Panel Discussion

CLINICAL TRIALS

Clinical Trials
1–13

DOI: 10.1177/1740774519853573
journals.sagepub.com/home/ctj
SAGE

University of Pennsylvania 11th annual conference on statistical issues in clinical trials: Estimands, missing data and sensitivity analysis (morning panel session)

Frank W Rockhold[1], Anne Lindblad[2], Jay P Siegel[3] and Geert Molenberghs[4]

Susan Ellenberg: In 1998, the International Conference on Harmonization (ICH) issued a Guidance document entitled Statistical Issues in Clinical Trials (the document is also known as E9 as it was the ninth document issued under the Efficacy umbrella).[1] The ICH is a collaboration of scientists from regulatory authorities and industry in multiple regions of the world; the purpose of ICH is to standardize to the extent possible the approaches to regulated drug development programs in order to increase efficiency.

After almost 20 years, it seemed it's time to rethink some aspects of this document. In particular, concerns emerged about the need for better clarity in defining what a trial is trying to estimate. The problem arises because of the unavoidable problem of missing data. How to handle missing data in the analysis of clinical trial results has been a huge issue not just for the Food and Drug Administration (FDA) and industry but also for everyone conducting, analyzing and reporting clinical trials. The National Academies of Science issued an important report on this issue in 2010,[2] and this motivated thinking about the need for an addendum to the ICH E9 Guidance. The process was initiated in 2014, and the final revised Guidance document (E9 (R1)) was issued in 2017.[3] This morning's discussion panel includes two of the people who were involved in writing this addendum. Other speakers and panelists, who have spent a lot of time thinking about methods for missing data, will comment on the document and present some other ideas about handling missing data.

Frank W Rockhold[4]: In 2008, the 10-year anniversary of ICH E9 Guidance, Dr Stephen Ruberg and I wrote a review covering both how it was working and what we saw as some of the gaps.[4] We surveyed colleagues in industry and academia on such issues as how to handle missing observations and/or follow-up data, and defining the estimate of interest. The whole concept of intention-to-treat (ITT) seemed to need further discussion. (By the way, in the ICH E9 Guidance you'll see a term "full analysis set," and that came from the fact that "intention-to-treat" didn't translate well into Japanese, and the Japanese representatives on the committee wanted us to come up with a different term. So, "full analysis set" was a way of saying include as many patients as possible in the analysis.) Terms like "modified intention-to-treat" also arose. It is clear that ITT, while it was developed originally in the context of mortality trials, didn't always apply in obvious ways in other contexts.

One advantage of being around awhile is that you see things change over time. Thus, when people started using newer methodological approaches like general estimating equations and repeated measures analysis, it was pointed out that they had in fact just changed the hypothesis (and estimand). You're now asking a different question. You're not answering the precise question that was in the trial design. So, I think these discussions and debates around the estimand of interest have been building for several years. An additional perspective comes from the opportunity I had, in the last 9 years of my industry career, to be the chief safety officer for a large pharmaceutical company, so my view of data is in a broader context. I try and understand how to take E9 (R1) and make it accessible in the real world.

[1]Duke University
[2]Emmes
[3]Janssen (ret)
[4]Universiteit Hasselt and Ku Leuven

Focus of R1

This new addendum to E9 is really focused on *efficacy*. It's how to determine whether or not an intervention is performing as hypothesized and how to get the best estimate of the effect defined by the hypothesis of interest. If I were performing a phase II trial, I might consider a number of options suggested in R1 because I'd want to understand how the drug is working and how I can best plan phase III. In a confirmatory setting, some of these estimand options may be more useful than others. One thing that the Guidance doesn't do (and was not intended to do) is consider safety and benefit-to-risk assessment. It's focused on estimating *efficacy*, but in determining whether or not a drug is useful, effectiveness is an important consideration. Parenthetically, when I first started in clinical trials many many years ago, I learned that the most useful table in a clinical trial report was the Consolidated Standards of Reporting Trials (CONSORT)[5]-like diagram (CONSORT didn't exist yet) that showed patterns of withdrawals and reasons for withdrawal patterns. Those data could do a pretty good job of predicting whether the treatment was effective or not.

