

Research Techniques Made Simple: Interpreting Measures of Association in Clinical Research



Michelle R. Roberts^{1,2}, Sepideh Ashrafzadeh^{1,2} and Maryam M. Asgari^{1,2}

To bring evidence-based improvements in medicine and health care delivery to clinical practice, health care providers must know how to interpret clinical research findings and critically evaluate the strength of evidence. This requires an understanding of differences in clinical study designs and the various statistical methods used to identify associations. We aim to provide a foundation for understanding the common measures of association used in epidemiologic studies to quantify relationships between exposures and outcomes, including relative risks, odds ratios, and hazard ratios. We also provide a framework for critically assessing clinical research findings and highlight specific methodologic concerns.

Journal of Investigative Dermatology (2019) 139, 502–511; doi:10.1016/j.jid.2018.12.023

CME Activity Dates: 20 February 2019

Expiration Date: 19 February 2020

Estimated Time to Complete: 1 hour

Planning Committee/Speaker Disclosure: Maryam Asgari, MD receives research support from Valeant Pharmaceuticals and Pfizer, Inc. All other authors, planning committee members, CME committee members and staff involved with this activity as content validation reviewers have no financial relationships with commercial interests to disclose relative to the content of this CME activity.

Commercial Support Acknowledgment: This CME activity is supported by an educational grant from Lilly USA, LLC.

Description: This article, designed for dermatologists, residents, fellows, and related healthcare providers, seeks to reduce the growing divide between dermatology clinical practice and the basic science/current research methodologies on which many diagnostic and therapeutic advances are built.

Objectives: At the conclusion of this activity, learners should be better able to:

- Recognize the newest techniques in biomedical research.
- Describe how these techniques can be utilized and their limitations.
- Describe the potential impact of these techniques.

CME Accreditation and Credit Designation: This activity has been planned and implemented in accordance with the accreditation requirements and policies of the Accreditation Council for Continuing Medical Education through the joint providership of Beaumont Health and the Society for Investigative Dermatology. Beaumont Health is accredited by the ACCME to provide continuing medical education for physicians. Beaumont Health designates this enduring material for a maximum of 1.0 *AMA PRA Category 1 Credit(s)*[™]. Physicians should claim only the credit commensurate with the extent of their participation in the activity.

Method of Physician Participation in Learning Process: The content can be read from the *Journal of Investigative Dermatology* website: <http://www.jidonline.org/current>. Tests for CME credits may only be submitted online at <https://beaumont.cloud-cme.com/RTMS-Mar19> — click ‘CME on Demand’ and locate the article to complete the test. Fax or other copies will not be accepted. To receive credits, learners must review the CME accreditation information; view the entire article, complete the post-test with a minimum performance level of 60%; and complete the online evaluation form in order to claim CME credit. The CME credit code for this activity is: 21310. For questions about CME credit email cme@beaumont.edu.

INTRODUCTION

Epidemiology is the study of the distribution and determinants of disease and other health-related outcomes within populations. As the basic science of public health, epidemiologic studies can describe patterns of disease within specific populations (descriptive epidemiology) or investigate etiology and risk factors for health outcomes (analytic epidemiology). A core feature of analytic epidemiology is the presence of an appropriate comparison group. Using analytic epidemiologic methods, we can

investigate hypotheses about exposure-outcome relationships by comparing exposure status between groups of people. A sound understanding of epidemiologic principles enables health care providers to consider if the effects of an exposure could warrant changes in clinical practice, treatment protocols, or community program management. In this article, we describe several measures of association frequently encountered in analytic epidemiology and discuss factors to consider when interpreting clinical research.

¹Department of Dermatology, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA; and ²Department of Population Medicine, Harvard Pilgrim Healthcare Institute, Boston, Massachusetts, USA

Correspondence: Maryam M. Asgari, Department of Dermatology, Massachusetts General Hospital, 50 Staniford Street, Suite 230A, Boston, Massachusetts 02114, USA. E-mail: PORES@mgh.harvard.edu

Abbreviations: CI, confidence interval; HR, hazard ratio; I, incidence; O, odds; OR, odds ratio; RR, relative risk

SUMMARY POINTS

- *Measures of association* refers to a wide variety of statistics that quantify the strength and direction of the relationship between exposure and outcome variables, enabling comparison between different groups.
- The measure calculated depends on the study design used to collect data. Odds ratios should be used for case-control and cross-sectional studies, whereas relative risk should be used in cohort studies.
- When interpreting measures of association in clinical practice, consider whether the results may have been affected by sources of bias and confounding, as well as how generalizable the study sample is to the target population.
- Confounding may be addressed through randomization, matching, stratification, or statistical adjustment, although unmeasured confounders or residual confounding may still affect the observed association.
- Effect sizes and measures of variability, such as confidence intervals, may be more informative than *P*-values for interpreting epidemiologic data.

MEASURES OF ASSOCIATION

Epidemiologic study designs are differentiated by the presence or absence of an intervention, randomization of participants, and the temporal relationships among comparison groups. Common observational designs, including cohort, case-control, and cross-sectional studies, are shown in Table 1 (Besen and Gan, 2014; Silverberg, 2015).

Relationships between exposures and outcomes are quantified using various measures of association, which are statistics that estimate the direction and magnitude of associations among variables. Commonly used measures are described in Table 2 and Figure 1. The reported measure of association depends on the study design used to collect the data and the statistical method used to analyze it (Pearce, 1993). A useful way to visualize the calculation of several measures of association is by constructing a basic 2×2 contingency table (Figure 2), which shows the cross-tabulation of exposed and unexposed participants (rows) by those with and without an outcome of interest (columns).

Relative risk

Relative risk (RR) is often calculated in cohort studies, where participants with and without exposure(s) are followed for particular outcome(s). This design allows for the calculation of incidence (*I*), found by dividing the number of new cases of an outcome by the number of people at risk for the outcome during a specified period (Figure 2): $I_{\text{exposed}} = A/(A + B)$ and $I_{\text{unexposed}} = C/(C + D)$. The RR is the ratio of the incidence among exposed participants to the incidence among

unexposed participants: $RR = I_{\text{exposed}}/I_{\text{unexposed}}$. By comparing incidence rates between the exposed and unexposed groups, it is possible to determine if an exposure increases or decreases risk of an outcome.

