# A Taxonomy of Estimands for Regulatory Clinical Trials with Discontinuations

Thomas Permutt*

**Abstract**

When patients in trials discontinue treatment because of toxicity, lack of efficacy, or death, it is not always clear how to define the effect of treatment. We might follow them, if they are alive, and record the outcome notwithstanding discontinuation of treatment. We might (and must, for death) consider the discontinuation itself to be the outcome. We might (rarely) consider the values before discontinuation to be the outcome. We might try to estimate the effect in a subset of patients who complete the course of treatment, though this requires careful definition.

I shall talk about how to define and estimate these effects. I shall also consider and dismiss certain other definitions of effect, notably the effect if the patients who discontinued had not done so.

**Key Words:** missing data, clinical trials, causal inference, direct effects

## 1. Two and a Half Problems, With Answers

Consider a clinical trial in a very bad infectious disease. The treatment is given as a single shot. The outcome of primary interest is whether the patients are alive or dead a month later. You lose track of some patients and don't know whether they are alive or dead. You could just look at the ones you did manage to follow, but they might be systematically different from the ones you lost. For example, if part of your follow-up involves phoning patients, dead people are more likely not to answer the phone. Even this might not matter much, unless the test and control groups are differently different: that is, unless the difference between followed and unfollowed patients is different for the test and control.

This is a statistical problem, and statisticians love to talk about it. We write papers about methods for dealing with it mathematically, but all those papers say you should try very hard to follow the patients, and only use our wonderfully elegant methods in cases where your very hard efforts fail. A few years ago, the Food and Drug Administration contracted with the National Research Council (NRC) to convene a panel of expert statisticians to advise us, and they wrote us such a paper, filled with elegant mathematical methods but titled, *The Prevention and Treatment of Missing Data,* [1] prevention first.

Now consider a trial in diabetes. The outcome of primary interest is the blood sugar, measured as hemoglobin A1c, at the end of six months of treatment. Some of the patients die during the six months, so of course they don't have HbA1c measured at six months.

These two problems have very little in common. The first one has a binary outcome, the second one has a continuous outcome, but that's not important here. In the first case you failed to observe something you might possibly have been able to observe; in the second case you observed everything you possibly could have. In the first case the thing you didn't observe nevertheless exists: each patient is in fact alive or dead, whether you know it or not. In the second case, the "missing" value of HbA1c does not exist. If we knew in the first case whether the missing patients were alive or dead, a straightforward analysis would be meaningful: count how many are alive; in the second case there is no meaningful, straightforward analysis. In the first case, failures to ascertain vital status, if there are a lot,

*U.S. Food & Drug Administration, 10903 New Hampshire Ave., Silver Spring, MD 20993

might be seen as an indication of faulty design or conduct of the trial. In the second case, I'm sure you did everything you could to keep the patients alive: there is no implication of bad design or conduct. In the first case, if you use other data to guess whether each patient is alive or dead, you will guess right or you will guess wrong. In the second case, if you estimate the HbA1c for dead patients, you won't be right or wrong; you'll be talking about some sort of alternative reality.

About all the two problems do have in common is the structure of the database. There's a place to enter vital status or HbA1c, and there's no value for vital status or HbA1c, and you're going to enter a missing value code, even though the interpretation of that code is completely different in the two cases. So you can turn the same missing-data crank and get an answer. The methods assume you have some problem like the first one, with a meaningful but unobserved outcome. If you have the second problem, where the outcome does not exist, the methods produce nothing of any interest.

NRC wrote us a book about the first case, filled with hard statistics but still called *Prevention and Treatment*. The second case, they opined, was not really a missing data problem at all, and they just left it at that. This is unfortunate. I read it as scholarly diffidence, saying, We know how to do one of these problems, so that's what well write about; we don't presume to advise on the other. Others read it differently: Here are some of the world's greatest experts on missing data, here's their book, it only has methods for problem 1, they can't possibly have meant to leave problem 2 out, so they must want us to use these methods anyway for problem 2.

Now let me turn to a third problem, which I'll call problem one and a half. You have an inhaled treatment for cystic fibrosis (CF). The outcomes of primary interest would be progression and survival, but pulmonary function is a surrogate, and you're going to measure it at the end of six months. The treatment is unpleasant: it makes patients cough a lot. Coughing in CF patients can be good up to a point—it helps clear the airways of the secretions that partially block them—but some patients cough so violently that they are unable or unwilling to continue.

Is this problem 1.5 like either problem 1 or problem 2? NRC suggests that many such problems can and should be handled like problem 1. Whether patients continued or discontinued treatment, they still have lung function after six months, and it could be measured. Furthermore, NRC suggests, this measurement is meaningful to address what may be the most important medical question: is this treatment helping patients, on average? That is, are patients treated with the test article better or worse off than those treated with control, regardless of whether they adhered perfectly, imperfectly, or not at all to the assigned treatment? Many people would use the term *intent to treat* in this connection, but other people use this term in different ways.

I think this intent-to-treat approach is reasonable, but I also think it's reasonable to handle this problem 1.5 like problem 2. Patients who discontinue are not dead, but you might think, even without measuring their pulmonary function, you already know what you need to know about them. Their pulmonary function at six months is unlikely to be much affected by the treatment they stopped taking, and that was a surrogate anyway; their long-term progression and survival are not going to be affected at all. So the outcome that matters for these patients is the very fact that they discontinued the treatment, just as in the diabetes trial, the outcome was the fact that the patients died. There's nothing missing that should have been measured. The fact that some patients were unable to tolerate the drug does not indicate any defect in design or conduct of the trial.

