with risk of ischemic stroke: a cohort study. Psychosom Med 2015;77:200–8.
31. Spiegelman D, Hertzmark E. Easy SAS calculations for risk or prevalence ratios and differences. Am J Epidemiol 2005;162:199–200.