When is an answer not an answer?

David L. Streiner, PhD, CPsych,[1,2] and Geoffrey R. Norman, PhD[2]

[1]Department of Psychiatry, University of Toronto, Ontario, Canada, and [2]Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada

When your beloved authors were studying research and statistics, around the time that Methuselah was celebrating his first birthday, we thought we knew the difference between hypothesis testing and hypothesis generating. With the former, you begin with a question, design a study to answer it, carry it out, and then do some statistical mumbo-jumbo on the data to determine if you have reasonable evidence to answer the question. With the latter, usually done after you've answered the main questions, you don't have any preconceived idea of what's going on, so you analyze anything that moves. We know that's not really kosher, because the probability of finding something just by chance (a Type I error) increases astronomically as you do more tests.[1] So, in the hypothesis generating phase, you don't come to any conclusions; you just say, "That's an interesting finding. Now we'll have to do a real study to see if our observation holds up."

Well, we *thought* we knew the difference, but something must have changed over the past few centuries when we weren't paying too much attention. The reason for our puzzlement is an article by Hurvitz et al[2] about the relative effectiveness of trastuzumab emtansine (T-DM1) compared with trastuzumab plus docetaxel (HT) in patients with metastatic breast cancer. First, a bit about the study itself. This was a phase 2, multicenter open label randomized controlled trial. "Phase 2" means it's not yet ready for prime time,[3] and "open label" means that nobody was blinded regarding who got what. ("Multicenter" means a great opportunity for the investigators to rack up frequent flier points.) There were 137 women with HER2-positive metastatic breast cancer or recurrent locally advanced breast cancer, randomly divided between the 2 groups. The primary endpoints were progression-free survival (PFS) and safety, both assessed by the investigators. Key secondary endpoints were overall survival (OS), objective response rate (ORR), quality of life (QOL), and a handful of others. What they found was that the median PFS was 9.2 months with HT, compared with 14.2 months for T-DM1; and an ORR of 58.0% in the HT group and 64.4% in the T-DM1 group; both were statistically significant. However, "preliminary OS results were similar between treatment arms."

So, let's begin looking at the study. It may help if you jotted down all of the abbreviations that were used; we listed 15 before we ran out of lead for our pencils, and we never even got to the ones from statistics we were familiar with. We can't fault the authors for this; it appears to be an editorial policy to abbreviate everything and not provide a table of them to help readers. Ink must be a very precious commodity. But back to the study.

The paper states that "This study had a hypothesis-generating statistical design." If you go back over all of the papers we have written for this journal, looking for a definition of "hypothesis-generating statistical design," you will look in vain. If you think that we have been remiss in not discussing all research designs (actually, we have been, and haven't mentioned many of them) and check textbooks of research design, your search will again prove to be fruitless. In fact, we had to resort to that salvation of all serious academic researchers, Google. What we found, among all the hundreds of millions of Web pages, was only one mention of the term – in the article we are reviewing! So what does the term mean? Given our vast knowledge of statistics and research design, we feel safe in saying, "We don't have the foggiest idea." There are indeed many research designs, and we should know; we've written books about them (OK, so maybe only one book[4]). Different designs depend on how the subjects were located, how (and if) they were followed up, whether or not the researchers had any control over who got what, and a host of other factors, but not whether the study was meant to test hypotheses or to generate them – that depends solely on whether the analyses were specified beforehand or not.[5]

Commun Oncol 2013;10:189-190
DOI: 10.12788/j.cmonc.0037

Actually, we ran across a similar issue in a previous critique we did for this journal.[6] There the authors claimed that they did a "noncomparative" study and wouldn't do any tests of statistical significance. However, they then reported the results for their two groups along with confidence intervals, so that anyone with a modicum of statistical knowledge could do the math. We wonder if this is the same bait-and-switch tactic; saying on the one hand, "Don't take these results too seriously, because we're only generating hypotheses," and on the other hand, "Here are our conclusions and we've done statistical tests, so take them seriously (and approve our drug so we can market it)."

