

Drug trial phases

David L. Streiner, PhD, CPsych,¹ and Geoffrey R. Norman, PhD²

¹Department of Psychiatry, University of Toronto, Senior Scientist, Kunin-Lunenfeld Applied Research Unit Baycrest Centre, Toronto, Ontario, Canada, ²Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada

All drugs go through three, and sometimes four, testing phases. This paper discusses what those phases are and why they're needed.

Editor's note

One part irreverence to two parts statistics

COMMUNITY ONCOLOGISTS are increasingly asked to practice evidence-based medicine. But what is evidence-based medicine other than studies in certain patient populations, from which we are asked to make conclusions for our patients who may or may not exactly fit that population? Enter the dreaded biostatistics. We all have some basic understanding of what a *P* value means, if it is greater or less than 0.05. But who among us could differentiate between a randomized phase II or III trial, write a noninferiority trial, or perform a calculation to determine the power of a study?

Since understanding statistical jargon and concepts has become such an important part of what we do, the editors of *Community Oncology* have decided to run a continuing column to introduce and review practical statistics for oncology practitioners. In the months ahead, this series will cover many of the common types of studies, integral to evidence-based medicine, and help us better interpret and apply these data and trials.

Our authors, Drs. David L. Streiner and Geoffrey R. Norman, a University of Toronto professor in the department of psychiatry, and McMaster University professor of clinical epidemiology and biostatistics, respectively, are much beloved by their students as—and I quote—“the only people on the planet who can make me laugh when trying to figure out statistics.”

Enjoy the series. You will learn something valuable—and have fun doing it.

— David H. Henry, MD, FACP
Pennsylvania Hospital, Philadelphia, PA

Before a drug comes on the market, it has gone through three different phases of testing with humans, denoted, with utmost originality, I, II, and III. After some drugs have been on the market for some time, there may or may not be a phase IV trial. We presume that the Roman numerals are used to convey a sense of gravity and seriousness that wouldn't be given by Arabic numbers; their use cannot be justified on the basis of the number of trials done in ancient Rome. These phases are stipulated by the US Food and Drug Administration (FDA) to “assure the safety and rights of subjects.” So, what are those phases, and how do they differ from one another?

Phase I trials

Even though the term “phase I” makes it sound like it is the first phase in drug testing, it actually isn't. Before an agent makes it to this stage, there have been preclinical studies, in which the drug has been tested in

vitro and in vivo in animals to see whether it has promise in terms of shrinking tumors or achieving whatever endpoint we're looking for (although it may be difficult for mice to fill out quality-of-life scales). If the results look good, the company files an investigational new drug (IND) request with the FDA; if approved, it can go on to a phase I study. Phase I means that it is the first phase in testing with humans.

Almost all phase I trials are alike, in that they are independent of the intended use of the drug; that is, whether the drug is used to shrink a tumor,

Manuscript received November 14, 2008; accepted December 4, 2008.

Correspondence to: David L. Streiner, PhD, CPsych, Professor, Department of Psychiatry, University of Toronto, Senior Scientist, Kunin-Lunenfeld Applied Research Unit Baycrest Centre, 3560 Bathurst Street, Toronto, Ontario, Canada M6A 2E1; telephone: 416-785-2500, ext. 2534; fax: 416-785-4230; e-mail: dstreiner@klaru-baycrest.on.ca.

Commun Oncol 2009;6:36–40 © 2009 Elsevier Inc. All rights reserved.

or it is used by a shrink to treat depression, the aims are similar. In oncology, one aim is to determine the best route to deliver the drug—orally, intramuscularly, or intravenously (or using some other orifice that is best left unmentioned in case any young children pick up this journal). In addition to the best route, this phase also gives at least an initial estimate of the maximum tolerated dose. Perhaps the most important aspect of this phase is to determine the ADME of the drug—its absorption, distribution, metabolism, and elimination—the stuff near and dear to the hearts of pharmacologists. This is usually done first for a single dose; then a rising dose; and finally for multiple, repeated doses. The pharmacokinetic and pharmacodynamic data that are gathered are useful in the planning of the later testing phases. Finally, during this phase, the researchers would also be looking for common but serious side effects. Because phase I trials are often quite small, using only 10–20 participants, uncommon adverse reactions will likely not be picked up during them.

