Generalizability

The trees, the forest, and the low-hanging fruit

Walter A. Kukull, PhD
Mary Ganguli, MD, MPH

Correspondence & reprint requests to Dr. Kukull: kukull@u.washington.edu

ABSTRACT

Clinical and epidemiologic investigations are paying increasing attention to the critical constructs of "representativeness" of study samples and "generalizability" of study results. This is a laudable trend and yet, these key concepts are often misconstrued and conflated, masking the central issues of internal and external validity. The authors define these issues and demonstrate how they are related to one another and to generalizability. Providing examples, they identify threats to validity from different forms of bias and confounding. They also lay out relevant practical issues in study design, from sample selection to assessment of exposures, in both clinic-based and population-based settings. ***Neurology***® **2012;78:1886–1891**

GLOSSARY

AD = Alzheimer disease; **ADRC** = Alzheimer's Disease Research Center; **PD** = Parkinson disease.

> Only to the extent we are able to explain empirical facts can we obtain the major objective of scientific research, namely not merely to record the phenomena of our experience, but to learn from them, by basing upon them theoretical generalizations which enable us to anticipate new occurrences and to control, at least to some extent, the changes in our environment.[1(p12)]

"This study sample is not representative of the population!" "Our results are not generalizable …" Such comments are increasingly familiar but what exactly do they mean? How do study design, subject ascertainment, and "representativeness" of a sample affect "generalizability" of results? Do study results generalize only from statistically drawn samples of a common underlying population? Has "lack of generalizability" become the low-hanging fruit, ripe for plucking by the casual critic?

INTERNAL AND EXTERNAL VALIDITY Confusion around generalizability has arisen from the conflation of 2 fundamental questions. First, are the results of the study true, or are they an artifact of the way the study was designed or conducted; i.e., is the study is internally valid? Second, are the study results likely to apply, generally or specifically, in other study settings or samples; i.e., are the study results externally valid?

Thoughtful study design, careful data collection, and appropriate statistical analysis are at the core of any study's internal validity. Whether or not those internally valid results will then broadly "generalize," to other study settings, samples, or populations, is as much a matter of judgment as of statistical inference. The generalizability of a study's results depends on the researcher's ability to separate the "relevant" from the "irrelevant" facts of the study, and then carry forward a judgment about the relevant facts,[2] which would be easy if we always knew what might eventually turn out to be relevant. After all, we generalize results from animal studies to humans, if the common biologic process or disease mechanism is "relevant" and species is relatively "irrelevant." We also draw broad inferences from randomized controlled trials, even though these studies often have specific inclusion and exclusion criteria, rather than being population probability samples. In other words, generalization is the "big picture" interpretation of a study's results once they are determined to be internally valid.

SAMPLING AND REPRESENTATIVENESS The statistical concepts of sampling theory and hypothesis testing have become intermingled with the notion of generalizability. Strict estimation of quantities based on a probability sample of a "population," vs assessing all members of that population, remained an object of

CME

From the Department of Epidemiology (W.A.K.), University of Washington School of Public Health, Seattle; and Departments of Psychiatry, Neurology, and Epidemiology (M.G.), University of Pittsburgh School of Medicine and Graduate School of Public Health, Pittsburgh, PA.

Study funding: Supported in part by NIH grants U01 AG016976 (W.A.K.) and K24AG022035 (M.G.).

Go to Neurology.org for full disclosures. Disclosures deemed relevant by the authors, if any, are provided at the end of this article.

considerable argument among statisticians until the early 20th century.[3] Sampling was adopted of necessity because studying the entire population was not feasible. Fair samples must provide valid estimates of the population characteristics being studied. This quite reasonable concept evolved in common usage so that "population" became synonymous with "all persons or all cases." It followed that to achieve representative and generalizable sample estimates, a probability sample of "all" must be drawn. Logically, then, "all" must somehow be enumerated before representative samples can be drawn. The bite of the vicious circle becomes obvious when "all" literally means all in a country or continent. Yet enumeration may be achievable when care is taken to establish more finite population boundaries.

