

If Not Now, It's Too Late: Clinical Science Is Futile If We Study the Wrong Population

written by Robert McNutt, M.D. | March 7, 2019



In 1936, the *Literary Digest*, a respected national magazine, undertook a public opinion poll. Who would win the race between Republican Alfred Landon, governor of Kansas, and Democratic incumbent Franklin D. Roosevelt? Mock ballots were mailed to 10 million Americans. About 2.4 million responded—one of the largest survey samples ever created. Their prediction? Landon would carry the day.

They were wrong—by a landslide for FDR. That's because respondents were biased toward Landon and did not accurately represent the distribution of presidential preferences across all voters. Notably, George Gallop accurately predicted FDR's victory using a smaller representative sample of about 50,000 people.

While that slice of presidential election history provides an excellent example of polling error, it also illustrates a significant issue in clinical science: reliable and accurate extrapolations from clinical studies depend upon the relevance of the studied population to those for whom the results are generalized.

Who Gets Studied Defines Relevance of Results

Clinical science seeks to find statistically and clinically significant differences in outcome probabilities between one plan and another. The *size* and *accuracy* of the measured difference, and the ability to use that difference number to *extrapolate* a study's result, are the paramount duties of clinical science.

As researchers plan clinical science, much time and effort is devoted to plotting study design, ruminating about statistical “power” and presaging interpretation. These considerations reveal a focus on “internal validity,” which denotes how well the study, *per se*, is done. As a result, the “population studied” can become an after thought—not ignored, but subjugated—a major reason that studies fail to properly inform.

Patients in a study are a part of an “eligible to be studied” whole. Clinical science uses information from the partial, studied group to infer to the whole. If the part does not share characteristics of the whole, inference is weak, or wrong. A study with flawless internally validity means little if results do not translate. An egregious example is the overreliance on clinical studies of heart disease among men as the basis for treatment protocols for women, despite [gender disparities in how women experience heart disease](#) and what treatments work best for them.

The Gap Between Eligible and Accepted Populations Can Limit Study Applicability

Let's break this down further. There are three populations in research:

The *entire* population of patients with the condition who are eligible to be studied.

The part of the whole *invited* into a study.

The group of invited patients who *accept* being in the study.

Unfortunately, the path from the entire eligible population to the group that actually participates may lead to a clinical trial with limited applicability. As an editor I once saw a group of nearly 10,000 eligible people drop to only 300 accepting to participate in the study. Based on the numbers, alone, it's hard to imagine that 300 could properly inform the 10,000.

Whether the small part can be generalized to the whole begs a question: Is the best study one with large numbers of patients? We've already seen the flaws in the 1936 *Literary Digest* poll, one of the largest studies ever done. How about a contemporary clinical science example? In 2011, the [National Lung Screening Trial \(NLST\)](#) was published; 53,454 patients were enrolled

from 33 centers to test if CT scanning saved lives from lung cancer.

As with all trials, inclusion and exclusion criteria created the population of patients eligible for study; but we don't know if all potential patients who would have been eligible were taken into account at outset of the trial. Also, there is no description, either in the published report or in the study protocol sent to ClinicalTrials.gov, of *how* patients from the eligible population were invited. We don't know if a consecutive sample was invited, a systematic sample scheme was used, a random sample of eligible patients was asked, or if doctors picked whom to ask.

This failure to *describe* which people got invited and who accepted being in the study is a profound omission—even, I would argue, a disqualifying omission, in terms of using the study to make policy or patient decisions. If the people were haphazardly, rather than systematically, invited, the large sample is nothing more than a convenience sample of handpicked patients. Random assignment to treatments after haphazard recruitment does not help us generalize results. It would be better to have a random sample of all eligible people at outset.

Large Sample Size Does Not Guarantee More Accurate Results

Is there evidence that the NLST study population was not generalizable, and, therefore, of limited value to individual patients?

After publication, CT scanning was promoted based on the trial results, and centers began screening. The experiences of other sites did not replicate the NLST findings. For example, the Veterans Administration found that their screened patients were older than those in the NLST (53 percent over 65 years of age versus 27 percent for the NLST), were more likely to be current smokers, had more abnormalities on CT requiring follow-up, found fewer lower stage cancers, and had a [complication rate over twice as high](#) as reported in the NLST. They also noted variations in patients' experiences with outcomes, process, costs and complications across 8 study sites, none with results similar to the NLST.

Large studies are large for a reason. There is little anticipated difference in the outcomes of a randomized trial; the base rates for outcomes are small. Some argue that a random sample of a population is not needed when the base rates of outcome events are small, but the examples above nix that debate. Outcome rates, especially complication rates, vary by patients' clinical and personal characteristics, and their means.

Any large study, including the NLST, that fails to include all eligible patients or fails to randomly invite people from all who are eligible is off on the wrong foot from the get-go. Simple random

samples of patients may also be inadequate for the future advancement of clinical studies—but that is a topic for a future post.

Clinical research must be more like the 1936 Gallop poll than the convenience sampling of even huge numbers of people. If clinical science can't get the right population to study at outset, advancing care via science will be slow and dangerous to some. Generalizability, not internal validity, should dominate study planning.

Founded as ICLOPS in 2002, Roji Health Intelligence guides health care systems, providers and patients on the path to better health through [Solutions](#) that help providers improve their value and succeed in Risk. Roji Health Intelligence is a CMS Qualified Clinical Data Registry.

Image: [chuttersnap](#)