The R1 Guidance also briefly mentions prevention, but I would posit that prevention trials (including vaccine trials) may be viewed a little bit differently from treatment trials. In my view, the ITT estimand is the only sensible one in an event prevention trial.

The estimand

In choosing an estimand, one needs to think about real-world extrapolation of the results. A few interesting pieces of information provide some perspective on how we look at treatment estimates. First, it is well known that about a third of prescriptions written are never filled.[6] So if you're thinking about treatment policy, you need to ask how the chosen estimand will actually reflect what's going on in the real world. For many treatments, ITT may make the most sense. Second, if you're studying a regimen of baby aspirin every day, then treatment policy matters because you're trying to impact the population. Patients and physicians in the real world often don't follow dosing instructions based on clinical trials, so regardless of what you tell them—how many weeks you have to take a drug before the dose can be adjusted, or what its effect is, or who should take the drug—the formal instructions are not necessarily followed. Third, the New York Times had a recent article[7] on the enrollment criteria that determine who participates in clinical trials. Patient advocacy groups are concerned about how entry into clinical trials is restricted, because the ultimate inference from the trial population doesn't necessarily apply to the target population.

The reason for highlighting these pieces of information is that to come up with a really comprehensive definition of a relevant estimate for treatment *efficacy* (effectiveness) in the population, the robustness of the estimate is important and one can learn as much from the sensitivity analyses as the primary analysis in that regard.

Anne Lindblad: When I was asked to be a panelist for this conference, the first thing I did was investigate the origin of the word "estimand." Wiktionary says that it comes from a Latin word meaning I estimate, I value, and claimed that it was first mentioned in a 1939 JASA publication. In 1968, Tukey and Mosteller[8] used it in a paper probably in a manner that's most similar to the way we're talking about it today. I believe the first step to implementation is establishing a common understanding of the purpose and intent of what an estimand is and what impact it may or may not have on our inference. My next thought when reading the E9 Guidance was whether or not an estimand is a new concept. My conclusion? Yes and no. I say yes because I think (and I haven't done this) if I asked a very large group of clinical trialists to tell me what an estimand is and why it's important, I think I'd get, "I don't know" from the majority. This is why this guidance is important. It will begin the conversation that must occur in order to encourage incorporation into our language and increase our understanding of how adoption might help us in terms of what we're all here to do, which is try to make improvements in public health.

I also believe it is not a new concept because composite outcomes have been around for a long time. They have been and continue to be used heavily in cardiovascular studies, for example. We also thought about intercurrent events before the E9 document came out when we considered the effect on drop-in and drop-out on our delta and the impact of such events on power calculations. What I really like about E9 (R1) is that it provides a logical framework that is repeatable. The document includes four attributes to describe an estimand that we can and should start seeing in our protocols:

> … an estimand defines in detail what needs to be estimated to address a specific scientific question of interest. A description of an estimand includes four attributes: a) the population, that is, the patients targeted by the scientific question; b) the variable (or end point), to be obtained for each patient, that is required to address the scientific question; c) the specification of how to account for intercurrent events to reflect the scientific question of interest; d) the population-level summary for the variable which provides, as required, a basis for a comparison between treatment conditions. Together these attributes describe the estimand, defining the treatment effect of interest.

The specification of how the analysis will account for intercurrent events has the potential to generate

continued debate. The one that gives me the most pause is the principal stratification strategy, defined in E9 (R1) as:

Principal Stratification: Is the classification of subjects according to the potential occurrence of an intercurrent event on all treatments. With two treatments, there are four principal strata with respect to a given intercurrent event: subjects who would not experience the event on either treatment, subjects who would experience the event on treatment A but not B, subjects who would experience the event on treatment B but not A, and subjects who would experience the event on both treatments[3]

I worry that it will be the approach most prone to misinterpretation and misuse. As well-written as the Guidance is, the interpretation and implementation are what we, as clinical trialists, have to watch and be sure to recalibrate as needed so we stay true to the overriding principles of the document.