Statisticians Kruskal and Mosteller[3–6] conducted a detailed examination of nonscientific, "extrastatistical scientific," and statistical literature to classify uses of the term *representative sample or sampling.* Those meanings are 1) "general, unjustified acclaim for the data"; 2) "absence (or presence) of selective forces"; 3) "mirror or miniature of the population"; 4) "typical or ideal case … that represents it (the population) on average"; 5) "coverage of the population … (sample) containing at least one item from each stratum …"; 6) "a vague term to be made precise" by specification of a particular statistical sampling scheme, e.g., simple random sampling. In statistical literature, representative sampling meanings include a) "a specific sampling method"; b) "permitting good estimation"; and c) "good enough for a particular purpose."[4] The conflicts and ambiguities among the above uses are obvious, but how do we seek clarity in our research discourse?

POPULATIONS, CLINICS, AND BOUNDARIES

So is there in fact any value to population-based studies (Indeed there is!), and if so, how should we define a "population"? We first define it by establishing its boundaries (e.g., counties, insurance memberships, schools, voter registration lists). The population is made up entirely of members with disease (cases) and members without disease (noncases), leaving nobody out. Ideally, we would capture and study all cases, as they occur. As a comparison group, we would also include either all noncases, or a probability sample of noncases.[7] The choice of "boundaries" for a study population influences internal and external validity. If we deliberately or inadvertently "gerrymander" our boundaries, so that the factor of interest is more (or less) common among cases than among noncases, the study base will be biased and our results will be spurious or misleading.

Adequately designed population-based studies minimize the possibility that selection factors will have unintended adverse consequences on the study results. Further, since any effect we might measure depends as much on the comparison group as it does on the case group, appropriate selection is no less important for the noncases than it is for cases. This is true whether the study is clinic-based or population-based. Population-based research anchors the comparison group to the cases.

Clinic-based investigations are exemplified by those conducted at Alzheimer's Disease Research Centers (ADRCs). They typically examine high-risk, family-based, clinic-based, or hospital-based groups, to observe association with treatment or disease. This is an efficient means to facilitate in-depth study of "clean" diagnostic subgroups. The external validity of these studies rests on the judgment of whether the subject selection process itself could have spuriously influenced the results. This determination is often harder in clinic-based studies than in population-based studies. Replication in an independent sample is therefore key, but replication is more elusive and difficult with clinic-based studies, as we discuss later.

Regardless of whether the study sample is clinic-based or population-based, how well and completely we identify "disease" (including preclinical or asymptomatic disease), not only in our case group, but also among those in our comparison group, can adversely impact results. For example, consider a study of Alzheimer disease (AD) in which, unbeknownst to the subjects as well as the investigators, the cognitively normal control group includes a large proportion of persons with underlying AD pathology. The resulting diagnostic misclassification, caused by including true "cases" among the noncases, would spuriously distort and weaken the observed results. This distortion can happen in clinic-based or population-based studies; it is a matter of internal validity tied to diagnostic accuracy, rather than an issue of representativeness or generalizability.

BIAS

Bias causes observed measurements or results to differ from their true values because of systematic, but unintended, "errors," for example, in the way we ascertain and enroll study subjects (selection bias), or the way we collect data from them (information bias). Statistical significance of study results, regardless of p value, is completely irrelevant as a means of evaluating results when bias is active.