For the vast majority of drugs, the participants in these trials are normal volunteers. (Actually, in a significant proportion of studies, they are inmates in prisons, hoping to make some money or be granted early parole. Whether these people are either normal or voluntary is a point best left for ethicists to decide.) However, because of the extreme toxicity of most drugs used in oncology, participants in cancer trials are, for the most part, patients with various types of cancer. Indeed, because most of the patients won't benefit from the new drug during this phase of testing, the ones who are enrolled usually have advanced cancer and have already been tried on a number of other treatments.

Despite their small size, phase I studies often take a long time, because patients are enrolled slowly. The first few patients (usually about 3 or 4) are started on a very low dose of the drug.

If nothing untoward happens, a few more people are placed on a higher dose, and so on. The dose escalation schedule often follows what's called a "modified Fibonacci scheme."¹ If the initial dose is not high enough (and it rarely is), the next level doubles it. If no serious adverse events develop, the third dose increases by another 67%, the fourth by 50%, and all subsequent doses by 33%. This scheme balances increasing the dosage (and hence toxicity) too rapidly against depriving patients of any benefit by not increasing rapidly enough. This process is continued until either the pharmacologic data or the patients say "Enough, already," which is usually where the toxicity is moderate but reversible.

Phase II trials

Agents that get past phase I—and, in oncology, only about 5% do—go on to the next stage, which is, not surprisingly, called phase II. The aims during this phase of testing are to (a) see whether the agent is efficacious enough that it's worth going on to a much more expensive phase III trial; (b) determine in what types of cancer the drug works best; (c) find out more about side effects and how to manage them; and (d) nail down the optimal dose. Dosing studies are sometimes referred to as phase IIA, and efficacy studies as phase IIB, although they're often combined in the same trial.

In *efficacy* studies, we stack the deck in our favor in order to maximize the chances that the agent will look good. (In a later article in this series, we'll explain in more detail the differences between this type of study and *effectiveness* trials; for now, suffice it to say that effectiveness studies are more like the real world.) How do we do this? First, we select patients who are most likely to benefit from the treatment, so we screen out those with any comorbid conditions, those whom we feel may not respond to the agent, and those whom we feel may not adhere to the treatment regimen.

We also do everything possible to maximize the effect of the drug, ranging from having a nurse hold the patient's hand when the side effects kick in, to carefully supervising how the drug is given, to calling patients to remind them of their appointments or to take the drug. Is this cheating? Not really. In an efficacy study, the aim is to see whether the agent *can* work. Later studies (the effectiveness trials) will see whether it *does* work under real-world conditions with more realistic criteria for selecting people and administration methods that more closely approximate those in routine clinical practice.²

One question that arises is whether the patients should be homogeneous or heterogeneous with respect to what type of cancer they have. The answer is unequivocally "It all depends." If an agent has been developed to target a specific type of tumor, it would make sense to enroll only patients with that specific form of cancer. In other cases, it may be known that the agent operates in a certain way, such as by restricting the blood supply, but it's unclear what particular types of cancer would respond best, so a more heterogeneous mix of patients is studied, at least initially.

Because one goal of a phase II trial is to determine efficacy, another question that arises is "Efficacious as compared to what?" For this reason, phase II studies are often randomized controlled trials (a topic for the next paper in this series), in which patients are randomly assigned to get either the new agent or treatment as usual (TAU). The goal is not to find clear, unequivocal evidence in favor of the new drug—that's left for the next stage—but simply to get enough data to satisfy the regulatory agencies that it makes sense to proceed to phase III studies.

