Research Techniques Made Simple: Choosing Appropriate Statistical Methods for Clinical Research

Noori Kim[1,4], Alexander H. Fischer[1,4], Beatrice Dyring-Andersen[2], Bernard Rosner[3] and Ginette A. Okoye[1]

The statistical significance of results is an important component to drawing appropriate conclusions in a study. Choosing the correct statistical test to analyze results is essential in interpreting the validity of the study and centers on defining the study variables and purpose of the analysis. The complexity of statistical modeling makes this a daunting task, so we propose a basic algorithmic approach as an initial step in determining what statistical method will be appropriate for a particular clinical study.

Journal of Investigative Dermatology (2017) **137,** e173–e178; doi:10.1016/j.jid.2017.08.007

CME Activity Dates: 20 September 2017
Expiration Date: 19 September 2018
Estimated Time to Complete: 1 hour

Planning Committee/Speaker Disclosure: All authors, planning committee members, CME committee members and staff involved with this activity as content validation reviewers have no financial relationship(s) 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 William Beaumont Hospital and the Society for Investigative Dermatology. William Beaumont Hospital is accredited by the ACCME to provide continuing medical education for physicians. William Beaumont Hospital 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-Oct17 – 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

Choosing the correct statistical method when analyzing clinical data can be a daunting task. We propose an algorithmic approach to organizing the basic key elements in a clinical study that will guide which statistical test is best (Altman, 1991; Rosner, 2015).

This guide is not meant to be a comprehensive guide for data analysis. Instead, this set of instructions is meant to act as a general overview to give the researcher a starting point to help determine what statistical test is appropriate for data analysis. This article contains little discussion of the assumptions of the various statistical tests or the nuances of statistical modeling. Thus, we encourage the reader to investigate the specific tests to be used to ensure that the assumptions are appropriate and to consider consulting a biostatistician with questions about the appropriate analytic approach. This guide also does not go in depth about choosing the appropriate study design to address the research question. However, we include a few resources that may help with this aspect of the investigatory process,

[1]*Department of Dermatology, Johns Hopkins University School of Medicine, Baltimore, Maryland;* [2]*Department of Dermatology, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts; and* [3]*Channing Division of Network Medicine, Department of Medicine, Harvard Medical School, Boston, Massachusetts*

[4] *These authors contributed equally to this work*

Correspondence: Noori Kim, 601 North Caroline Street, 8th Floor Dermatology, Baltimore, Maryland 21287, USA. E-mail: nkim34@jhmi.edu

SUMMARY POINTS

It is important to initially define two main elements to help identify the appropriate statistical method for a study:

- What is being measured in the study? (Study variables)
- How are these variables related? (Purpose of analysis)

It is important to identify the assumptions of various statistical tests to ensure selection of an appropriate method for a study.

LIMITATIONS

This reference offers a basic algorithmic approach to choosing a statistical test in clinical research. This reference cannot capture all of the nuances of statistical testing, and it should not serve as a substitution to the consultation of a biostatistician.

including clinical research study design (Besen and Gan, 2014), clinical trials (Williams et al., 2015), and comparative effectiveness research (Nambudiri and Qureshi, 2013).

The two main elements to determining the correct statistical test are defining the study variables and defining the purpose of the analysis, which will be explained in more detail. To help contextualize these concepts, we refer to two specific arbitrary examples throughout the text:

- the association between sunburn and the number of pigmented nevi a person has and
- the association between tanning bed use and risk of melanoma.

DEFINING THE STUDY VARIABLES

It is important to have a good grasp of the study variables. The variable characteristics will dictate which statistical tests can be performed.

Generally, when examining an association, variables fit one of two types. The *outcome variable,* synonymous with the *dependent variable,* refers to the variable that we want to explain or predict as a result of the variation in the *explanatory variable,* or *independent variable*. In this case, the number of nevi in example 1 or risk of melanoma in example 2 would be the outcome variables of interest. The explanatory variables would be sunburn in example 1 and tanning bed use in example 2.