Perhaps the authors meant that this is only a pilot study; it was, after all, a phase 2 study and not a phase 3 one. But this would be equally puzzling. Pilot studies are designed to determine if there's anything worth following up and if a grownup study is feasible. As such, they aren't designed to have enough subjects to demonstrate statistical significance. But this article found significance for the primary outcomes, so a larger study isn't needed for those. Perhaps a clue comes from their abstract, where they state that "preliminary" results for OS were similar between treatment arms (the hazard ratio [HR] was 1.06 and not significant, where 1.0 means no difference). Does that mean they should do a larger study so that the HR will become statistically significant? That's not a statistical question, but rather a clinical one. Increasing the sample size enormously may result in statistical significance for a HR of 1.06; the question is, even if it becomes significant, would you change your practice and switch to T-DM1 for a HR that small? Neither would we, and we're not even oncologists.

This study is also similar to the previous one we mentioned[6] in another regard; the choice of the outcomes that the authors chose to report. Most of the significant differences here – PFS, ORR, and safety – were judged by the researchers. Don't forget that this was an open-label study, so the research team knew who was getting what. Most phase 3 trials are double-blind, meaning that nobody knows who got what except for the statistician or pharmacist (and they'd best not lose the list if they value their valuables). There are reasons for this, and they all have to do with potential biases that can creep – or stampede – in: patients may report improvement if they believe that they are getting a newer, better, more expensive intervention (the placebo effect); and clinicians may "see" improvement if they have an investment – emotional as well as financial – in the study (expectation bias). It is precisely for these reasons that the patients, clinicians, and raters are kept blind whenever possible. There was one outcome that could not be biased in this way – overall survival. And what did they find here? The statistical/Yiddish term is "bupkis," which means "absolutely nothing." (For those with an interest in etymology, the full term is "kozabupkis," which means "goat droppings," but we didn't want to be crude). Progression-free survival may look objective too, but then someone has to judge progression. This is similar to the previous article which also had an impressive outcome with respect to PFS but no difference with OS. The two saving graces are first, that there were only half as many serious adverse events (AEs) with T-DM1; and second, that there was a delayed decline in QOL by 5 months for those in the T-DM1 group; that is, it took a bit longer to feel terrible, but judging when something starts to decline is also subjective and very tricky.

So what's our bottom line? First, we don't begin to understand what this study was about; was it hypothesis generating (which is what they said) or hypothesis testing (which is what they did)? Second, from the patients' point of view, T-DM1 *may* have some promise in terms of AEs and QOL but does SFA (that's one of the few abbreviations you won't find in the paper) for life expectancy. Now it's up to the clinicians to make a decision: even if we believe it (and this is a time for agnosticism), how much is that delay in the decline of QOL worth, compared with the increased cost of the new treatment? We may say that no cost is too great to improve the lives of those with a terminal illness, but bear in mind what economists call "opportunity costs" – given finite resources, money spent on one thing means that less is available to be spent on something else. So what are you willing to give up to achieve 5 months of better QOL for these patients – the ability to buy another MRI machine, or to open another bed on the oncology unit, or being able to do more operations? Kinda sorta makes you think twice before you rush out to buy some of that T-DM1. In our opinion, we yet again have to fall back on that delightful Scottish legal equivocation: "Not proven."

References

1. Norman GR, Streiner DL. P less than 0.05: Statistical inference. *Community Oncol*. 2009;6:284-286.
2. Hurvitz SA, Dirix L, Kocsis J, et al. Phase II randomized study of trastuzumab emtansine versus trastuzumab plus docetaxel in patients with human epidermal growth factor receptor 2-positive metastatic breast cancer. *J Clin Oncol*. 2013;31:1157-1163.
3. Streiner DL, Norman GR. Drug trial phases. *Community Oncol*. 2009;6:36-40.
4. Streiner DL, Norman GR. *PDQ Epidemiology*. 3rd ed. Shelton, CT: PMPH USA; 2009.
5. Streiner DL, Norman GR. Correction for multiple testing: is there a resolution? *Chest*. 2011;140:16-18.
6. Streiner DL, Norman GR. Counting what really counts: the irinotecan in recurrent glioblastoma trial. *Community Oncol*. 2011;8:425-426.