Phase II trials fall in between phase I and III studies numerically, chronologically, and also in terms of sample size. They generally involve between 100 and 500 volunteers, and the studies last

Why we need statistics

Geoffrey R. Norman, PhD, and David L. Streiner, PhD, CPsych

Statistics is a way of dealing with chance. Anytime we observe anything in a sample of patients—mean value of a lab test, mortality rate, average survival time—the observation will differ from the long-term average derived from many studies with larger samples simply because of randomness. As a result, if we wish to compare two treatments or examine groups of patients who are and are not exposed to some risk factor, we must have some way to separate the differences between the two groups. Some differences may arise from random variation; others are beyond random variation and may reflect real effects. That's what statistics is all about—nothing more and nothing less.

As any health researcher knows, the *sine qua non* of every research paper is the statistics. It is the scary part. You understand the background very well, you can follow most of the methods, but when you get to the section labeled Analysis, it might as well be subtitled “Mere mortals may skip this step.”

Statistics, with its talk of confidence intervals, Cox models, and Kaplan-Meiers, is guaranteed to put the fear of a deity in most clinicians. But much of that is just another kind of jargon. Every field has its own specialized language, and outsiders need secret decoder rings to figure it out. You're not innocent! Remember your last NHL pt whose ANC took a dive and you had to order ESAs? We'll trade you one WBC for two ANOVAs and a GLM. And in statistics, as in medicine, once you understand the words behind the abbreviation, things aren't all that bad. In fact, as two whose toes have dipped into both ponds, our opinion is that fluids and electrolytes make even the hairiest stats seem pretty simple.

The starting point is to figure out why all the fuss. What is so essential about statistics? Simply this: clinical research almost always suffers from too few participants and too small effects. Long gone are the days when Edward Jenner could prove his point with one milkmaid, or Banting and Best could

make their case with one child suffering from type I diabetes. Back then, when everyone who had diabetes died, all you needed to do to “prove” a point and be heroic was save just one life.

Because of modern medicine and modern methods, over the typical 3-to-5-year life of a study, you can expect that even with relatively serious conditions, few participants will die. Of course, there are situations in oncology where this is not the case and mortality is high, but in the grand scheme, this is probably an accurate statement. And you can be virtually certain that whatever you do, whatever new drug is in the experimental arm, the net benefit over the conventional therapy will also be small. Consequently, you are almost inevitably looking at very small numbers to try to decide whether or not a treatment worked.

For example, suppose the overall 5-year mortality for a given condition is 10%. Further suppose that the drug is a typical good new drug and reduces mortality by 20%. If we have 100 patients in the experimental arm and 100 in the control arm, on average we'll be looking at 8 deaths in the former group and 10 in the latter.

But that's what you would see on average. The reason you really need statisticians is this: Even though we can work out that the expected number of deaths in each group would be 8 and 10, we'll get fluctuations in the observed numbers of deaths. For example, in the 100-

sample size study, we can expect (based on our arcane knowledge of statistics) to observe variations between 2 and 14 deaths in the treatment group and 4 and 16 deaths in the control group. So it would not be all that impossible to get a mortality rate of, say, 12/100 in the treatment group and 5/100 in the control group, even if the drug had a 20% risk reduction.

Be objective about chance

That's the essence of the problem. If we observed 8 deaths in the treatment group and 10 in the control group, few among us would be prepared to conclude that the evidence shows the drug worked. With the same relative risk and mortality rate and a sample size of 10,000, we would see 800 deaths versus 1,000 deaths on average—a difference of 200 deaths. Probably most of us would say this is evidence of effectiveness.

But the problem is that the dividing line between success and failure, a real difference versus a statistical fluctuation, is gray indeed. And if we left it to judgment, there is no doubt that those judgments would be seriously influenced by prior beliefs. In particular, if you're a shareholder in a drug company, you have lots of prior belief that the drug worked; if you're running a drug insurance plan, you have good reason to hope the expensive new drug didn't work. So we need

some objective way to decide whether the difference that was observed falls within the range of chance fluctuation or is unlikely to have arisen by chance. (Incidentally, it would take about 2,000 trial participants per group to achieve the magical P value of 0.05 and to be confident that the difference did not arise by chance.)

And that's what statistics does for us. That magical number " $P < 0.05$ " is a precise statement about the role of chance. When you unpack it, what it says is "The likelihood that a difference this large or larger could have arisen by chance, if there was

no treatment effect, is less than 5%." Conventionally, if it's less than 5%, we presume that the difference was a consequence of a real treatment effect.