Both explanatory and outcome variables are further subcategorized by the distribution of the data as categorical or continuous variables. Categorical variables with only two categories (e.g., yes, no) are called *dichotomous* or *binary variables* (e.g., history of tanning bed use, history of melanoma, sex). Categorical variables with more than two qualitative, nonvalue measurements are referred to as *nominal variables* (e.g., race, state of residence). Categorical variables with ordered ranges in which the differences between values have no well-defined meaning are referred to as *ordinal variables* (e.g., pain score, patient satisfaction scale, or Likert item [e.g., strongly agree, agree, neutral, disagree, strongly disagree]). In contrast, continuous variables are quantitative, where the differences in values are meaningful (e.g., age, body mass index). Parametric statistical tests can be used to analyze continuous variables that follow a particular distribution. A histogram can be used to get a general sense of the distribution. Knowing whether a continuous variable follows a normal distribution (i.e., a symmetric bell-shaped Gaussian distribution) is important for choosing the appropriate statistical test for a continuous variable. Sometimes regression analyses include variables that follow a count distribution, meaning that there are a discrete number of events over a time period (e.g., number of sunburns, number of total body skin examinations). Table 1 can be used as a reference for summarizing these different types of variables.

Table 1. Summary statistics to describe a study population or group

Variable Distribution	Summary Statistic(s)
Continuous (normally distributed)	Mean (standard deviation)
Continuous (not normally distributed)	Median (interquartile range)
Ordinal	Median (interquartile range)
Dichotomous	Proportion
Nominal	Relative proportions

An additional consideration that should be made is whether the different groups that make up a categorical explanatory variable are related. Examples could include applying topical treatment and placebo to two different lesions on each individual in the study population. Another example could be a baseline and follow-up visit(s) of the same group of individuals in the study population. Analytic points to keep in mind in this situation are discussed in the section on "Independence of Observations."

DEFINING THE PURPOSE OF THE ANALYSIS

It is important to determine the purpose of the analysis to choose the appropriate statistical test to support the research question. It is important to ensure that the statistical analysis is appropriate for the way that the study was designed and the data were collected. This article offers a logical way to approach selection of the appropriate statistical test.

- If the purpose is to determine if two continuous variables in the study population are correlated, a Pearson correlation should be used if both variables are normally distributed or if the relationship between the two variables is linear, and a Spearman correlation should be used if at least one variable is not normally distributed. For example, we could examine whether the number of lifetime sunburns correlates with the number of pigmented nevi that a person has.
- If the purpose is to determine if the distribution of a variable (outcome variable) is different across two or more subgroups (explanatory variable), then Table 2 can be used. For

Table 2. Choosing a statistical test to determine if the distribution of the outcome variable is different across two or more explanatory subgroups

	Explanatory Variable			
Outcome Variable	**Dichotomous (Unrelated)**	**Dichotomous (Related)[1]**	**Three or More Subgroups (Unrelated)**	**Three or More Subgroups (Related)[1]**
Continuous (normally distributed)	Two-sample *t* test	Paired *t* test	Analysis of variance (ANOVA)	Mixed-effects model for repeated measures
Ordinal, Continuous (not normally distributed)	Wilcoxon rank sum test	Wilcoxon signed rank test	Kruskal-Wallis test	Friedman test, Skillings-Mack test
Categorical	Chi-square test, Fisher exact test[2]	McNemar test	Chi-square test, Fisher exact test[2]	Cochran *Q* test[3]

[1]Analyses in which the groups of the explanatory variable are related may be better addressed using multilevel modeling techniques, and thus the investigator should consider consulting with a biostatistician to help appropriately incorporate these analytic techniques.

[2]Fisher exact test should be used if at least one expected count is less than 5.

[3]For dichotomous outcome only.

example, we could examine whether the number of pigmented nevi differs between individuals who have ever sunburned in their lifetime versus individuals who have never sunburned in their lifetime.

- If the purpose is to determine if two variables are associated, with or without adjustment for other variables, then the information regarding regression analysis in Table 3 may be used (Wakkee et al., 2014). For example, we could examine whether number of years of tanning bed use was associated with risk of melanoma, adjusting for differences in age, Fitzpatrick skin type, and frequency of tanning bed use among individuals.

ADDITIONAL ITEMS TO KEEP IN MIND

Finally, analyses are rarely completely straightforward. Below, we discuss a few topics that have commonly come up in our research when determining the correct statistical test to use.

