

## What Is Wrong with What Is Said?

Hasan Yazici, M.D.

### Abstract

*The scientific flaws in medical research, especially as they relate to randomized controlled trials are rather uniform and the common denominator of such flaws is the attempt to prove, rather than falsify, the study hypothesis. This interactive educational exercise is an attempt to highlight these flaws in short scenarios.*

The randomized controlled trial is the backbone of evidence-based medicine as it relates to patient management. The aim of this interactive exercise is to highlight and briefly discuss some errors, most of them quite common, in conducting, reporting, and peer reviewing randomized controlled trials (RCT) for drug research. A portion of the issues brought forward specifically pertains to RCTs, while others relate to any scientific trial design and communication.

For each point discussed, a brief scenario is presented followed by a discussion of what is wrong within the scenarios. Finally, please note that the scenarios 1A to F concern a single, hypothetical RCT of the use of vitamin E in the management of gout.

### Scenario 1A

Several recent reports suggest that vitamin E can lower serum uric acid levels.<sup>1-3</sup> These are either single case reports or uncontrolled, open studies with small groups of patients. Therefore, we tested the hypothesis of whether vitamin E was an effective treatment for gout in a 6-month double-blind, placebo-controlled study.

*What is wrong with what is said?*

Hasan Yazici, M.D., is from the Division of Rheumatology, Department of Medicine, University of Istanbul, Turkey.

Correspondence: Hasan Yazici, M.D., Sefa Sokak, Sen Apt 17/7, Kadikoy, Istanbul, Turkey 81310; hyazici@attglobal.net.

### Response

In any scientific experiment, the hypothesis is better expressed as an affirmative statement rather than as a question. Since the aim of any scientific experiment is falsification,<sup>1</sup> as we will reiterate several times during this discussion, an affirmative statement from the start clearly defines what needs to be falsified.

### Scenario 1B

Initially 60 patients each had been allocated to the active drug and the placebo groups. During the course of the study 30 patients had to leave the study for various reasons. Twenty patients from the active drug group and 10 patients from the placebo group left the study before its completion (not significant). Thus, the final analysis for drug efficiency was made based on 90 patients.

*What is wrong with what is said?*

### Response

In analyzing the results, the investigators do not take into account the patients who left the trial. The so-called “intention-to-treat” principle<sup>2</sup> compels us to account for the fate of each randomized patient when making our efficacy analyses. Furthermore, replacing a dropout with another eligible patient diminishes the effect of the initial randomization, and, if the reason for dropout had been harm, this replacement process has the potential of erroneously inflating the denominator when assessing for harm at the end of the trial.

### Scenario 1C

The allocation of 60 patients to the active drug and the placebo arms ensured an 86% power to detect a difference rate of 25% between the two groups in the number of patients whose uric acid levels were lowered by 1 mg/dL or more, using a two-sided Fisher’s exact test with  $\alpha$  set at 0.05.

*What is wrong with what is said?*

**Response**

In a power calculation,<sup>2</sup> it is not enough to state the magnitude of change. The baseline value should also be given. A 25% change in the number of respondents could be the difference between 40% and 30% or the difference between 20% and 15%, two sets of values having different degrees of statistical power.

**Scenario 1D**

The uric acid levels decreased to  $5.2 \pm 1.5$  mg/dL from a baseline of  $9.3 \pm 2.3$  mg/dL in the treatment group and from  $9.7 \pm 1.9$  mg/dL in the placebo group. There were no differences in the lowering of uric acid between the two groups ( $p > 0.05$ ). There were also no differences in the number of gouty attacks between the two groups of patients. However, a subgroup analysis was also done. Among those patients in the treatment group who gave a history of more than five gouty attacks per year, as compared to those who had less than three, the uric acid levels were significantly lower after treatment ( $p < 0.05$ ).

*The peer reviewer said:* The findings in the subgroup analysis, conducted among a small number of patients, should be interpreted with caution. Even though the investigators found a statistically significant difference in efficacy, the study was not primarily planned to assess this difference. This lessens the external validity of the results.

*What is wrong with what the reviewer said?*

**Response**

The main problem with this subgroup analysis is not that the subgroup is small. The difficulty lies with the effort made in the direction of proving that the new drug at hand is efficacious and basically against the falsification process.

**Scenario 1E**

Table 3 gives the rate of observed adverse events. Indigestion was rather frequent in both groups: 15/45 patients in the treatment group and 14/45 in the placebo group. A subgroup analysis was also done after the patients who participated in the trial were questioned about a pretrial history of indigestion. It turned out that 20/45 patients in the treatment and 18/45 in the placebo group had pretrial indigestion. A further analysis among these patients revealed that 14/20 in the treatment group and 2/18 in the placebo group ( $p < 0.03$ ) with a history of pretrial indigestion also reported indigestion during the trial.

*The peer reviewer said:* The findings in the subgroup analysis, conducted among a small number of patients, should be interpreted with caution. Even though the investigators found a statistically significant difference in harm, the study was not primarily planned to assess this difference. This lessens the external validity of these results.

*What is wrong with what the reviewer said?*

**Response**

Again, the issue is not the size of the subgroup. This subgroup analysis looks at the possibility that this new drug

might not be all that harmless. Thus, the exercise is in the direction of falsification and, therefore, it is justified.

**Scenario 1F**

In conclusion, vitamin E is an innocuous agent, with the possible additional benefit of an antihypertensive effect, as was also serendipitously noted in this study. Such studies are surely warranted. We are about to start an investigative study along these lines at our institution.