When RR is equal to 1, the incidence is the same among those exposed and unexposed. An RR less than 1 suggests that the exposure is protective ($I_{\text{exposed}} < I_{\text{unexposed}}$), and an RR greater than 1 suggests that the exposure is a risk factor for the outcome ($I_{\text{exposed}} > I_{\text{unexposed}}$). For example, the relationship between dietary vitamin D intake and risk of melanoma was investigated in a cohort study, and a RR of 1.31 (95% confidence interval [CI] = 0.94–1.82) was observed for the highest quartile of vitamin D compared with the lowest quartile (Asgari et al., 2009b). The point estimate indicates a 31% increased risk of melanoma (or 1.31 times the risk) among participants with the highest level of vitamin D intake, but because the CI includes the null value of 1, we would not consider the finding statistically significant.

Odds ratio

In case-control or cross-sectional studies, where we cannot calculate incidence rates, the odds ratio (OR) is typically calculated. The OR is the ratio of the exposure odds (*O*) among the case group to the exposure odds among the control group (Figure 2): $O_{\text{case}} = A/C$, $O_{\text{control}} = B/D$, $OR = O_{\text{case}}/O_{\text{control}}$, and it is interpreted similarly to the RR. An OR equal to 1 indicates no association, an OR less than 1 suggests that the exposure is protective (exposure is less likely among the case group), and OR greater than 1 suggests that the exposure is a risk factor (exposure is less likely among the control group). For example, in a case-control study examining the association between infection with human papillomavirus β and risk of squamous cell carcinoma, an OR of 4.0 (95% CI = 1.3–12.0) was observed (Asgari et al., 2008). This OR indicates that the odds of being exposed (i.e., having this human papillomavirus subtype) were 4 times greater among the case group than the control group or, put another way, that cases were 4 times more likely to have this human papillomavirus subtype than controls.

When the outcome is rare, the OR approximates the RR. This assumption, known as the *rare disease assumption*, can be visualized in Figure 2. When the proportions in cells A and C are small, $A + B \approx B$ and $C + D \approx D$. Therefore, $RR = [A/(A + B)]/[C/(C + D)] \approx (A/B)/(C/D) = (A/C)/(B/D) = OR$. When the outcome is more common (>10%), however, the OR provides more extreme estimates than the RR. In Figure 2, where 44% of the study population has the outcome, the OR is much smaller than the RR.

Hazard ratio

The hazard ratio (HR) is the ratio of the rate at which the exposed group experiences an outcome to the rate at which the unexposed group experiences an outcome, and it provides the instantaneous risk at a given time rather than the cumulative risk over the length of a study. It is calculated in survival or time-to-event analyses, in which the outcome variable is the time (days, months, years, etc.) until the occurrence of the event of interest, such as development of a disease, disease complication (e.g., cancer recurrence), death, or other outcome. Participants who do not experience

Table 1. Study designs in clinical research¹

Study Design	Description	Strengths/Utility	Weaknesses
Meta-analysis	Analysis in which multiple RCTs and/or observational studies are combined	Larger sample size and higher statistical power than individual studies	Limited by the quality and potential heterogeneity of the individual studies they combine
Experimental studies			
Randomized controlled trial	Prospective design in which participants are randomly allocated to intervention and control groups Control group may be placebo or a comparison intervention	Random assignment balances confounding variables between groups (even unmeasured variables) Identification of causality between an exposure/intervention and outcome	Expensive May not capture etiologically relevant time period Potential lack of generalizability Differential loss to follow-up may introduce bias Potential ethical issues
Quasi-experimental	Nonrandomized intervention study	Can assess the effects of an intervention Useful when randomization is not possible for practical or ethical reasons	Lack of random assignment Potential loss of internal validity
Observational studies			
Cohort	Longitudinal design in which participants are followed up over time May be prospective or retrospective	Possible to evaluate multiple exposures and outcomes in the same study population Temporal sequence of events is more clearly indicated Permits the calculation of disease incidence Facilitates examination of rare exposures Reduces the potential for selection bias at enrollment	Expensive and time consuming May be inefficient for rare outcomes or diseases with long latent periods Differential loss to follow-up may introduce bias For retrospective designs: May be difficult to identify appropriate exposed cohort and comparison group Data on important confounding variables may be absent Potential for reduction in data quality if records not designed for the study are used
Case-control	Design in which participants with an outcome (case group) and participants without the outcome (control group) are sampled from a defined source population and compared with respect to the frequency of one or more exposures May be prospective or retrospective May be nested within a cohort study	Facilitates the study of rare diseases/outcomes or those with long latency periods Less expensive and time consuming than cohort designs More efficient when exposure data are expensive or difficult to obtain Advantageous for dynamic populations in which long-term follow-up may be difficult	Inefficient for rare exposures Do not permit the calculation of disease incidence May be subject to selection bias, particularly due to nonrepresentative sampling of control individuals More susceptible to information biases, including recall and observer biases May be more difficult to establish temporality
Cross-sectional	Descriptive design in which data are collected from a population at a specific point in time Provides a “snapshot” of exposures and outcomes	Inexpensive and less time-consuming than other designs Can estimate prevalence of exposures and outcomes simultaneously Useful for monitoring health status and needs of a particular population	Temporality is difficult to ascertain Tends to identify prevalent cases of long disease duration (e.g., more serious cases may not be captured because of death) Potential for nonresponse bias
Ecologic	Design in which data are collected at the population, rather than individual, level Populations may be defined geographically or temporally	Useful for examining rare diseases Inexpensive and easy to conduct using routinely collected data Useful for monitoring population health, making comparisons between populations, or when individual-level data are unavailable	Prone to bias and confounding, both within and between groups The <i>ecologic fallacy</i> , in which effects observed at the population level do not accurately reflect effects at the individual level Methodologic weaknesses limit causal inference
Case study or series	A descriptive analysis of an individual case or series of cases, with no comparison group	Can describe new trends or rare characteristics of diseases May detect previously unreported adverse effects or potential new uses of medications Useful in teaching clinical lessons learned from patient care	May lack generalizability Potential confounding may not be addressed Difficult to establish causality

Abbreviation: RCT, randomized controlled trial.