Independence of observations

In general, most statistical tests hinge on the assumption that the observations (individual patients) are independent, or unrelated, unless otherwise stated. As we alluded to earlier when describing the different types of variables, a variable may include observations that are related. Examples include conducting analyses on multiple samples or sites from the same person, repeated samples, studies with longitudinal follow-up of the same patient cohort, and matched case-control studies. It is important to note that related observations should not be treated as independent observations. The statistical analysis must reflect the assumption that related observations are likely to be more similar to each other than independent observations. Examples that appropriately take these assumptions into account include the tests for related

Table 3. Choosing an appropriate statistical model to examine if an outcome variable is associated with one or more explanatory variables

Outcome Variable	**Measure of Association**	**Regression Model[1]**	**Notes**
Continuous variable	Difference in means	Linear regression	Residuals should meet assumptions, otherwise the continuous outcome may need to be broken into categories and analyzed using multinomial logistic regression.
Ordinal variable	Odds ratio	Ordinal logistic regression	Model should meet proportional odds assumption.
Count variable	Incidence rate ratio	Poisson regression	The count variable should meet the Poisson assumptions; otherwise, other models may be used including negative binomial regression, zero-inflated or zero-truncated models, etc.
Dichotomous variable (case-control analysis)	Odds ratio	(Unconditional) logistic regression	Because of the case-control study design and thus the inability to determine prevalence of the outcome of interest, the odds ratio is the only measure of association that can be used.
Dichotomous variable (matched case-control analysis)	Odds ratio	Conditional logistic regression	If additional confounding is still present after matching, multivariable conditional logistic regression may be used.
Dichotomous variable (cross-sectional analysis)	Prevalence ratio (relative risk)	Log binomial regression	Prevalence odds ratio using logistic regression may also be used; however, if the prevalence of the outcome exceeds 10%, the odds ratio will overestimate the relative risk. Poisson regression with a robust variance estimator can be used if the log binomial regression fails to converge.
Dichotomous variable (longitudinal analysis with a discrete time interval)	Cumulative incidence ratio (relative risk)	Log binomial regression	Poisson regression with a robust variance estimator can be used if the log binomial regression fails to converge.
Dichotomous variable (longitudinal time-to-event analysis)	Hazard ratio (relative risk)	Cox proportional hazards model	Model should follow the proportional hazards assumption.
Nominal variable	Odds ratio	Multinomial logistic regression	Odds ratios for the different categories will be compared with a common reference category.

[1]Regression models can incorporate one or more categorical or continuous explanatory variables.

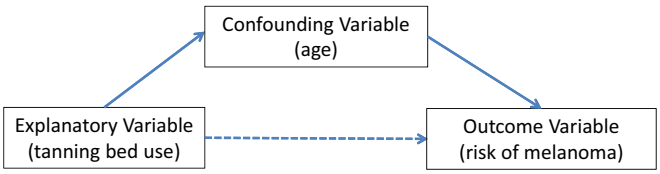

Figure 1. Confounding variables. When examining the potential association between the main explanatory variable and the outcome variable of interest, confounding variables to consider are those variables that are risk factors for the outcome variable of interest, associated with the main explanatory variable, but not in the causal pathway between the main explanatory variable and the outcome variable of interest.

categorical variables (Table 2) and conditional logistic regression (Table 3). If the study design is more complicated than these tests allow, more complex multilevel modeling approaches will likely be required, and the investigator should consider consulting a biostatistician to appropriately incorporate these analytic methods.

Confounding

One of the most fundamental ideas behind a research experiment is having a controlled environment in which only one variable is changed at a time. Although this idea is much more attainable in a randomized controlled trial, in observational studies we often see that the subgroups being compared are not completely comparable. This lack of comparability, or confounding, will skew results so that the true association between the outcome variable and the explanatory variable of interest is biased. We often try to avoid confounding either in the study design stage (e.g., matched case-control study, randomized controlled trial) and/or data analysis (e.g., adjustment for potential confounding variables using multivariable regression analysis [Wakkee et al., 2014] or stratified analyses, explained in the section on "Interactions").

Confounding variables are classically identified as variables that are (i) risk factors for the outcome variable, (ii) associated with the main explanatory variable, and (iii) not in the causal pathway between the main explanatory variable and the outcome variable. When adjusting analyses for confounding using multivariable regression analysis, the idea is to look at analyses between the main explanatory variable and the outcome variable of interest, assuming that the distribution of the confounding variable is the same across the various groups of the explanatory variable. After adjusting for confounding variables, the relationship between the main explanatory variable and the outcome variable of interest is thus considered independent of the confounding variable. Examples of potential confounding variables when examining the association between indoor tanning use and risk of melanoma may include age and Fitzpatrick skin type, because the distribution of age and Fitzpatrick skin type among individuals who tan indoors is likely to be different from individuals who do not tan indoors (Figure 1).

