If Not Now, It's Too Late: Simple Randomization Can Lead to False Inferences About Treatment Decisions

written by Robert McNutt, M.D. | April 4, 2019

Medical decisions are best made on the basis of clinical science. Accurate research, shared between physician and patient, enables the patient to make an informed choice about risks and outcomes of treatment options.

That's how it should work, in theory. But in practice, even with the best shared medical decision-making, far too much clinical research employs faulty methodologies that limit the relevance of findings. This must change.

In a recent blog post, I suggested that clinical science can improve by choosing more representative groups of people for study.
Many clinical studies use convenience samples of patients rather than samples chosen either randomly or systematically from full populations. This compromises our ability to generalize insights from a sample to the full population.

But the methodological flaws don't end there. The next problem with clinical science is how we randomize patients after the population to be studied is constituted. Simply randomizing people, like a coin flip, to a new treatment group versus the usual comparison group potentially fails on two aspects: First, it blunts our ability to inform individuals in the trial and the population in general who vary on clinical and personal characteristics, and, second, simple randomization fails to assure balance in factors that may influence interpretation.

Remember, science aims to find the *independent contribution* of a new therapy over another. This requires a comparison of the frequency of clinical outcomes between two groups. If one of the groups has more people who are ill, for example, then a comparison of the new treatment is weakened, since any difference we find may be due to the imbalance in prognostic factors rather than the treatment being tested. Randomization intends to balance these factors.

Simply Randomizing Gives an Average Difference but Little Help to Individuals

Let's address the first issue with an example. A woman is 64, has invasive breast cancer, two nodes positive. Her estrogen /progesterone receptor status (ER/PR) is positive; true for most cancer cells. Additionally, 100 percent of her cancer cells are positive for HER2, another tumor maker. She did not take the usual chemotherapy, opting to choose a less toxic regimen than the usual. Now she asks if she should take the drug for HER2.

Her uniqueness may influence the decision. In a randomized trial (RT) of nearly 5,000 women regarding the HER2 drug in question versus placebo, those taking the drug did better in terms of recurrence and survival. So, she has information about benefit and harm, *on average*. How did those in the trial *who are like her do*? We don't know. Only 200 or so of the 5,000 subjects did not take the first line chemotherapy, and there is no reported distribution of the percent receptor positivity for ER/PR/HER2. Hence, I could not isolate "her" unique profile of age, node status, receptor status and level, and prior treatment status.

There is a hint in the data that ER/PR status modified the effect of the drug for HER2, but the trial was too small to say for certain, and, additionally, the distribution of breast-cancer-related outcomes for those who did not take the usual chemotherapy prior to the RT varied from somewhat better to 1½ times worse. Hence, this woman is uncertain of the benefit of the HER2 drug.

After this RT, the drug became the usual protocol for women with HER2, but it is not an easy drug to take; it is costly and comes with significant side-effects. Randomization did not capture enough of the variation in the people involved in the trial. More was needed from this study to

help this woman.

The “more” is called *stratified randomization*. People with characteristics that might influence the outcome of a RT are identified at outset and grouped; then, randomization is carried out within groups. For example, since previous evidence shows that people are helped by treatment with estrogen-blocking agents when ER/PR markers are positive, grouping women with various combinations of ER/PR/HER2 could have been completed prior to randomizing. If the benefit of the drug is the same for all groups, a general inference can be made. If the drug helps different groups differently, however, then those characteristics are said to modify the value of the treatment being tested in the trial. And, if the new drug offers little, perchance, for those who were ER/PR positive, women with those characteristics would have useful information for them.

In short, simple randomization is a flawed methodology for a practice of medicine aimed to care for individuals. There is no assurance that there will be a significant number of patients in important prognostic subgroups to offer relevant insights to those individuals.

Simple Randomization May Fail to Balance Prognostic Factors

Besides failing to address individuals’ variations, simple randomization may fail to assure that what we study is better *for anyone*. This is a contentious statement. The RT is purported to be the gold standard of clinical science. The following examples show, however, that important prognostic factors may not balance enough for the clinician and patient to make a reasoned inference. This phenomenon is called “chance bias” and is underappreciated as a flaw of RT’s.

A striking example of imbalance in prognostic characteristics was a RT of a drug for patients in the intensive care unit (ICU). The new drug (activated protein C) versus placebo benefited patients; those who got the drug were more likely to survive. Researchers presumed that conditions that adversely affect survival were balanced.

However, these factors did not balance. For eight of nine measured prognostic factors that portend a poor outcome, more people with these got the *placebo*. For example, 3 percent more with hypertension got the placebo, 2 percent more with a myocardial infarction, 2 percent more with cancer, 5 percent more with liver disease, 5 percent more on mechanical ventilation, and others. This imbalance makes the drug look better, but better was an illusion, born out by future studies.

Chance bias is common with simple randomization. A recent expose evaluated ten top-cited

RT's and found chance imbalance in nearly all. (Journals track how often a paper is referenced, cited, by other papers. The theory is that more impactful published ideas get cited more often). For example, a top-cited trial of stroke care found that factors that affected mortality after a stroke were not balanced: study patients allocated to the new treatment were less ill (3 percent fewer had congestive heart failure, 8 percent fewer were smoking, 14 percent more took aspirin therapy, as examples). These unbalanced factors cloud a trial's main outcomes.

And this problem extends to the largest trials. It is assumed that trials with large numbers of patients are best. This is not true; trials with small and large numbers of patients face the same chance problem. Large trials are planned to look for small differences in the numbers of patients with measured outcomes; it follows that even small imbalances can negate small differences in outcomes. For example, in the National Lung Screening Trial, discussed in the previous blog, the difference in the number of people dying of lung cancer between the CT scan group versus the CXR comparison group was 76 people (out of 50,000+). However, 26 more people were in older age groups in the CXR arm of the study and 38 more were current smokers. These sound like small amounts, but adding those two imbalanced subgroups equals 64 more people with adverse prognostic characteristics in the CXR group to compare to the 76 with better outcomes. This sort of imbalance is worrisome, even in large trials.

When I read a paper, I count the number of people who differ in all prognostic groups and find, too often, imbalanced numbers of people. When I compare the number of imbalanced people to the total number of people different between studied groups, the numbers are often close to each other. This should raise concern for the veracity of the RT. Statisticians, sometimes, look to see if imbalance is a factor, but these types of analyses are often weak due to small numbers of subjects.

A better solution than statistical or counting efforts to look for imbalance after a RT is to *stratify and randomize* only after planning for people with known confounding clinical factors and combinations of factors that make a difference to individual people. Future clinical trials must consider how best to randomize in order to help people who are not like the average in the RT to make informed choices, and researchers must, before doing a trial, make sure prognosis is balanced. The present methods of clinical science are, unfortunately, not good enough for individual patients who must decide.

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: Nick Fewings