¹This table lists advantages and disadvantages common to clinical study designs but is not exhaustive. Readers are referred to the many excellent published reviews of epidemiologic study design principles, including [Besen and Gan \(2014\)](#) and [Silverberg \(2015\)](#).

an event during the follow-up period are censored. This occurs if the participant is lost to follow-up, the follow-up period ends and the participant is event-free, or the participant experiences another outcome. At the time of censoring, the participant stops contributing follow-up time to the analysis.

This type of censoring is known as *right-censoring*, because the true unobserved event lies to the right of the censoring time. For example, in a survival analysis of acral lentiginous melanoma, both melanoma-specific survival and overall survival, or all-cause mortality, were examined. In the

Table 2. Examples of measures of association in clinical research

Measure of Association	Definition	Exposure	Outcome	Effect Estimate	Interpretation
Relative risk (RR) ¹	The ratio of the incidence in the exposed group to the incidence in the unexposed group	Vitamin D intake	Melanoma	RR = 1.31 (95% CI = 0.94–1.82)	When compared with the lowest quartile of dietary vitamin D intake, participants with the highest quartile of intake had 1.31 times the risk of melanoma. This may also be phrased as having a 31% increase in melanoma risk. Because the 95% CI includes 1 (the null value, indicating no association between exposure and outcome), the results are not statistically significant (Asgari et al., 2009b).
Odds ratio (OR)	The ratio of the exposure odds among the case group to the exposure odds among the control group	Presence or absence of HPV	Squamous cell carcinoma	Any HPV species: OR = 0.9 (95% CI = 0.4–1.8) HPV β -papillomavirus: OR = 4.0 (95% CI = 1.3–12.0)	This study compared tissue from patients with squamous cell carcinoma to tissue from control individuals with no history of skin cancer. No statistically significant association between patients (cases) and control individuals was observed when all HPV species were considered as the exposure. In the subgroup analysis, however, tissue from patients was 4 times more likely to contain the β -papillomavirus species compared with tissue from control individuals (Asgari et al., 2008).
Hazard ratio (HR)	The ratio of the rate at which patients with a risk factor experience an event to the rate at which patients without the risk factor experience an event	Systemic immune suppression	Merkel cell carcinoma-specific survival	HR = 3.8 (95% CI = 2.2–6.4)	The rate of death from Merkel cell carcinoma for people with systemic immune suppression was 3.8 times higher than for nonimmunosuppressed individuals (Paulson et al., 2013).
Pearson correlation coefficient (<i>r</i>)	Measures the strength and direction of the linear association between two continuous variables	<i>GOLPH3L</i> gene expression	<i>HORMAD1</i> gene expression	<i>r</i> = 0.991	There is a strong, positive linear relationship between <i>GOLPH3L</i> and <i>HORMAD1</i> gene expression, indicating that when one gene is expressed, the other is often expressed as well (Ioannidis et al., 2018).
Spearman correlation coefficient (ρ)	Measures the monotonic relationship between two variables	Individual typology angle	Melanin index	ρ = -0.98	There is a strong, negative monotonic relationship between individual typology angle and melanin index, indicating that when one is low, the other is high (Wilkes et al., 2015).
Beta coefficient (linear regression)	Measures the association between a continuous outcome variable and continuous and/or categorical predictor variable(s)	Pain (self-rated from 0–10)	Sleep quality score (range = 8–40, with higher scores indicating more disturbed sleep)	β = 0.21 <i>P</i> < .001	There is a positive relationship between self-rated pain and sleep disturbance. For each 1-unit increase in self-rated pain, sleep quality score increases by 0.21. The <i>P</i> -value indicates that this association is statistically significant (Milette et al., 2013).
Chi-squared test	Measures the association between two categorical variables by assessing whether there is a significant difference between observed and expected data	Training level of clinician	Treatment type	<i>P</i> < 0.0001	Patients treated with Mohs surgery were almost exclusively treated by attending physicians (98.8% vs. 1.2% resident/nurse practitioner). Patients receiving excision were treated slightly more frequently by resident physicians (51% vs. 46.8% attending and 2.1% nurse practitioner). Patients treated with destruction by electrodissection and curettage were more commonly treated by attending physicians (57.1% vs. 33.8% resident and 9.1% nurse practitioner). The <i>P</i> -value from the chi-squared test indicates that these differences are statistically significant (Asgari et al., 2009a).
Risk difference (RD)	Measures the difference in risk between exposed and unexposed groups	UV light therapy	Psoriasis	RD = -0.06	After receiving UV light therapy, 2% of patients continued to experience psoriasis, compared with 8% of patients not receiving this treatment. The RD indicates that patients who received light therapy had 6 fewer cases of persistent psoriasis per 100 people compared with patients not receiving light therapy. ²

(continued)

Table 2. Continued

Measure of Association	Definition	Exposure	Outcome	Effect Estimate	Interpretation
Relative risk reduction (RRR)	The proportion of risk reduction attributable to the exposure/intervention	UV light therapy	Psoriasis	RRR = 0.75	Using the data from the UV light/psoriasis example, the relative risk may be calculated as 0.02/0.08 = 0.25 (the incidence in the exposed group divided by the incidence in the unexposed group). The RRR is therefore 0.75(1 - RR), which can be interpreted as UV light therapy resulting in a 75% reduction in psoriasis incidence, relative to patients who did not receive light therapy. ²
Number needed to treat (NNT)	The number of patients who must be treated for one patient to benefit	UV light therapy	Psoriasis	NNT = 16.7	Using the data from the UV light/psoriasis example, the NNT may be calculated as 1/(incidence among the unexposed - incidence among the exposed), or 1/(0.08 - 0.02). Therefore, the NNT equals 16.7, indicating that 17 patients need to be treated with UV light therapy for one patient to benefit. ²

Abbreviations: CI, confidence interval; HPV, human papillomavirus.

¹Relative risk may also be referred to as the risk ratio, rate ratio, or relative rate.