Interactions

When examining whether an association between an explanatory variable and an outcome is different in one

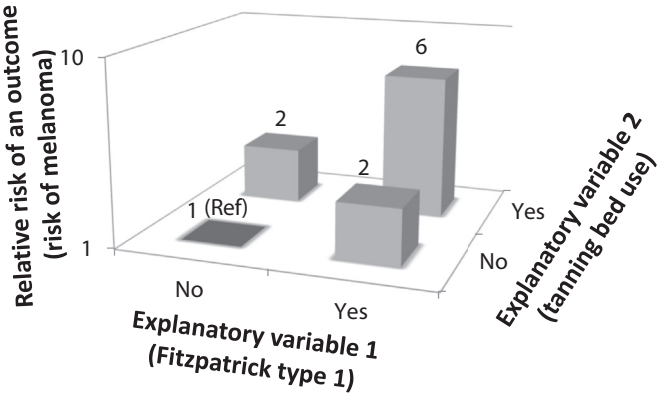

Figure 2. Interaction analysis. When examining the interaction between two explanatory variables to predict a given outcome, we would consider an interaction to be observed if the presence of both explanatory variables resulted in risk for the outcome that was different (higher in this case) than the risk that would be expected from both individual variables combined, in reference to risk for the outcome without both explanatory variables.

subpopulation versus another, it is not enough to simply compare the magnitudes of association. Essentially, there are two explanatory variables that need to be accounted for in this analysis, and thus this joint association should be formally tested with an interaction term. An example in which an interaction analysis would be necessary would be if we were to examine whether the magnitude of association between tanning bed use and risk of melanoma was greater in individuals with Fitzpatrick skin type 1 versus 2 (Figure 2). Using this example, a general stepwise approach would be to examine the magnitudes of association between tanning bed use and risk of melanoma in all individuals. Next, examine this association in individuals with Fitzpatrick skin type 1 and separately in individuals with Fitzpatrick skin type 2. If the magnitudes of association were significantly different by an interaction test, then we would consider this a significant interaction, and we would want to present this interesting finding in stratified analyses. If the magnitudes of association changed in both subpopulations by more than 10% and in the same direction (interaction test was not significant), then we would consider Fitzpatrick skin type to be a confounding variable, and we would want to adjust analyses for Fitzpatrick skin type. If the magnitudes of association did not change by more than 10%, we would generally not consider Fitzpatrick skin type to be a confounding variable and would not adjust analyses for this variable, unless we deemed this variable to be important to our question of interest.

P-values in relation to sample size and multiple comparisons

Although a *P*-value less than 0.05 is generally used as a cutoff in determining statistical significance, there are factors that may affect what *P*-value cutoff is used. If multiple pairwise associations are examined, results should be adjusted for multiple comparisons. The idea behind this is that if we take a *P*-value less than 0.05 to be significant and test multiple pairwise comparisons, we would expect that 5% of these comparisons would end up being significant just by chance alone. Thus, a more stringent cutoff for statistical significance should be used in this case. Examples of correcting for multiple comparisons include the Bonferroni correction, weighted Bonferroni corrections, false

MULTIPLE CHOICE QUESTIONS

1. Ordinal variables
   A. are called dichotomous or binary variables.
   B. have no well-defined meaning between values, for example, pain score.
   C. include nonvalue measurements such as race.
   D. are outcome variables.
2. Most statistical tests hinge on the assumption that the observations are independent. An example of this is
   A. multiple samples or sites from the same person.
   B. repeated samples from the same person.
   C. studies with longitudinal follow-up of the same patient cohort.
   D. one sample per patient in the target group and control group.
3. What is NOT a characteristic of a confounding variable?
   A. Risk factors for the outcome variable
   B. Can skew the true association between the outcome and explanatory variable
   C. Measures whether the association between an explanatory and outcome variable is different in one subpopulation versus another
   D. Can be adjusted for by using a multivariable regression analysis
4. If our study were comparing the number of nevi among individuals based on whether they have ever or never had a sunburn, what would be the most appropriate statistical method to analyze the difference, assuming that the number of nevi did not follow a normal distribution in each group?
   A. Two-sample *t* test
   B. Analysis of variance (ANOVA)
   C. Ordinal logistic regression
   D. Wilcoxon rank sum test
5. We are trying to evaluate whether a history of tanning bed use is associated with melanoma risk. Data on the number of months from when a patient is enrolled in the study until the patient is diagnosed with melanoma or censored from the study are available. What is the most appropriate statistical test?
   A. Ordinal logistic regression
   B. Cox proportional hazards model
   C. Linear regression
   D. Chi-square test