We have heard this morning about pre-specifying sensitivity analyses to be conducted as part of our final analyses. What I did not hear today is the importance of that discussion occurring in the design phase. How well are we assessing our parameter assumptions when calculating sample size and power? Is our design and our sample size robust enough to handle events both known and unknown, as we conduct our trials? How much confidence do we have in our control estimates, the variability of the outcome measure and our expected delta? Even with a perfect design, the importance of the quality of how we operationalize our designs can have a huge impact on the outcomes of our studies and undermine all that we pre-specify. For example, as we learned early in our careers, the best way to adjust for missing data is not to have any. That being said, we are dealing with people and there will certainly be missing data and intercurrent events. We can and should be placing high importance in the design phase on understanding how a protocol is best implemented in a clinic, and training study staff on implementation and methods to reduce missing data and intercurrent events when such events can be influenced without impact on patient safety. Trying to mitigate sloppy implementation with more refined and defined statistical methods is not the solution for better estimates and inference. For example, consider the intercurrent event of premature treatment discontinuation. Who made that decision to discontinue treatment and how was it made? Did we specify in our protocol the parameters under which it's allowable to stop, and was there agreement among the investigators? Do we have buy-in from all the physicians who may be influencing the study participant, whether they are part of the participating clinic or are a referring physician from a different practice? If we do and there is training regarding the circumstances under which a participant may be removed from treatment and a rationale, we can minimize losses related to inadequate trial staff training. Although many of these concepts are not new, our implementation of them is a bit like the medical community implementing new treatments shown to be efficacious in clinical trials. Balas and Boren[9] suggested that it takes 17 years for a clinical result to be put into practice. My hope is it will not take us 17 years to have these discussions. We must be thoughtful and take adequate time and have adequate discussion on design and implementation before we launch trials. It is the time taken to consider the "what-ifs" and create the "what thens" that will improve our studies. Often we and the sponsors are in a rush to enter the first patient. We must recognize that work we do up front will have long-term payoffs. So, I do hope the document will encourage more thought in this space.

My final point is this: we can keep using our statistical techniques to generate more and more precise outcome estimates that we use for inference, but think about what we're doing as a clinical trial community. We talk about a population, and that's what we're inferring to, but are we really considering the target population in practice? We create such tight inclusion/exclusion criteria that we may no longer have representation of the population that will likely be prescribed the treatment and to whom the treatment will be marketed. In addition, we select clinics and physicians who come with cohorts of patients who may not be representative of the population at large that we intend to target. We try to overcome that selectiveness by conducting multicenter and multicountry trials. It helps but does not solve the problem. So perhaps we can improve precision by defining estimands and intercurrent event strategies and through excellence in operations, but let us not be fooled that somehow this gets us closer to the truth for the actual population that will be targeted. We operate our trials in a flawed system and if we accept that concept we will not lose our skepticism even when we adopt the principles espoused in this new E9 addendum to refine our estimates and inferences.

Jay P Siegel: In the presence of rapidly advancing science and evolving thinking regarding missing data, the draft ICH E9 (R1) addendum is the third major effort to address the missing data issues for clinical trials in the regulatory setting. About 20 years ago, ICH E9 presented a high-level overview, highlighting some of the issues with missing data and ITT approaches. Subsequently, after seeing many problematic approaches, e.g., misuse of "last observation carried forward," and being aware of methodological advances, FDA asked the National Research Council of the National Academies to put together an expert panel and prepare a report. Many of the members of that panel are

here, including myself; our report addressed many of the same issues that are addressed in R1. The panel report speaks at some length about estimands and about sensitivity analysis, but it really focuses on missing data and how best to avoid it or to impute it. ICH E9 (R1) provides a lot more new focus and insight in an important, related area: not where the data are missing and needed, but where the data, whether present or missing, are not deemed relevant to the effect one wishes to estimate. For example, in a study targeting preservation of renal function using a serum creatinine end point, creatinine measures after progression to dialysis, while not missing, would be less informative than imputation of what creatinine would have been without dialysis. While this is not a missing data problem per se, it has much of the same flavor.