²Mock data are used for these examples.

melanoma-specific survival analysis, only melanoma-related deaths were considered events, and participants who died of causes not related to melanoma were right-censored at the time of death. In the overall survival analysis, however, deaths from any cause were considered events (Asgari et al., 2017). In contrast to right-censoring, left-censoring occurs when the event has already taken place before the observation period begins, and the true unobserved event lies to the left of the censoring time. Estimation of the HR, as with Cox proportional hazards regression, accounts for only right-censored data (Clark et al., 2003).

When the HR is equal to 1, instantaneous event rates at a particular time are the same in the exposed and unexposed groups. When the HR is equal to 0.5, half as many people in the exposed group have experienced an event compared with the unexposed group, and when HR is equal to 2, twice as many people have experienced an event. For example, in a study examining the association between systemic immune suppression and Merkel cell carcinoma-specific survival, an HR of 3.8 was observed (95% CI = 2.2–6.4) (Paulson et al., 2013). This estimate indicates that the rate of death from Merkel cell carcinoma was 3.8 times higher in people with

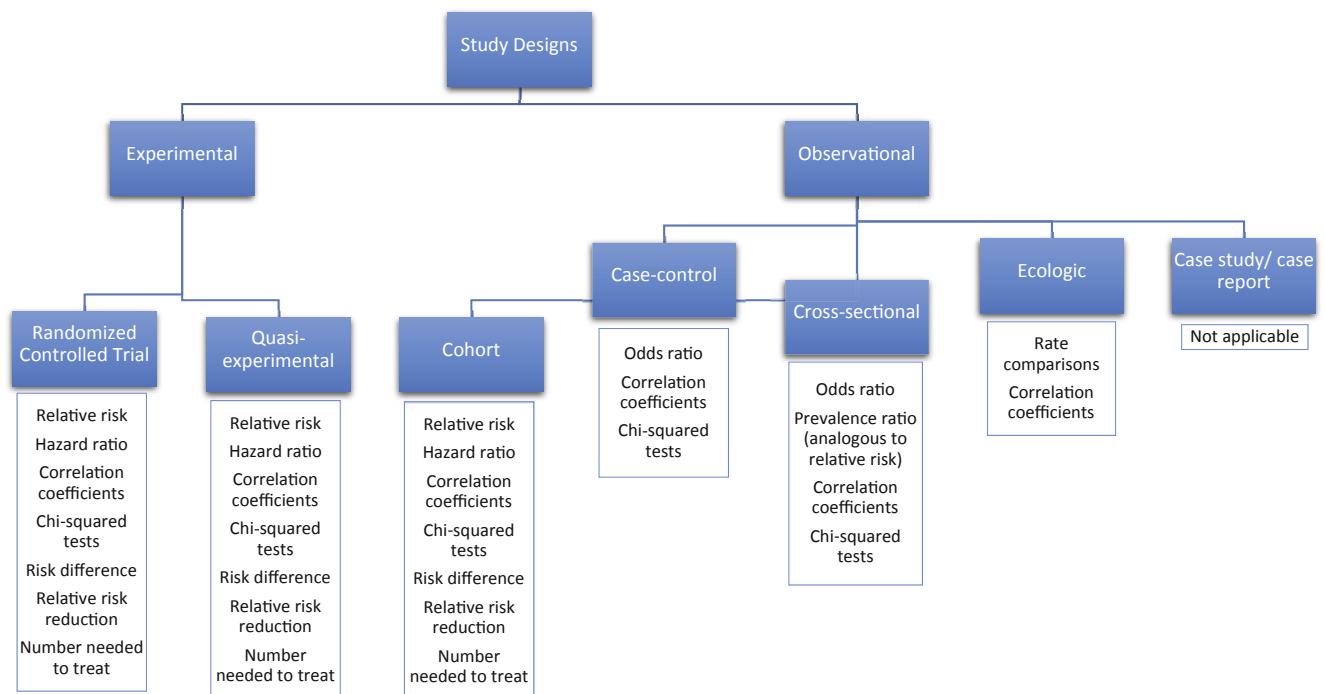


Figure 1. Measures of association used in common clinical research study designs. Measures of association commonly encountered in each type of study design are depicted.

Exposure	Outcome		Total
	Present	Absent	
Present	A (90)	B (210)	A + B (300)
Absent	C (350)	D (350)	C + D (700)
Total	A + C (440)	B + D (760)	A + B + C + D (1000)

Relative risk = $A/(A + B) / C/(C + D)$
 = $(90/300) / (350/700)$
 = $0.3 / 0.5$
 = 0.6
The exposed group has 0.6 times the risk of the outcome (or 40% less risk of the outcome), compared to the unexposed group

Odds ratio = $(A/C) / (B/D)$
 = $(90/350) / (210/350)$
 = $0.26 / 0.6$
 = 0.43
The case group has 0.43 times the odds of exposure (or are 57% less likely to have the exposure), compared to the control group

Risk difference = $A/(A + B) - C/(C + D)$
 = $0.3 - 0.5$ [or 30% - 50%]
 = -0.2 [-20%, or -20/100]
Participants with the exposure had 2 fewer instances of the outcome per 10 people, compared to participants without the exposure

Relative risk reduction =
 $((C/C + D) - A/(A + B)) / (C/C + D)$
 = $(0.5 - 0.3) / 0.5$
 = 0.4
There is a 40% reduction in risk of the outcome in the exposed group, relative to the unexposed group

Number needed to treat =
 $1 / ((C/C + D) - A/(A + B))$
 = $1 / (0.5 - 0.3)$
 = 5
5 individuals must receive the exposure to prevent one outcome from occurring

Figure 2. Calculation of common measures of association. A 2×2 contingency table displays the number of individuals with and without the exposure by the number of individuals with and without the outcome. This information can be used to calculate several commonly encountered measures of association.

systemic immune suppression. Because the 95% CI excludes the null value of 1, we can conclude that this HR is statistically significant.

Other measures of association

Other frequently encountered statistics include correlation coefficients, beta coefficients (linear regression), chi-squared/Fisher exact tests, risk difference, relative risk reduction, and number needed to treat (NNT) (Table 2).