That's all statistics does for us. It deals explicitly with the operation of chance and random error. But it does not deal with any of the external and internal biases that arise in research studies (a topic for a future column). Yes, we can sometimes try to correct for biases with "adjustment" methods, but no one would really believe a claim that a treatment worked only "after adjustment." So however essential statistics is for dealing with

the potential of random variation to mislead, that's all it's good for.

Still, that relatively minor role is good for a lot of discussion. Over the next dozen or so Practical Biostatistics columns, we'll further explore some basic ideas of quantitative analysis and the underlying logic of statistics. We'll then take up some of the more common tests, talk about what they are and how they work, when they're used, and how they're missed. After we're done, hopefully you may never fear statistics again.

Dr. Norman can be reached at norman@mcmaster.ca.

somewhere between 1 and 2 years.

Phase III trials

Again, only a small proportion of agents make it through phase II and go on to phase III studies.³ The goal at this stage is ambitious: to show that the drug (or combination of drugs) is either superior to TAU with regard to shrinking the tumor or saving lives or is equivalent to TAU in terms of effectiveness but has fewer side effects. Phase III trials are closer to the *effectiveness* end of the design spectrum—that is, does the intervention work under real-world conditions? Consequently, the patients are more similar to those encountered in daily practice. They may suffer from one or a number of other disorders, they may not be completely adherent to all instructions, they're not going to have research nurses calling them up to remind them to take their pills or to show up for their next appointment, and so on. Similarly, while the physicians in phase I and II trials may live in the cloistered environment of a university- or pharmacy-based lab, those who participate in phase III trials are more likely to be community-based oncologists. This arrangement may have implications in terms of their training, the amount of time they have to devote to each patient, and who actually sees the patients—they or medical

students and residents.

There are very few major breakthroughs in cancer research (or much of medical research, for that matter). This means that most often we are comparing a new treatment with TAU, that is, an already existing treatment that itself has proven to be at least somewhat effective. Consequently, we are looking for relatively small differences between the groups in terms of time to remission, the number of patients still alive at some given time, the severity of side effects, or whatever endpoint has been chosen. The result is that we need a large number of participants in each group in order to demonstrate a statistically significant difference, because sample size is akin to the magnification of a microscope. To see smaller things, we need more magnification; to see smaller differences, we need more subjects.⁴

There are other factors that influence the size of a trial. Outcomes that are dichotomous—alive/dead, in remission/not in remission, and so on—require much larger trials than continuous outcomes, such as size of the lesion, time to remission or death, or severity of adverse reactions.⁴ For these reasons, it's not unusual for phase III trials to have more than 1,000 participants and involve dozens or even a few hundred

sites. Moreover, new agents are rarely approved if only one phase III trial has been submitted to the FDA as supporting evidence. There are usually a minimum of two of them, varying in terms of types of patients, dosing regimens, and the like, although this requirement may be waived for compassionate reasons (compassion for the patients, not the drug companies).

Phase IV trials

Even though phase III trials may be large, they are not usually large enough to detect serious but rare side effects, especially those that may manifest themselves only after a number of months or years. The usual rule of thumb is that if an adverse event happens in 1 patient in 100, we'd have to look at 300 patients to be 95% confident that we'd detect it. So, even phase III trials with 1,000 patients can't reliably find adverse outcomes that occur in fewer than 1 out of 300 patients. Furthermore, although phase III trials may try to approximate the real world, it is still the case that patients who are enrolled in trials are not truly representative of patients in general, and the way a treatment is delivered during a study is more controlled than what occurs in actual practice, once the agent has been approved.

Commentary

How *not* to build a trial

David L. Streiner, PhD, CPsych, and Geoffrey R. Norman, PhD

Here, the authors dissect two studies—a phase II trial they consider to be flawed and a phase III trial they think was well done.