ICH E9 (R1) provides a valuable conceptual framework and taxonomy for such issues. It's an outstanding document, in significant part because it will sharpen people's thinking as they plan trials, thus making for better trials; it will lead to better dialogue between sponsors and regulators; and it will facilitate advancement of innovative approaches to address problems that we've always faced but we have not really been able to articulate well. A taxonomy and definitions of conceptual framework terms are really useful for all those purposes.

There are a few terms that need more work. One of them is "missing data" itself. I believe there is general agreement that data are not missing if they do not exist, such as laboratory values after an intercurrent event of death. But there is also need to distinguish those areas, such as in the post-dialysis creatinine example, where we have the data, but because of a rescue medication or procedure, they are not really the data we want. There needs to be some terminology for such data—I would prefer that it not be "missing data"—so that we can sharpen our thinking and our understanding in our discussions. Two other terms that have been used in discussions of missing data, and that clearly need sharper definitions, are "symptom trial" and "outcome trial." While a trial with a primary end point of HbA1c, cholesterol or blood pressure is not an "outcome trial," referring to it as a "symptom trial" is confusing because none of those measures are symptoms. Are "symptom trials" trials of symptoms? Are they trials of interventions intended to affect symptoms but not the disease course? Are they trials of drugs that have a short-term effect but don't have any persisting effect after they are discontinued? All those different situations have different implications. So as we move toward a clearer taxonomy, that area could use some attention. That said, ICH E9 (R1) is an important step in moving toward clearer terminology and, as a result, clearer thinking and communication.

I want to comment next on Dr Permutt's discussion of trimmed means as I think these encompass some very useful approaches and merit more attention. First, in clinical practice there's a common setting that calls for a specific estimand, perhaps using a principal stratum. In many diseases or chronic conditions where there are several therapeutic options, a physician and patient will decide to start one for a limited trial period. If the patient tolerates the drug and symptoms or markers appear to respond, the drug is continued; otherwise it is switched. That approach is very common not only for symptomatic therapy as for pain but also in disease-modifying therapies where there is an early indicator of response, such as serum glucose in diabetes control, cholesterol in hypercholesterolemia and blood pressure in hypertension. The therapeutic trial approach is used a lot even with disease-modifying therapies for rheumatoid arthritis, where the primary therapeutic goal may be to preserve the joints in the long term, but the treatments are judged in the short term, often by the relief of pain. In the therapeutic trial setting, you really want to know two things before starting a therapy. What is the likelihood the patient will tolerate the therapy and respond, and, if he or she does tolerate and respond, how large is the response likely to be? Assuming, as is often the case, that a brief, failed therapeutic trial is not likely to cause lasting harm, those two factors—the probability of success (tolerating and responding) and the extent of benefit in those achieving success—are what are going to help decide the therapeutic strategy. So a trial estimating them would be valuable, but estimating the latter (i.e. the extent of benefit in responders/tolerators) is a challenge with many of our classic approaches due to the lack of an appropriate comparator. One simply cannot identify on the control arm, the subset who would have been responders/tolerators had they received the study drug.

I believe that difficulty has prevented us from identifying what could be some very important therapies. For example, if you had a pain medication that was extremely safe and was far more effective than narcotics in 10% of people with severe cancer or joint pain, but ineffective in 90%, a lot of approaches would fail to identify efficacy of such a drug; yet such a drug could be extremely valuable medically and potentially in the market as well, for the 10% who do respond well. Identifying such a drug could be facilitated by quantitating efficacy in responders, and trimmed mean approaches[10] are among those that offer promise in such settings.