discovery rate corrections, and Tukey corrections. Sample size also affects the ability to observe statistically significant results. Studies using a larger sample size may achieve statistical significance at a smaller magnitude of association compared with studies using a smaller sample size. Thus, although uncommon, some studies with small sample sizes have used different *P*-value cutoffs such as 0.10. We would advise investigators to consult a biostatistician to conduct power calculations during the design phase of their clinical studies to determine what sample size would be necessary to achieve statistical significance for a certain magnitude of association.

Although *P*-values are the probabilities of observing an effect size as large as or larger than that detected in a study purely by chance alone, *P*-values do not account for the direction or size of the difference or relative risk in a particular study. In this instance, confidence intervals may provide more information, particularly when the results are not significant. Confidence intervals are a range of possible values for a target population calculated by various statistical methods, including the probability with which this range covers the real value. The probability is usually defined in advance at 95%, meaning that the confidence interval includes the true value in 95 of 100 studies performed. Like *P*-values, the size of the confidence interval will depend on sample size. Larger sample sizes lead to narrower confidence intervals, whereas smaller sample sizes lead to wider confidence intervals. A higher probability of including the true value means the confidence intervals will be wider. Whether there is statistical significance can be assumed by determining if the confidence interval does not include the value of zero effect within the range, such as the value of 0 for difference or a relative risk of 1. Confidence intervals are generally more informative than *P*-values, because they provide information on the certainty of the population estimate of interest (Gardner and Altman, 1986).

Examples of biostatistics in recent *Journal of Investigative Dermatology* literature

Here we have compiled a brief list of studies from recent *Journal of Investigative Dermatology* literature that have exemplified biostatistical methods used in clinical research in dermatology:

- Dusingize JC, Olsen CM, Pandeya NP, Subramaniam P, Thompson BS, Neale RE, et al. Cigarette smoking and the risks of basal cell carcinoma and squamous cell carcinoma. J Invest Dermatol 2017;137:1700–08.
- Egeberg A, Gislason GH, Nast A. Birth outcomes in children fathered by men treated with immunosuppressant drugs before conception—a Danish population-based cohort study. J Invest Dermatol 2017;137:1790–2.
- Hamer MA, Pardo LM, Jacobs LC, Ikram MA, Laven JS, Kayser M, et al. Lifestyle and physiological factors associated with facial wrinkling in men and women. J Invest Dermatol 2017;137:1392–9.
- Walker JL, Siegel JA, Sachar M, Pomerantz H, Chen SC, Swetter SM, et al. 5-fluorouracil for actinic keratosis treatment and chemoprevention: a randomized controlled trial. J Invest Dermatol 2017;137:1367–70.

CONCLUSIONS

In conclusion, we offer an algorithmic approach to help identify the appropriate statistical test for an investigator's clinical research question. Because this reference offers a basic approach to choosing a statistical test in clinical research, this reference cannot capture all of the nuances of the methodology. We strongly encourage readers to take biostatistics and/or epidemiology courses at their local medical centers or online (e.g., Stanford University's self-paced version of *Statistics in Medicine* [https://lagunita.stanford.edu/courses/Home/MedStats/Medicine/about]). Finally, we encourage investigators to collaborate with statistical experts for guidance on study design and appropriate analytic approaches to answer their clinical research questions of interest.

CONFLICT OF INTEREST

The authors state no conflict of interest.

SUPPLEMENTARY MATERIAL

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

REFERENCES

Altman DG. Practical statistics for medical research. Oxford: Chapman and Hall; 1991.

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

Gardner MJ, Altman DG. Confidence intervals rather than P-values: estimation rather than hypothesis testing. Br Med J 1986;292:746–50.

Nambudiri VE, Qureshi A. Comparative effectiveness research. J Invest Dermatol 2013;133:e5.

Rosner B. Fundamentals of Biostatistics, 8th ed. Pacific Grove, CA: Brooks Cole; 2015.

Wakkee M, Hollestein LM, Nijsten T. Multivariable analysis. J Invest Dermatol 2014;134:e20.

Williams HC, Burden-Teh E, Nunn AJ. What is a pragmatic clinical trial? J Invest Dermatol 2015;135:e33.