Selection bias. Selection bias is often subtle, and requires careful thought to discern its potential effect on the hypotheses being tested. For example, would selection bias render clinic-based ADRC study results suspect, if not invalid? Unfortunately, the answer is not simple; it depends on what is being studied and whether "selection" into the ADRC study distorts the true association. There are numerous advantages to recruiting study participants from specialized memory disorder clinics, as in the typical ADRC. Both AD cases and healthy controls are selected (as volunteers or referrals) under very specific circumstances that ensure their contribution to AD research. They either have (cases) or do not have (controls) the clinical/pathologic features typical of AD. Cases fulfill the research diagnostic criteria for AD, they have "reliable informants" who will accompany them to clinic visits; neither cases nor controls can have various exclusionary features (e.g., comorbid stroke or major psychiatric disorder); all are motivated to come to the clinic and participate fully in the research, including neuroimaging and lumbar puncture; many are eager to enter clinical trials, and many consent to eventual autopsy. AD cases who fit the above profile are admirable for their enthusiasm and altruism, but may not be typical, nor a probability sample of all AD cases in the population base from whence they came. The differential distribution of study factors between AD cases who did and did not enroll could give us an indication of whether bias may be attenuating or exaggerating the specific study results, if we were able to obtain that information. Therefore, the astute reader asks*:* "Can the underlying population base, from which the subjects came, be described? Might the population base's established boundaries or inclusion characteristics have influenced the results? Was subject enrollment in any way influenced by the factors being studied?" In a clinic-based study it is seldom easy to describe the unenrolled cases (or unenrolled noncases) from the underlying population base in order to make such comparisons. It helps internal validity very little to claim that the enrollees' age, race, and sex distributions are in similar proportions to the population of the surrounding county, if age, race, and sex have little to do with the factor being studied, and if participation is differentially associated with the factors being studied.

Note that population-based studies are not inherently protected from bias; individuals sampled from the community, who are not seeking services, may consent or refuse to participate in research, and their willingness to participate is unlikely to be random. If we were concerned about selection bias in a study examining pesticide exposure as a risk factor for Parkinson disease (PD), we might ask, "Were PD cases who had not been exposed to pesticides more (or less) likely to refuse enrollment in our study than PD cases who had been exposed?"

Selection bias may be not just inadvertent but also unavoidable. Some years ago, a startling finding[8] was reported that AD cases who volunteered or were referred to an ADRC were significantly more likely to carry the *APOE*4* genotype than were newly recognized AD cases captured through surveillance of a health maintenance organization population base within the same metropolitan area. The ADRC sample had yielded a biased overestimate of *APOE*4* allele frequency, and of its estimated relative risk, because ADRC cases were inadvertently selected on the basis of age, and it was unnoticed that the likelihood of carrying an *APOE*4* allele decreases with age. There is no way the ADRC investigators could have detected this inadvertent selection bias had they not also had access to a population sample from the same base. A later meta-analysis of *APOE*4* allele effects quantified the relationship between age and risk of AD associated with *APOE* alleles, and showed that AD risk due to *APOE*4* genotype is lower in population samples than in specialty clinic samples.[9] *APOE* allele frequency also could be influenced by study recruitment. Family history of AD seems to promote participation in both clinical and population-based studies involving memory loss, and is also associated with *APOE*4* frequency, thereby potentially biasing the magnitude of *APOE* effect.

Survival bias is a form of selection bias that is beyond the control of the selector. For example, some African populations have high *APOE*4* frequency but have not shown an elevated association between *APOE*4* and AD.[10,11] While there could be multiple reasons for this paradox, one possibility is that individuals with the *APOE*4* genotype had died of heart disease before growing old enough to develop dementia.

Prevalence bias (length bias) is similar to survival bias. In the 1990s, numerous case-control studies showed a protective effect of smoking on AD occurrence.[12] Assume that both AD and smoking shorten life expectancy and that AD cases enrolled in those studies some time after symptom onset. If age alone was the basis for potential selection bias, smoking should cause premature mortality equally among those who are and those who are not destined to develop AD. However, there is another aspect of selection bias called prevalence or length bias: at any given time, prevalent, i.e., existing, cases are those whose survival with disease (disease duration) was of greater length. If smokers with AD die sooner after AD onset than nonsmokers with AD, those prevalent AD cases available for study would "selectively" be nonsmokers. A scenario known as "competing risks" occurs when smoking influences the risk both of death and of AD.[13] This would enhance the observed excess of smoking among "controls" and thereby inflate the apparent protective association between smoking and AD. Subsequently, longitudinal studies of smokers and nonsmokers showed an increased risk of AD incidence associated with smoking,[12] suggesting that selection bias might have explained the earlier cross-sectional study results.

