Study designs in dermatology

Practical applications of study designs and their statistics in dermatology

Jonathan I. Silverberg, MD, PhD, MPH
Chicago, Illinois

Learning objectives
After completing this learning activity, participants should be able to identify the (in)appropriate use of study designs and statistics in dermatology research, describe the levels of evidence for scientific research, and describe how improved study designs have resolved controversies in dermatology.

Disclosures
Editors
The editors involved with this CME activity and all content validation/peer reviewers of the journal-based CME activity have reported no relevant financial relationships with commercial interest(s).

Authors
The authors involved with this journal-based CME activity have reported no relevant financial relationships with commercial interest(s).

Planners
The planners involved with this journal-based CME activity have reported no relevant financial relationships with commercial interest(s). The editorial and education staff involved with this journal-based CME activity have reported no relevant financial relationships with commercial interest(s).

A proper understanding of study designs and related statistical methodology is necessary for clinical dermatologists to critically read scientific literature and incorporate this information into clinical practice. This review focuses on how to identify the appropriate use of study designs and the statistical methodology used therein. Topics covered include population sampling and generalizability, power and sample size calculations, correction for multiple statistical testing, and how to identify the appropriate use of statistics. The impact of improved study designs in previously controversial topics in dermatology will be discussed. (J Am Acad Dermatol 2015;73:733-40.)

Key words: bias; categorical; generalizability; interval-scaled; mean; median; nonparametric; ordinal; standard deviation; standard error of the mean; study design.

INTRODUCTION

A proper understanding of study designs and related statistical methodology is a necessary skill set for the clinical dermatologist. The constant development of cosmeceuticals, novel medications, and procedures demands critical evaluation before incorporating into clinical practice. Dermatologists must be able to recognize the practical limitations of various study designs and critically evaluate the presentation of research data in the dermatology literature. Part I of this continuing medical education article reviewed the fundamentals of study designs and level of evidence. Part II focuses on how to identify the appropriate use of study designs and the statistical methodology used therein. A number of common pitfalls and misuses of statistics previously identified throughout the medical literature[1-3] will be addressed. The impact of improved study designs in previously controversial topics in dermatology will also be discussed.

From the Department of Dermatology, Preventive Medicine and Medical Social Sciences, Northwestern University.
Funding sources: None.
Conflicts of interest: None declared.
Accepted for publication July 27, 2014.
Reprint requests: Jonathan I. Silverberg, MD, PhD, MPH, Department of Dermatology, Northwestern University, 676 N. St. Clair St., Ste 1600, Chicago, IL 60611. E-mail: JonathanISilverberg@gmail.com.

0190-9622/$36.00

http://dx.doi.org/10.1016/j.jaad.2014.07.062
Date of release: November 2015
Expiration date: November 2018

POPULATION AND GENERALIZABILITY

Key points

- **Research studies aim to form a representative sample of the population of interest**
- **Study findings are only generalizable to persons represented by the study sample**

Generally, it is not possible to study the entire population of interest because of time, cost, and/or lack of resources. In many cases, it makes sense to think of a population as being infinite in size (eg, a population of patients past, present, and future who might be given some treatment). Research studies aim to form a representative sample of the population of interest. If the observed sample is truly representative of the population, we can infer that the findings from the observed study groups can be generalized to the population. Selection bias may occur, where subjects may not be representative of the population of interest. For example, a study of patients from a dermatology clinic at an academic medical center may not be reflective of the general population.

Issues of generalizability (external validity) must be considered when interpreting a research study and developing clinical recommendations. For example, mutations of the filaggrin gene were found to be associated with atopic dermatitis (AD). This association has become widely accepted and is the cornerstone for the hypothesis that barrier disruption is the incipient event for AD. Mutations of the filaggrin gene were originally found in persons of Northern European descent and subsequently confirmed in many Asian subpopulations. However, filaggrin mutations are only found in 27.5% of white patients with AD and in 5.8% of African American patients with AD.[4] Therefore, the finding of filaggrin mutations in AD, while clearly associated with a subset of AD, may not be generalizable to all populations of AD patients.

Many RCTs set rigorous inclusion and exclusion criteria, thereby assembling a sample of "perfect patients" that may not be generalizable to the entire population with a disease. RCTs are therefore useful for establishing efficacy, whereas other study designs are more useful for establishing effectiveness (ie, how well a treatment works in practice).

POWER AND SAMPLE SIZE CALCULATION

Key points

- **A type II error occurs when the null hypothesis is falsely retained**
- **Statistical power is the probability of not making a type II error**
- **The more power required, and the lower the significance level, the larger the sample size required**
- **Power analysis and sample size determination should be documented in all observational and interventional studies that use statistical comparisons between groups**

It is important to consider issues of powering and sample size calculations when interpreting the results of a study. It is a fundamental principle that the null hypothesis is rejected with a statistically significant result, but is never proved by a statistically insignificant result.[5] Rather, insignificant results indicate merely that there is insufficient evidence to reject the null hypothesis. It is of course possible to incorrectly retain a null hypothesis that is actually false; this is known as a type II error. With a type II error, an insignificant finding in the observed study group would then be incorrectly generalized to the whole population.

The probability of retaining a null hypothesis if it is false is called beta. Beta values ≤ 0.2 are commonly deemed acceptable (ie, a 20% chance of missing a real difference among populations). Statistical power is the probability of rejecting a null hypothesis if it is false (ie, it measures the ability of a test to correctly reject a false null hypothesis). As a value, power is equivalent to 1 minus the beta value. If a test has high statistical power, it can be asserted with reasonable confidence that a nonsignificant study finding means that population differences are small or zero. On the other hand, a nonsignificant result for a test with low power allows one no confidence to make such a statement.

A priori sample size determination is typically used for designing a study to plan the number of required participants. Reviewers of RCTs will generally insist that this be done. Sample size and power calculations can be quite complex and are typically based on the primary study outcome and are specific to a particular statistical design. Briefly, information needed in order to conduct sample size calculations generally includes specification of: (1) study design; (2) test type; (3) whether the intent is to establish that populations do or do not differ; (4) significance level; (5) whether any difference must be in one direction or could be in both directions; (6) desired power; (7) estimate of how different the populations actually are; and (8) estimate of the extent of within-arm variability. Post hoc power determination can be performed after the completion of a study in scenarios where a priori sample size calculation was not performed or to determine the power for a secondary study