AJANI ET AL¹ carried out a phase II trial for patients with nonresectable localized esophageal cancer. In the trial, all of the patients, who were divided into two arms, were given definitive chemoradiotherapy. Group A received fluorouracil-based therapy before chemoradiotherapy; group B did not. The aim was to see whether either surpassed a goal of 77.5% 1-year survival (a previous trial reported a 66% survival rate). Eighty-four patients were randomized to the two groups, of whom 72 were assessed. At the end, the 1-year survival rate in group A was 76%; in group B, it was 69%. The conclusion was that “neither of the two treatments proved to be sufficiently superior to the historical control...to warrant further investigation.”

We think this conclusion is flawed for two reasons. First, why was the criterion set at 77.5%? Why not 75% or 80% or some other figure? Without justification, this number appears

arbitrary. Second, there were only 37 patients in group A (28 at the end of 1 year). Given this very small number, the difference between 77.5% and the observed 75.7% is trivial; repeat the study with a new group of 37 patients, and it is just as likely as not that the criterion will be surpassed. This phase II trial (which determines whether it's worth going forward with a larger trial) was interpreted as if it were a phase III trial (looking for superiority).

In contrast, Jones et al² conducted a phase III trial in which they compared the combination of cyclophosphamide plus doxorubicin (Doxil; AC) against cyclophosphamide plus docetaxel (Taxotere; TC) as adjuvant therapies for operable breast cancer. They were looking for a 10% improvement in 5-year disease-free survival in favor of TC and enrolled a total of 1,016 patients, about equally divided between the groups. The groups were well balanced at baseline; most patients completed all four

treatment cycles (93% in the TC arm, 95% in the AC arm), and all were followed until study discontinuation. At the end, there was a significantly better outcome for AC than TC.

Although one could argue that a 10% improvement is an arbitrary figure, at least in this study, the rationale was spelled out explicitly. The number makes more sense than the first study's 17.4%—the difference between the previous “gold standard” of 66% versus the goal of 77.5%, a figure that was never stated.

References

1. Ajani JA, Winter K, Komaki R, et al. Phase II randomized trial of two nonoperative regimens of induction chemotherapy followed by chemoradiation in patients with localized carcinoma of the esophagus: RTOG 0113. *J Clin Oncol* 2008;26:4551–4556.
2. Jones SE, Savin MA, Holmes FA, et al. Phase III trial comparing doxorubicin plus cyclophosphamide with docetaxel plus cyclophosphamide as adjuvant therapy for operable breast cancer. *J Clin Oncol* 2006;24:5381–5387.

Phase IV trials, which are often called “post-marketing surveillance studies,” are follow-ups of a large number of people who have been or are on the agent. They are naturalistic, in that nothing new is done to the patients. The data on adverse events—death, a life-threatening condition, hospitalization, disability, a congenital anomaly, or an intervention necessitated by one of these events—can come from a number of sources. First, the manufacturers are required to notify the FDA of any serious, adverse events that they discover. Second, front-line clinicians can file online reports with the FDA

under a program called MEDWatch at www.fda.gov/medwatch/. Finally, the manufacturers can be directed to fund a large epidemiologic survey of patients who received the drug.

In our next column, we'll talk about randomized studies and why they are considered the gold standard for evaluating treatments.

References

1. Crowley J. *Handbook of Statistics in Clinical Oncology*. New York: Marcel Dekker; 2001.
2. Streiner DL. The two Es of research: efficacy and effectiveness trials. *Can J Psychiatry* 2002;47:347–351.

3. Reichert JM, Wenger JB. Development trends for new cancer therapeutics and vaccines. *Drug Discov Today* 2008;13:30–37.

4. Norman GR, Streiner DL. *Biostatistics: the Bare Essentials*; 3rd ed. Toronto: B. C. Decker; 2007.

ABOUT THE AUTHORS

Affiliations: Dr. Streiner is Professor, Department of Psychiatry, University of Toronto, Senior Scientist, Kunin-Lunenfeld Applied Research Unit Baycrest Centre, Toronto, Ontario, Canada. Dr. Norman is Canada Research Chair in Cognitive Dimensions of Clinical Expertise; Assistant Dean, Programme for Educational Research and Development; Professor, Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada.

Conflicts of interest: None to disclose.