Information bias. Information bias (data inaccuracy) can occur if we measure or determine the outcome or the exposure with substantial error or if the outcome or exposure is measured differently between comparison groups. Here, the reader must ask "Was information on the study factors and covariates gathered in a fair and equal manner for all subjects?" For example, suppose we obtain the history of previous head trauma, from spouses of the cases, but by self-report from the controls. The frequency of head trauma could be systematically different between groups because of whom we asked, rather than because of their true occurrence. Many earlier case-control studies showed an association between AD and previous history of head trauma.[14] This finding was not replicated in a subsequent study based on prospective data from a comprehensive population-based record-linkage system.[15] Here, data about head injury were recorded in the same way from all subjects before the onset of dementia; when both selection bias (including length bias) and information bias were eliminated, the association was no longer present. More recently the issue has raised its battered head once again, but such studies should also be mindful of the methodologic lessons of the past.

CONFOUNDING Having done our best to avoid bias, how do we account for the simultaneous effects of other factors that could also cause the disease? Consider a study of diabetes as a risk factor for cognitive decline. Both diabetes and cognitive decline are associated with age, family history, and cerebrovascular disease. The effects of these other factors could distort our results, if they were unequally distributed between those with and without diabetes. This mixing of effects is called confounding. Similarly, in designing a study examining pesticide exposure as a risk factor for PD, we would be concerned about other risk or protective factors for PD which might themselves be associated with pesticide use.[16] A common additional exposure in rural farming areas is head trauma,[17] which arguably may increase risk of PD.[18] If head trauma was a causal risk factor and was distributed unequally between the pesticide-exposed and nonexposed groups, a spurious impression could be created about the risk associated with pesticide exposure.

If we proactively collected data on potential confounders, their effects could be "adjusted for" (equalized statistically between comparison groups) in the analysis, and can be similarly be "adjusted" in replication studies. Adjustment indicates ceteris paribus (holding all else constant): it statistically equalizes or removes the effect of the confounding factors (e.g., head trauma) so that the factor of interest (e.g., pesticide exposure) can be evaluated for its own effect. Note: bias (unlike confounding) can rarely be adjusted away.

REPLICATION Replication of results in independent samples supports both the internal validity and the generalizability of the original finding, and is now required for publication of genetic association studies. If 2 similar studies' results do not agree, one does not necessarily refute the other; however, several similar studies failing to replicate the original would weigh heavily against the original result. We do not expect all risk factor studies to have identical results because risk factor frequencies may be differentially distributed among populations. Sample variability does not rule out generalizability, a priori, but the potential effects of bias and confounding must not be ignored.

GENERALIZABILITY AND POWER Finally, another issue often wrongly subsumed under generalizability is related to the statistical power to observe an association if one truly exists. For example, a study of head trauma as a risk factor for dementia should be carried out in a sample where there is both sufficient head trauma and sufficient dementia for an association (if present) to be detected. A sample of young football players may have the former but not the latter[19]; a sample of elderly nuns[20] may have the latter but not the former; a sample of retired football players may have both[21]; a sample of aging military veterans may also have both, but there may be potential confounding factors associated with military service, such as other injuries, depression, or post-traumatic stress.[22] Thus, studies in different samples may not replicate one another's results with regard to head trauma and dementia not because the association changes but because of varying exposure or outcome frequency.