Correlation coefficients, including the Pearson *r* and Spearman rho statistics, measure the strength and direction between two variables and range from -1 (perfect negative correlation) to +1 (perfect positive correlation). A positive correlation coefficient indicates that both variables increase or decrease together, whereas a negative coefficient implies that as one variable increases, the other decreases (see examples in Table 2). The Pearson *r* statistic is generally used when data are continuous rather than categorical, and it assumes that the data are normally distributed and that the variables are linearly related. When these assumptions are not met, or when categorical data are involved, Spearman rho may be more appropriate. Spearman rho assumes a monotonic relationship between ranked variables and can be used for ordinal-level data. It is essentially a Pearson correlation using variable ranks rather than variable values. Spearman rho is the nonparametric version of Pearson *r*, and therefore it may be appropriate for nonnormally distributed data or when variables are not linearly related (McDonald, 2014a). For example, in a study examining cutaneous sarcoidosis, Rosenbach et al. (2013) calculated the correlations between disease severity and quality of life using several different instruments. The Physician’s Global Assessment of disease severity was found to be moderately positively correlated with Skindex-29 assessments of symptoms (Pearson *r* = 0.41) but weakly negatively correlated with the Sarcoidosis Health Questionnaire assessment of quality of life (Pearson *r* = -0.18). The Physician’s Global Assessment, Skindex-29, and Sarcoidosis Health

Questionnaire data were normally distributed. Because the data from another assessment, the Dermatology Life Quality Index, were not normally distributed and the sample size was small, the authors used the Spearman rho correlation coefficient to identify a weak positive correlation with the Physician’s Global Assessment ($\rho = 0.24$).

Linear regression is used to assess the relationship between a continuous outcome variable and one or more categorical or continuous predictor variables. For continuous predictors, a positive β coefficient represents the increase in the outcome variable for every 1-unit increase in the predictor variable. Conversely, a negative β coefficient represents the decrease in the outcome variable for every 1-unit increase in the predictor variable. Beta coefficients for categorical predictors have a similar interpretation, except that the coefficient represents the change in the outcome variable when switching from one category of the predictor variable to another. For instance, a study of patients with systemic sclerosis sought to investigate associations between demographic and medical variables and sleep disturbance, measured using a sleep quality scale. The number of gastrointestinal symptoms (continuous predictor) and sleep disturbance (continuous outcome) were positively associated ($\beta = 0.19, P = 0.001$). The beta coefficient indicates that for each 1-unit increase in the number of gastrointestinal symptoms, sleep quality score increases by 0.19 units. Female sex was also positively associated with sleep disturbance, although the association was not statistically significant ($\beta = 0.07, P = 0.164$). Because sex is a categorical variable, this beta coefficient indicates that being female, as opposed to being male, is associated with a 0.07-unit increase in sleep quality score (Milette et al., 2013).

The chi-squared and Fisher exact statistics are often used for testing relationships between categorical variables. These tests evaluate whether the proportions of one categorical variable differ by levels of another categorical variable (see example in Table 2). The null hypothesis for the chi-squared/Fisher exact test is that the variables are independent; that is,

Table 3. Points to consider when interpreting epidemiologic studies

Bias, confounding, and statistical significance

1. Can the presence of biases or confounding explain the results?
 - Biases and unaccounted for or unmeasured confounders may affect the validity of the point estimate
 - o Information bias: systematic errors in measurement that result in participants being misclassified with respect to exposure or outcome
 - Differential: classification errors are more likely in one group over another
 - Nondifferential: frequency of errors is roughly the same in the groups being compared
 - o Selection bias: results from the study population being nonrepresentative of the target population, and stems from
 - Control groups that are not representative of the population that produced the cases
 - Nonresponse or self-selection, whereby participation is related to exposure status
 - Differential loss to follow-up, in which the likelihood of being lost to follow-up is associated with exposure and/or outcome status
 - o Confounding: distortion of the true exposure-outcome relationship by independent variables that are associated with both exposure and outcome
 - Can include variables such as age, sex, socioeconomic status, etc.
2. What is the variability?
 - Wider confidence intervals indicate reduced precision of the point estimate
 - Sample size can affect the estimate of effect size and statistical significance—small studies should be interpreted cautiously

Replication and generalizability

1. Have the results been replicated?
 - Can methodologic weaknesses explain discrepancies in results between studies?
2. Is the exposure or intervention likely to have caused the outcome(s) reported?
 - Evaluating the body of evidence and methodologic concerns in individual studies can aid in assessment of potential causality
 - Although randomized controlled trials are often considered the standard for determining causality, they may be implausible for many exposures
3. Do the results of a study apply only to particular groups of people?
 - Differences between clinical and study populations may result from age, race, cultural factors, presence of comorbidities, etc.
4. Are there differences in the time course of the exposure or intervention under study compared with a clinical population?

the level of variable *A* does not predict the level of variable *B*. For each level of one variable, the expected frequencies at each level of the second variable are calculated. The chi-squared test statistic is based on the difference between the frequencies that are actually observed and those that would be expected if there were no relationship between the two variables. The more computationally intensive Fisher exact test is typically used only when sample sizes are small. These tests do not evaluate the magnitude of the association but indicate whether the association is statistically significant. For example, in a study examining patient satisfaction after treatment for nonmelanoma skin cancer with either destruction, excision, or Mohs surgery, categorical patient characteristics were compared among treatment groups using chi-squared or Fisher exact tests. The training level of the treating clinician (attending, resident, or nurse practitioner) differed significantly by treatment group ($P < 0.001$) (Asgari et al., 2009a).

The risk difference is the absolute difference in risk between exposed and unexposed groups, and it is useful for evaluating the excess risk of disease associated with an exposure. The relative risk reduction is the proportion of risk that is reduced in the exposed group relative to the unexposed group. The number needed to treat is the number of patients who must be treated for one patient to benefit. Calculations for risk difference, relative risk reduction, and

number needed to treat are shown in Figure 2, and examples are provided in Table 2.