So, I think you can get reasonable answers for this problem along the lines of problem 1 or along the lines of problem 2; but the statistical methods are completely different. For problem 1, you get the pulmonary function and use it. You use it in exactly the same way

for patients who are still on treatment and patients who are off treatment, because this gives you a perfectly valid answer to a meaningful question: Ignoring adherence, is it better to be assigned test drug or control? You try to measure pulmonary function in all patients, but you use methods such as NRC recommends for the few patients whose pulmonary function you are somehow unable to collect. For problem 2, you handle discontinuation itself as the outcome when patients discontinue, but pulmonary function as the outcome when they don't, so you have to build some kind of composite outcome to do statistics on.

## 2. Three Other Answers

Problem 1.5 (or 1.4 or 1.7) comes up very often. Patients discontinue treatments because of adverse events or lack of efficacy or some of each. Usually one or the other way of handling it will work: either collect the outcome notwithstanding discontinuation of treatment, or consider discontinuation to be the outcome in itself. Sometimes either method will be all right.

I need to discuss three other ways of looking at problem 1.5. The methods are distinguished by their approaches to observation after discontinuation of treatment. Method 1 was, values after discontinuation matter, so we should get them. Method 2 was, values after discontinuation don't matter because discontinuation itself is what matters. Now, if you look at many of the protocols I see, if you take them literally, they say something like this: I care about values after discontinuation, but I don't have to get them, because I have plenty of values already from people who complete the course of treatment. They say "intent to treat," and then they say "missing at random," and then they blithely pretend that the dropouts were just like the completers. (Sometimes they make a big fuss about the difference between Missing Completely At Random and Missing At Random, and then they still blithely pretend that the dropouts were just like the completers except for the effect of one covariate, not carefully chosen.) And, of course, I say, if the dropouts were just like the completers, then why did they drop out? And if your drug works at all, why would people who stop taking it do just as well as people who keep taking it?

And if you ask a statistician these questions, she will readily agree that of course the actual values after dropout are systematically different from the values for completers. She may then say, It's not the actual value that could have been retrieved that I care about. It's the value that would have been observed if the patient had adhered perfectly to the treatment regimen, because that's the effect of the drug. Now, remember, I have these patients with an invariably fatal illness, and they know coughing is good for them, and they simply can't stand the way this drug makes them cough. Do you want to estimate what would have happened to them if they could stand it, and call that the effect of your drug? I don't think you really do want that, but I'm quite sure that you can't meet the statutory standard for drug approval in the United States that way, which is "evidence that the drug will have the effect it purports or is represented to have": [2] will have, not could have in some counterfactual world. I also find this to be rather a bizarre distortion of the intent-to-treat principle. Yes, we like randomization-based analysis, so we like having a value for all patients randomized, but not at the expense of having to use a value that is not ascertained, does not exist, and could not exist. Some people cite the NRC report as justification for this approach, but NRC [1] explicitly says it cant be done (p. 25).

But what if you say, Look, I never said this treatment would work for everybody. Some people can tolerate it, though. Wouldn't you like to know whether, in those patients who tolerate it, it improves pulmonary function?

Yes, I would very much like to know that. The problem is whom to compare them to. You need to synthesize, from the control group, a set of patients who are like the patients

who tolerated the active treatment. This is not intent to treat. It's not pure randomization-based analysis. It has a flavor of observational epidemiology. In particular, it requires serious modeling of either the response or the propensity to drop out or both as a function of covariates and intermediate observations. Unlike randomization-based analysis, you can't just prespecify simple, doubly-robust models that work because of randomization even when they do not correctly model the outcome.

This is not the same thing, conceptually nor mathematically, as the other two non-intent-to-treat ideas. It makes sense, it answers an important question, but it's difficult, especially in a regulatory environment where we are used to prespecified, simple models.

## 3. Summary

I've posed two very distinct problems with very distinct answers: first, failure to ascertain a meaningful value; second, nonexistence of a meaningful value. I've also suggested that the third problem, discontinuation of treatment, can often be handled as one or the other of the first two, and sometimes either way will work.

I've also given two and a half answers. Treat problem 1 as problem 1: get the values and use them. Treat problem 2 as problem 2: if discontinuation makes the outcome irrelevant, then discontinuation itself is the relevant outcome. Treat problem 1.5, one way or the other as appropriate, or treat it as a third kind of problem. I've discussed three other ways of thinking about the third problem. Two of them are not useful in the regulatory environment. The third is potentially useful but hard.

You can, you should, ask yourself these questions, in the plainest possible language, before you start talking about statistical methods. Is there something actually missing? Like the vital status in the first problem? Why is it missing? Could you have found it out if you tried harder? In some cases, at least? And the other cases, where you failed to find it out, how are they different? They have to be different somehow, or you would have found out; so what systematic differences would you expect between missing and nonmissing cases?

Or do you have problem 2? Did we actually observe what we needed to observe, namely the fact of death or discontinuation? If so, the problem is not "missingness" in the sense that NRC meant, nor in the sense that pretty much any statistics book means, so that much of what you read there will be irrelevant. The problem is dealing with an outcome that is sometimes a number like HbA1c and sometimes the fact that there is no number. How do we compute with such outcomes in a way that represents benefit to the patient? Having asked and answered these questions, you should tell me your answers. NRC has some language about specifying causal estimands. It's rather obscure. I think they're talking about these questions. If you don't answer them, you risk, at best, confusion about what you've done, and at worst, doing something that makes no sense. If you do answer them, we can work together to make meaningful, defensible statements about the actual effects of your drugs.

## References

[1] National Research Council. *The prevention and treatment of missing data in clinical trials*. National Academies Press: Washington, 2010.

[2] Federal food, drug and cosmetic act. Section 505(d).