*What is wrong with what is said?*

**Response**

The declaration that more studies are being planned is potentially preempting. It should be avoided in that it might influence other investigators away from this line of research. The take home message of the whole exercise is that we (or the reviewers!) are usually *wrong* when we try to prove ourselves. The aim is or should be *falsification* of the hypothesis.<sup>1</sup>

**Scenario 2**

Ninety patients were initially randomized to 30 patients each of placebo, QXY 2 mg qd, and QXY 5 mg qd groups. One patient in the QXY 5 mg group developed pneumonia and had to be hospitalized 10 days after treatment started. The patient was withdrawn and, according to the protocol, another patient was recruited. Thus, the intention-to-treat analysis brought the number of patients analyzed to 31 in the QXY 5 mg group and the total number of patients to 91.

*What is wrong with what is said?*

**Response**

The replacement of a patient who leaves the trial after randomization is problematic because he bypasses the initial randomization process. A further issue is that he will inflate the denominator when looking at harm at the end. Thus in the present case it will distribute 1 case of pneumonia (a serious adverse event) among 31 instead of 30 patients.

**Scenario 3**

The primary end point was the achievement of at least 25% improvement according to the ACR criteria (ACR20) at week 20. The sample size of 180 patients per group was chosen to ensure an adequate safety evaluation. The sample size also ensured that there was 90% power to detect a significant difference in the proportion of ACR20 responders between the treatment groups using a significance level of 0.05, assuming 20% and 42% of the patients in the control and the treatment groups, respectively achieved ACR20 responses.

*What is wrong with what is said?*

**Response**

No rationale is given for baseline assumptions related to harm, and the power calculations are post hoc (after the study was completed), which then have limited meaning.

**Scenario 4**

A sample size of 300 patients each in the study drug and

the control groups was determined to demonstrate a specific adverse event rate of 1% or less, with 95% confidence.

*What is wrong with what is said?*

#### Response

If one screens “N” individuals and does not find an attribute “Y” among this group, one can conclude that the frequency of “Y” is less than  $1/0.33n$  among the “N” individuals surveyed, with 95% confidence. This approximation is true for prevalences less than 0.02 for the attribute surveyed.<sup>3</sup> On the other hand, the probability of an event *not happening* does not give us information, as is claimed in the case at hand about the likelihood of events *happening* in the different arms of the study.

### Scenario 5

In a placebo-controlled withdrawal study, an investigator showed that small doses of prednisone (1 to 4 mg/day) were significantly effective in the management of rheumatoid arthritis. The study was conducted among 31 patients.

*The peer reviewer said:* This beneficial effect of prednisone described among a small number of patients should be interpreted with extreme caution, even if the investigators found a statistically significant difference. The number of patients studied, and thus the study power, was simply too small.

*What is wrong with what the reviewer said?*

#### Response

The reviewer is partially right. The issue of power applies only if we are missing (type II error) an effect that would have been more evident in a larger group.

As for the external validity (the generalizability) of the results, findings among a small number of patients might be quite different from the larger population of rheumatoid arthritis patients. On the other hand (!), this is not simply an issue of numbers but of patient selection.

### Scenario 6

Six thousand patients with osteoarthritis of the knee were randomized to receive either the new coxib (NCB) or the traditional coxib (TCB). Forty percent of the NCB and 35% of the TCB patients found total pain relief. Hypertension was a problem in 2.5% of the NCB and 4.5% of the TCB patients. It was concluded that NCB decreased pain by 13% and there was 45% less hypertension with NCB.

*What is wrong with what is said?*

#### Response

Nothing is wrong; however, the findings need more interpretation. An NNT (number needed to treat) analysis<sup>2</sup> will tell you that you have to treat 10 patients to see the superiority of NCB over TCB in one patient. A NNH (number needed to harm) analysis will say that you have to treat 50 patients with TCB to harm one more patient with hypertension as compared to using NCB.

### Scenario 7

Two previous double-blind studies of colchicine in BS (Behcet’s syndrome) had shown no superiority of this agent over placebo in treating the oral ulcers in this condition. Recently, a withdrawal study was done among those patients who claimed benefit from colchicine. They were randomized to continue to receive the active drug or placebo. After 3 months, those who stopped taking colchicine had significantly more ulcers.

*The peer reviewer said:* 1. This study is interesting but of limited use. A problem with all withdrawal trials is that they seldom represent real life use. 2. The investigators used the “sign test” to analyze the differences in oral ulcers between the two groups. Is this a new test? I suspect it is not very powerful. Why not use the more powerful tests of significance to show the real differences?

*What is wrong with what the reviewer said?*

#### Response

The main problem with withdrawal studies is that they are not done often enough. They provide excellent information about possible type II errors in previous traditional RCTs. An important issue is that they do not provide a fair picture of drug-associated harm in that the trial starts among a group of drug responders who have been taking the drug, presumably with no important side effects. Moreover, the sign test is a time-honored, conservative tool. Finally, significance observed in a conservative test gives more validity to the differences observed.

### Scenario 8

Among the 96 patients allocated to the new medication there were three cases of myocardial infarction, while the same was true for 2/94 patients allocated to placebo ( $p = 0.86$ ).

*What is wrong with what is said?*

#### Response

This is an abuse of the “p” value. It is obvious that there were no significant differences in the rate of myocardial infarction between the two groups. In order to offer a formal statistic, here is an unfortunate attempt to give an aura of science.

#### Disclosure Statement

The authors has no financial or proprietary interest in the subject matter or materials discussed, including, but not limited to, employment, consultancies, stock ownership, honoraria, and paid expert testimony.

### References

1. Yazici H. Use and abuse of the controlled clinical trial. *Bull NYU Hosp Jt Dis.* 2007;65:132-4.
2. Schulz KF, Grimes DA. Sample size calculations in randomized trials: mandatory and mystical. *Lancet.* 2005;365:1348-53.
3. Yazici H, et al. The “zero patient” design to compare the prevalences of rare diseases. *Rheumatology (Oxford).* 2001;40:121-2.