METHODOLOGIC CONSIDERATIONS

Resources such as the US Preventive Services Task Force, Cochrane Library, International Agency for Research on Cancer monographs, UpToDate, and DynaMed Plus provide evidence-based guidelines for clinical practice. However, for many diseases, expert summaries may be unavailable, making the interpretation of clinical research critical for providers. Accurate interpretation requires a familiarity with methodologic considerations in epidemiology, outlined briefly in this section (Table 3).

Bias and confounding

Examining potential sources of biases or confounding is crucial for evaluating the validity of study findings (Figure 3) (Delgado-Rodríguez and Llorca, 2004; Sackett, 1979; Silverberg, 2015). Biases are systemic errors that result in incorrect estimation of the exposure-outcome association. Information biases are systematic errors in measurement, which result in participants being misclassified with respect to exposure or outcome. Selection biases stem from the study population being nonrepresentative of the target population. The presence of bias may result in an overestimation or underestimation of the true association.

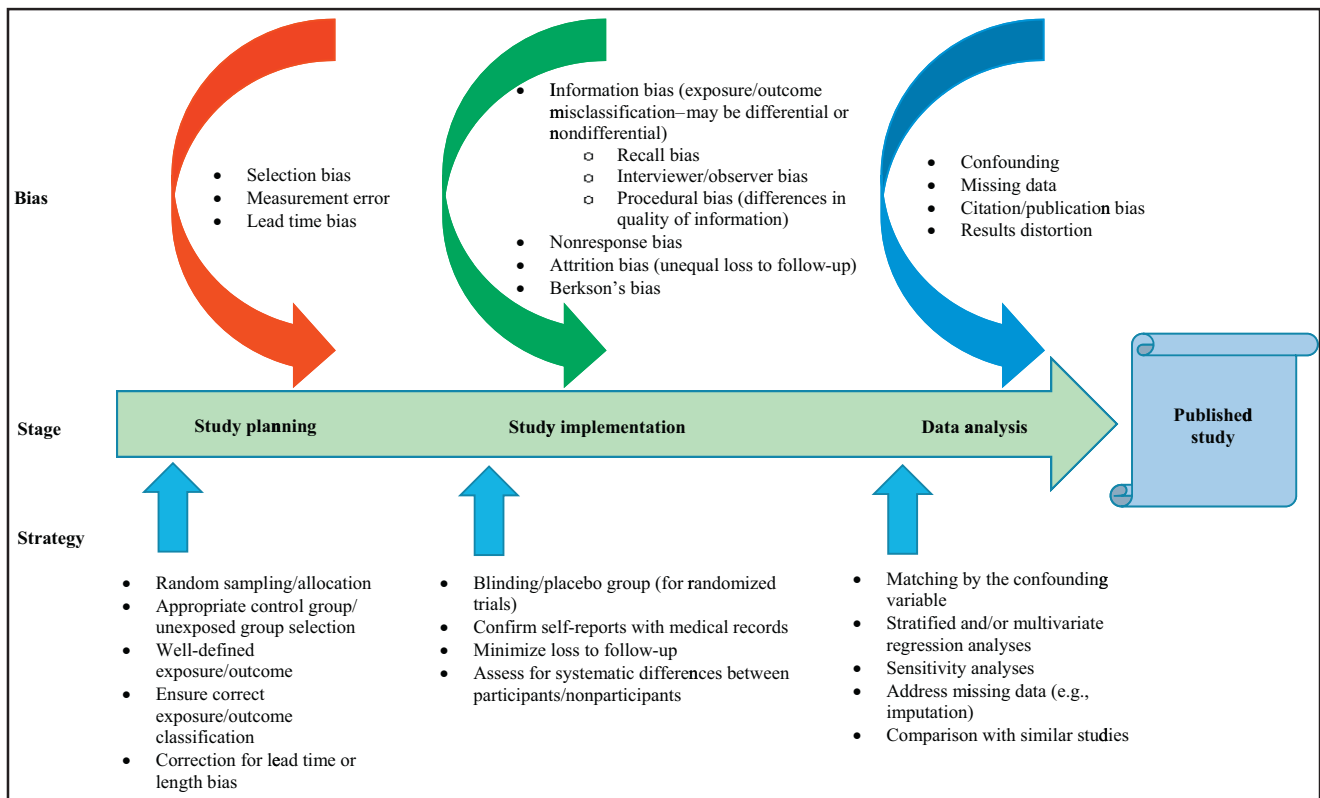


Figure 3. Strategies to minimize biases common to observational research. Methods for addressing various biases in epidemiologic research are shown, although this list is not exhaustive. Readers are referred to several excellent reviews, including [Choi and Pak \(2005\)](#), [Delgado-Rodríguez and Llorca \(2004\)](#), and [Sackett \(1979\)](#).

Confounding is a distortion of the exposure-outcome relationship by independent variables that are associated with both exposure and outcome. Confounding may be minimized through statistical adjustment, stratification, matching, or randomization. Methods to address confounding have been reviewed in detail elsewhere ([Greenland and Morgenstern, 2001](#); [Kim et al., 2017](#); [McNamee, 2005](#); [Wakkee et al., 2014](#)). Suppose that, when examining the association between serum vitamin D levels and skin cancer risk, we observe an OR of 1.85, indicating an 85% increased risk of skin cancer among participants with high serum vitamin D levels compared with those who have low levels. If participants with high vitamin D levels are also more likely to have increased sun exposure, it could erroneously appear that vitamin D increases the risk of skin cancer. In this hypothetical example, when sun exposure is addressed through statistical adjustment, we observe an OR of 1.15. The attenuated adjusted OR indicates that our unadjusted association was spurious and due to confounding caused by strong sun exposure-vitamin D and sun exposure-skin cancer associations. The likelihood of observing spurious associations may therefore be reduced by implementing methods to reduce confounding. Even when confounding is addressed, however, unmeasured confounders or residual confounding may distort the observed association.

Statistical significance

Although a *P*-value less than 0.05 is widely considered statistically significant, this cutoff is arbitrary and does not

necessarily equate to clinical significance. Effect sizes, which indicate the magnitude of the difference between groups, and measures of variability, such as confidence intervals, are more informative when interpreting epidemiologic data ([Greenland et al., 2016](#); [Sullivan and Feinn, 2012](#)). Wide confidence intervals indicate large variability and reduced precision of a point estimate. Other measures of variability or dispersion include range, interquartile range, variance, and standard deviation. These measures indicate the extent to which the mean of a given variable represents the study population as a whole.