THE SMOKING GUN We close with the one of the most influential articles of the 20th century, to demonstrate how even very narrowly defined study samples may provide widely generalizable results if conducted with an eye to rigorous internal validity. Entitled "The mortality of doctors in relation to their smoking habits: a preliminary report,"[23] this 1954 article by Doll and Hill concerned the association between lung cancer and cigarette smoking in British physicians. All 59,600 physicians in the Medical Register at the time were sent a questionnaire on their smoking habits. The investigators excluded physicians who did not return usable responses, and also women physicians, and physicians aged <35 years, because of their low expected frequency of lung cancer deaths. The remaining sample was a male, physician cohort of 24,389, about 40% of those in the Medical Register. During the 29-month follow-up, investigators observed only 36 confirmed lung cancer deaths, occurring at rates of 0.00 per 1,000 in nonsmokers, and 1.14 per 1,000 among smokers of 25 or more grams of tobacco per day. The lung cancer death rate was dose-dependent on amount smoked, but the same relationship with tobacco dose was observed neither for 5 other disease comparison groups, nor for all causes of death. Further, study cohort had an all-cause death rate of 14.0 per 1,000 per year as compared to 24.6 per 1,000 for men of all social classes and similar age.[23]

Surely, that study provided a veritable feast of low-hanging fruit for critics focused on generalizability. With such a select study sample, would the results not be so specific and isolated that none would generalize to groups other than male British physicians? Undaunted, Doll and Hill focused on internal validity, considering whether their study-defined boundaries and method of subject selection could have created a spurious association between smoking and lung cancer death. They reasoned that the initially nonresponding physicians may have overrepresented those already close to death, causing the observed death rate in the short term to be lower than the general population. More importantly, they asked whether such a difference in mortality within their sample could have caused the dose-response "gradient" between amount smoked and lung cancer death rate. "For such an effect we should have to suppose that the heavier smokers who already knew that they had cancer of the lung tended to reply more often than the nonsmokers or lighter smokers in a similar situation. That would not seem probable to us."[23(p1454)] This study has been replicated in many other population- and clinic-based studies. It has been generalized, in the broad scientific sense, to a variety of other groups, populations, and settings, despite the decidedly "nonrepresentative" nature of the study group and its specific boundaries. Doll and Hill focused on how, and how much, the definition of their study group and its characteristics could have influenced their results. That is, they considered how the effects of subject selection (i.e., selection bias), data accuracy (i.e., information bias), and unequal distribution of other risk/protective factors between comparison groups (i.e., confounding) could have threatened the study's internal validity. They also considered "power" when they excluded younger men and women. In this study, the "relevant" factor concerned the potential carcinogenic effect of tobacco smoke on human lung tissue.

Would the designs and findings of similar studies among restricted groups, nonrepresentative of the universe, be as readily accepted today? Would current readers question whether the results from British physicians would also apply to Wichita linemen or to real housewives from New Jersey? The British physicians were likely different in many ways from groups to which we might want to "generalize" the principal results. But they were not fundamentally different in ways that would affect our conclusions about the effect of tobacco smoke on lung tissue and ultimate mortality.

Science proceeds by replication and by generalization of individual study results into broader hypotheses, theories, or conclusions of fact. Establishing study boundaries and conducting "population-based" research within them enhances both internal validity and the likelihood that results may apply to similar and dissimilar groups. However, studies of specifically defined groups may also generalize to extend our knowledge. We could yield to temptation and seize the low-hanging fruit, vaguely challenging a study on grounds of generalizability. But then we would miss the forest for the trees.

AUTHOR CONTRIBUTIONS

Walter A. Kukull, PhD, made substantive contribution to the design and conceptualization and interpretations and was responsible for the initial draft and revising the manuscript. Mary Ganguli, MD, MPH, made substantive contribution to the design and conceptualization and interpretations and contributed to revising the manuscript.

DISCLOSURE

The authors report no disclosures relevant to the manuscript. **Go to Neurology.org for full disclosures.**

Received October 17, 2011. Accepted in final form January 23, 2012.