Statistical power is the probability of correctly rejecting the null hypothesis when it is false, or, alternatively, the likelihood of finding a statistically significant difference when one truly exists ([Sullivan and Feinn, 2012](#)). Power is dependent upon effect size and sample size. Overpowered studies with very large sample sizes may detect very small effect sizes that are not clinically meaningful ([Bhardwaj et al., 2004](#)). Results from underpowered studies should also be interpreted with caution, because true associations may be masked by small sample size, or conversely, spurious, inflated risk estimates may be detected ([Button et al., 2013](#)).

Finally, when a large number of statistical tests are performed, some will be significant at $P < 0.05$ by chance alone, even when the null hypothesis is true. Statistical corrections for multiple comparisons aim to reduce the number of false positive findings; they include the Bonferroni correction, which reduces the *P*-value threshold for significance; resampling methods; and adjusting the false discovery rate. More

MULTIPLE CHOICE QUESTIONS

1. A study follows adults with psoriasis treated with either retinoids alone or retinoids with corticosteroids. The relative risk of 6-month psoriasis recurrence is 0.8. What is the correct interpretation of this finding?
 - A. The incidence of psoriasis recurrence in adults who are dual-treated with retinoids and corticosteroids is 0.8 (80%).
 - B. Adults with psoriasis who are dual-treated with topical retinoids and corticosteroids have 0.8 times the risk of having 6-month psoriasis recurrence compared with those who receive only retinoid treatment.
 - C. The difference in risk of 6-month psoriasis recurrence between adults treated with only retinoids and those dual-treated with topical retinoids and corticosteroids is 0.8 (80%).
 - D. The difference in risk of 6-month psoriasis recurrence between adults treated with only retinoids and those dual-treated with topical retinoids and corticosteroids is 0.2 (20%).
2. In a case-control study, what measure of association should be used to calculate associations between the exposure and outcome?
 - A. Hazard ratio
 - B. Pearson correlation coefficient
 - C. Odds ratio
 - D. Relative risk
3. Can an odds ratio ever approximate the relative risk?
 - A. Yes, when the outcome (i.e., disease) being studied is rare.
 - B. Yes, when the exposure being studied is rare.
 - C. No, because the odds ratio is calculated using odds, and the relative risk is calculated using incidence rates.
 - D. No, because these measures are calculated using data from different study designs.
4. What are confounders?
 - A. Variables that are associated only with the exposure
 - B. Independent variables that are associated with both the exposure and the outcome
 - C. Variables that are associated only with an outcome
 - D. Variables that are the consequence of an exposure
5. A chi-squared test is used for what type of data?
 - A. Discrete quantitative
 - B. Ratio
 - C. Continuous quantitative
 - D. Categorical

detailed information about multiple comparisons may be found in [Bender and Lange \(2001\)](#), [Cao and Zhang \(2014\)](#), and [McDonald \(2014b\)](#).

Replication, causality, and generalizability

Replication is key in clinical research, and methodologic concerns that may explain discrepancies between studies should be considered. In observational studies, causality between an exposure and outcome is difficult to ascertain concretely. For many exposures, randomized controlled trials are implausible, and well-designed observational studies are the best alternative ([Rothman and Greenland, 2005](#)). Clinicians should also judge the degree to which a study simulates clinical practice and whether the results are generalizable to his/her own patient population ([Wu et al., 2014](#)). For example, mutations in *NCSTN*, *PSENN1*, and *PSEN1*, which affect the function of γ -secretase, have been strongly associated with familial hidradenitis suppurativa in Chinese individuals. In other populations, however, γ -secretase mutations affect only a minority of hidradenitis suppurativa patients ([Ingram, 2016](#)). In some instances, lack of generalizability may render study findings noninformative for populations with different characteristics.

SUMMARY

Measures of association quantify the relationship between an exposure and an outcome, enabling comparison between different groups, and their validity is highly dependent on the methodologic context in which they were calculated. Interpreting epidemiologic findings, therefore, requires an assessment of study methodology, including sources of bias and confounding, generalizability, and replication of results. Evaluating these factors enables clinicians to critically evaluate the strength of evidence and make informed decisions for patient care.

CONFLICT OF INTEREST

The authors state no conflict of interest.

ACKNOWLEDGMENTS

This work was supported by the National Institute of Arthritis and Musculoskeletal and Skin Diseases (K24 AR069760, principal investigator: MMA).

SUPPLEMENTARY MATERIAL

Supplementary material is linked to this paper. Teaching slides are available as supplementary material.

REFERENCES

Asgari MM, Bertenthal D, Sen S, Sahay A, Chren M-M. Patient satisfaction after treatment of nonmelanoma skin cancer. *Dermatol Surg* 2009a;35:1041–9.

Asgari MM, Kiviat NB, Critchlow CW, Stern JE, Argenyi ZB, Raugi GJ, et al. Detection of human papillomavirus DNA in cutaneous squamous cell carcinoma among immunocompetent individuals. *J Invest Dermatol* 2008;128:1409–17.

Asgari MM, Maruti SS, Kushi LH, White E. A cohort study of vitamin D intake and melanoma risk. *J Invest Dermatol* 2009b;129:1675–80.

Asgari MM, Shen L, Sokil MM, Yeh I, Jorgenson E. Prognostic factors and survival in acral lentiginous melanoma. *Br J Dermatol* 2017;177:428–35.

Bender R, Lange S. Adjusting for multiple testing—when and how? *J Clin Epidemiol* 2001;54:343–9.

Besen J, Gan SD. A critical evaluation of clinical research study designs. *J Invest Dermatol* 2014;134(3):e18.

Bhardwaj SS, Camacho F, Derrow A, Fleischer AB, Feldman SR. Statistical significance and clinical relevance. *Arch Dermatol* 2004;140:1520–3.

- Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci* 2013;14:365–76.
- Cao J, Zhang S. Multiple comparison procedures. *JAMA* 2014;312:543.
- Choi BCK, Pak AWP. A catalog of biases in questionnaires. *Prev Chronic Dis* 2005;2(1):A13.
- Clark TG, Bradburn MJ, Love SB, Altman DG. Survival analysis part I: basic concepts and first analyses. *Br J Cancer* 2003;89:232–8.
- Delgado-Rodríguez M, Llorca J. Bias. *J Epidemiol Community Health* 2004;58:635–41.
- Greenland S, Morgenstern H. Confounding in health research. *Annu Rev Public Health* 2001;22:189–212.
- Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol* 2016;31:337–50.
- Ingram JR. The genetics of hidradenitis suppurativa. *Dermatol Clin* 2016;34:23–8.
- Ioannidis NM, Wang W, Furlotte NA, Hinds DA, 23andMe Research Team, Bustamante CD, et al. Gene expression imputation identifies candidate genes and susceptibility loci associated with cutaneous squamous cell carcinoma. *Nat Commun* 2018;9(1):4264.
- Kim N, Fischer AH, Dyring-Andersen B, Rosner B, Okoye GA. Research techniques made simple: choosing appropriate statistical methods for clinical research. *J Invest Dermatol* 2017;137(10):e173–8.
- McDonald J. Regressions. In: *Handbook of biological statistics*. 3rd ed. Baltimore, MD: Sparky House Publishing; 2014a. p. 209–12. <http://www.biostathandbook.com/spearman.html>.
- McDonald J. Multiple tests. In: *Handbook of biological statistics*. 3rd ed. Baltimore, MD: Sparky House Publishing; 2014b. p. 254–60. <http://www.biostathandbook.com/multiplecomparisons.html>.
- McNamee R. Regression modelling and other methods to control confounding. *Occup Environ Med* 2005;62:500–6.
- Milette K, Hudson M, Korner A, Baron M, Thombs BD. Canadian Scleroderma Research Group. Sleep disturbances in systemic sclerosis: evidence for the role of gastrointestinal symptoms, pain and pruritus. *Rheumatology* 2013;52:1715–20.
- Paulson KG, Iyer JG, Blom A, Warton EM, Sokil M, Yelistratova L, et al. Systemic immune suppression predicts diminished Merkel cell carcinoma-specific survival independent of stage. *J Invest Dermatol* 2013;133:642–6.
- Pearce N. What does the odds ratio estimate in a case-control study? *Int J Epidemiol* 1993;22:1189–92.
- Rosenbach M, Yeung H, Chu EY, Kim EJ, Payne AS, Takeshita J, et al. Reliability and convergent validity of the cutaneous sarcoidosis activity and morphology instrument for assessing cutaneous sarcoidosis. *JAMA Dermatology* 2013;149:550.
- Rothman KJ, Greenland S. Causation and causal inference in epidemiology. *Am J Public Health* 2005;95(Suppl. 1):S144–50.
- Sackett DL. Bias in analytic research. *J Chronic Dis* 1979;32:51–63.
- Silverberg JI. Study designs in dermatology. *J Am Acad Dermatol* 2015;73:721–31.
- Sullivan GM, Feinn R. Using effect size-or why the P value is not enough. *J Grad Med Educ* 2012;4:279–82.
- Wakkee M, Hollestein LM, Nijsten T. Multivariable analysis. *J Invest Dermatol* 2014;134(5):e20.
- Wilkes M, Wright CY, du Plessis JL, Reeder A. Fitzpatrick skin type, individual typology angle, and melanin index in an African population: steps toward universally applicable skin photosensitivity assessments. *JAMA Dermatol* 2015;151:902–3.
- Wu AW, Bradford AN, Velanovich V, Sprangers MAG, Brundage M, Snyder C. Clinician's checklist for reading and using an article about patient-reported outcomes. *Mayo Clin Proc* 2014;89:653–61.

RESEARCH TECHNIQUES MADE SIMPLE

DETAILED ANSWERS

1. A study follows adults with psoriasis treated with either retinoids alone or retinoids with corticosteroids. The relative risk of 6-month psoriasis recurrence is 0.8. What is the correct interpretation of this finding?

Correct answer: B. Adults with psoriasis who are dual-treated with topical retinoids and corticosteroids have 0.8 times the risk of having 6-month psoriasis recurrence compared with those who receive only retinoid treatment.

Adults with psoriasis who are dual-treated with topical retinoids and corticosteroids have 0.8 times the risk of having 6-month psoriasis recurrence compared with those who receive only retinoid treatment, indicating that dual treatment reduces the risk of psoriasis recurrence. The relative risk compares the incidence of disease in the exposed group relative to the incidence of disease in the unexposed group. It does not describe the incidence of the outcome. The risk difference is the difference between risk in the exposed group and the risk in the nonexposed group.

2. In a case-control study, what measure of association should be used to calculate associations between the exposure and outcome?

Correct answer: C. Odds ratio

In a case-control study, we cannot calculate incidence because we start with a specific number of patients (cases) and control individuals. Consequently, we compare the odds of the exposure between the case and control groups to determine the association between exposure and outcome.

3. Can an odds ratio ever approximate the relative risk?

Correct answer: A. Yes, when the outcome (i.e., disease) being studied is rare.

We can use a 2×2 contingency table to see how the odds ratio can approximate the relative risk when the rare disease assumption is met. When the proportions in cells A and C are small, $A + B \approx B$, $C + D \approx D$. Therefore, due to algebraic rearrangement, $RR = [A/(A + B)]/[C/(C + D)] \approx [A/B]/[C/D] = [A/C]/[B/D] = OR$.

4. What are confounders?

Correct answer: B. Independent variables that are associated with both the exposure and the outcome

Because confounders have a relationship with both the exposure and the outcome, they can distort the true exposure-outcome relationship. For example, a study may find an association between physical activity level and weight gain. However, age may act as a confounding variable, because older individuals may have decreased physical activity but also slower metabolic activity, which can contribute to weight gain.

5. A chi-squared test is used for what type of data?

Correct answer: D. Categorical

The chi-squared test is used to determine if there is a relationship between two categorical variables, such as sex or skin color. The null hypothesis of the chi-squared test is that there is no relationship between the categorical variables (i.e., the variables are independent of each other).