REFERENCES

1. Brody BA. Readings in the Philosophy of Science. Englewood Cliffs, NJ: Prentice-Hall; 1970.
2. Rothman KJ, Greenland S, Lash TL. Modern Epidemiology, 3rd ed. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2008.
3. Kruskal W, Mosteller F. Representative sampling: 4: the history of the concept in statistics, 1895–1939. Int Stat Rev 1980;48:169–195.
4. Kruskal W, Mosteller F. Representative sampling: 3: current statistical literature. Int Stat Rev 1979;47:245–265.
5. Kruskal W, Mosteller F. Representative sampling: 2: scientific literature, excluding statistics. Int Stat Rev 1979;47: 111–127.
6. Kruskal W, Mosteller F. Representative sampling: 1: nonscientific literature. Int Stat Rev 1979;47:13–24.
7. Koepsell TD, Weiss NS. Epidemiologic Methods: Studying the Occurrence of Illness. Oxford: Oxford University Press; 2003.

8. Tsuang D, Kukull W, Sheppard L, et al. Impact of sample selection on APOE epsilon 4 allele frequency: a comparison of two Alzheimer's disease samples. J Am Geriatr Soc 1996;44:704–707.
9. Farrer LA, Cupples LA, Haines JL, et al. Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease: a meta-analysis: APOE and Alzheimer Disease Meta Analysis Consortium. JAMA 1997;278:1349–1356.
10. Hendrie HC, Ogunniyi A, Hall KS, et al. Incidence of dementia and Alzheimer disease in 2 communities: Yoruba residing in Ibadan, Nigeria, and African Americans residing in Indianapolis, Indiana. JAMA 2001;285:739–747.
11. Ogunniyi A, Baiyewu O, Gureje O, et al. Epidemiology of dementia in Nigeria: results from the Indianapolis-Ibadan study. Eur J Neurol 2000;7:485–490.
12. Kukull WA. The association between smoking and Alzheimer's disease: effects of study design and bias. Biol Psychiatry 2001;49:194–199.
13. Chang CC, Zhao Y, Lee CW, Ganguli M. Smoking, death, and Alzheimer disease: a case of competing risks. Alzheimer Dis Assoc Disord Epub 2011.
14. Mortimer JA, van Duijn CM, Chandra V, et al. Head trauma as a risk factor for Alzheimer's disease: a collaborative reanalysis of case-control studies: EURODEM Risk Factors Research Group. Int J Epidemiol 1991;20(suppl 2):S28–S35.
15. Chandra V, Kokmen E, Schoenberg BS, Beard CM. Head trauma with loss of consciousness as a risk factor for Alzheimer's disease. Neurology 1989;39:1576–1578.
16. Tanner CM, Kamel F, Ross GW, et al. Rotenone, paraquat, and Parkinson's disease. Environ Health Perspect 2011;119:866–872.
17. Seidler A, Hellenbrand W, Robra BP, et al. Possible environmental, occupational, and other etiologic factors for Parkinson's disease: a case-control study in Germany. Neurology 1996;46:1275–1284.
18. Goldman SM, Tanner CM, Oakes D, Bhudhikanok GS, Gupta A, Langston JW. Head injury and Parkinson's disease risk in twins. Ann Neurol 2006;60:65–72.
19. Macciocchi SN, Barth JT, Alves W, Rimel RW, Jane JA. Neuropsychological functioning and recovery after mild head injury in collegiate athletes. Neurosurgery 1996;39:510–514.
20. Snowdon DA. Healthy aging and dementia: findings from the Nun Study. Ann Intern Med 2003;139:450–454.
21. Guskiewicz KM, Marshall SW, Bailes J, et al. Association between recurrent concussion and late-life cognitive impairment in retired professional football players. Neurosurgery 2005;57:719–726.
22. Plassman BL, Havlik RJ, Steffens DC, et al. Documented head injury in early adulthood and risk of Alzheimer's disease and other dementias. Neurology 2000;55:1158–1166.
23. Doll R, Hill AB. The mortality of doctors in relation to their smoking habits: a preliminary report. BMJ 1954;1:1451–1455.