Other design approaches have been useful in estimating efficacy in people who tolerate and appear to respond. We can do run-in trials to find responders/tolerators and then randomize them. We can use enrichment designs when we can identify likely responders/tolerators prior to the trial. We can do randomized

withdrawal trials, in which we only randomize people who appear to be responding to and tolerating the drug. All those designs are probably underutilized, but all have limitations; they are part of a solution, not a whole solution. I'm quite attracted to the use of a unilateral trimmed mean as an additional part of the solution. From the study drug arm, you can trim off all study participants except those who tolerated the drug and who had a response that, if observed in a therapeutic trial setting, would have been sufficient to keep them on the drug. Lack of tolerability and inadequate efficacy, both leading to trimming in the study analysis and to discontinuation in the therapeutic trial setting, can be considered as equivalent outcomes with regard to efficacy, assuming there isn't persistent harm to a failed therapeutic trial. Either way, the patient is not going to stay on the drug; they will proceed to a different therapy. In the control arm, after trimming any non-tolerators, you would rank responses and trim the lowest responses such that the proportion trimmed on both arms is the same. The remaining group comprises the participants on control with the best responses, and thus their response is a least as good as those on the control arm who would have tolerated and responded to the study drug, had they been on the study drug arm. So, comparing them with the study drug tolerators/responders may underestimate drug effect in the principal stratum of those who respond and tolerate (especially when the placebo effect is large and variable), but, importantly, could not overestimate the drug effect. In such a manner, the trimmed mean allows you to focus on the population you're really interested in, the people who are going to stay on the drug when given a therapeutic trial. If you can estimate efficacy in that population and you know the size of that population, you know a lot of useful information.

Finally, I would like to speculate a little about possible uses of trimmed means approaches in other settings—outcome trials and trials where there's substantial chance of both harm and of benefit (Figure 1). For example, a fibrinolytic therapy in stroke might improve blood flow and help some recover, but might also increase bleeding and cause some to die or have severe disability. Both benefit and harm are reflected on the same stroke outcome scales and the effects of drugs that benefit some and harm others are not well characterized by central measures. A bilaterally trimmed mean approach, allowing estimation of the harm in those who are harmed and the efficacy in those who are benefited, would be interesting to consider in that setting.

Geert Molenberghs: Dr Permutt talked about the differences between clinical trials and survey sampling. There has been quite a bit of commentary in the survey arena on the very word "estimand," which is not a word in the English language (at least not in the Oxford English Dictionary[12]), but it is a word nevertheless because it's jargon now and has been for almost a century. In many survey sampling contexts, people would start from the estimand; a basic but important point of our discussion is that in clinical trials research we have now started to think about the basics and where we get started, and that is indeed the case with the estimand. There are classical concepts in survey methodology that to some extent translate to the clinical trials world, but as Dr Permutt said, there are also very important differences. Let me take you through three situations that will help to offer a viewpoint.

The first situation is your standard randomized clinical trial, where we have two outcomes, the patient outcome on the placebo and the patient outcome on the active treatment. From a simplified perspective, the typical quantity you are interested in is the individual treatment effect, and then of course the expected value of that. No covariates here, so this example is a little oversimplified. Further assume that there is 50% missing information, not because you failed to record it, but because it's unobservable. So, if the estimand here is the average or expected treatment effect in the population as a whole, where do we get our information from since we only observe half of what we really want to observe? The strength of the randomization paradigm allows us to reliably and meaningfully estimate the average effect of treatment. If properly conducted following the state of the art, we do have an estimator. This is a very simple case where we get our information, and that means not just that we can estimate quantities, but we can do so reasonably unambiguously without there being unverifiable assumptions. Indeed, we get information from the design, data and assumptions in general, but (unverifiable) assumptions are now not needed. We have the data, and we have the powerful resource: design. We should never forget that and this is maybe the reason why 50–60 years ago so many people, including some founding fathers of our field, were vigorously promoting randomization as the only way to draw valid conclusions. For example, in *The Design of Experiments* Fisher states "The purpose of randomisation … is to guarantee the validity of the test of significance, this test being based on an estimate of error made possible by replication."[13] For a more detailed account, see the study by Armitage.[14]

Let me go on to surrogate end points in the same setting as a moment ago, with treatment under placebo and under active treatment. But now, in addition, we have a surrogate and for now let's say the surrogate is just measured once (we expand on this later). So there are still the same quantities as before, but the quantity we're interested in now is a so-called predictive causal association. Whereas for the treatment effect as before, we still have all the information we need, this is not true for the correlation. This correlation between the true

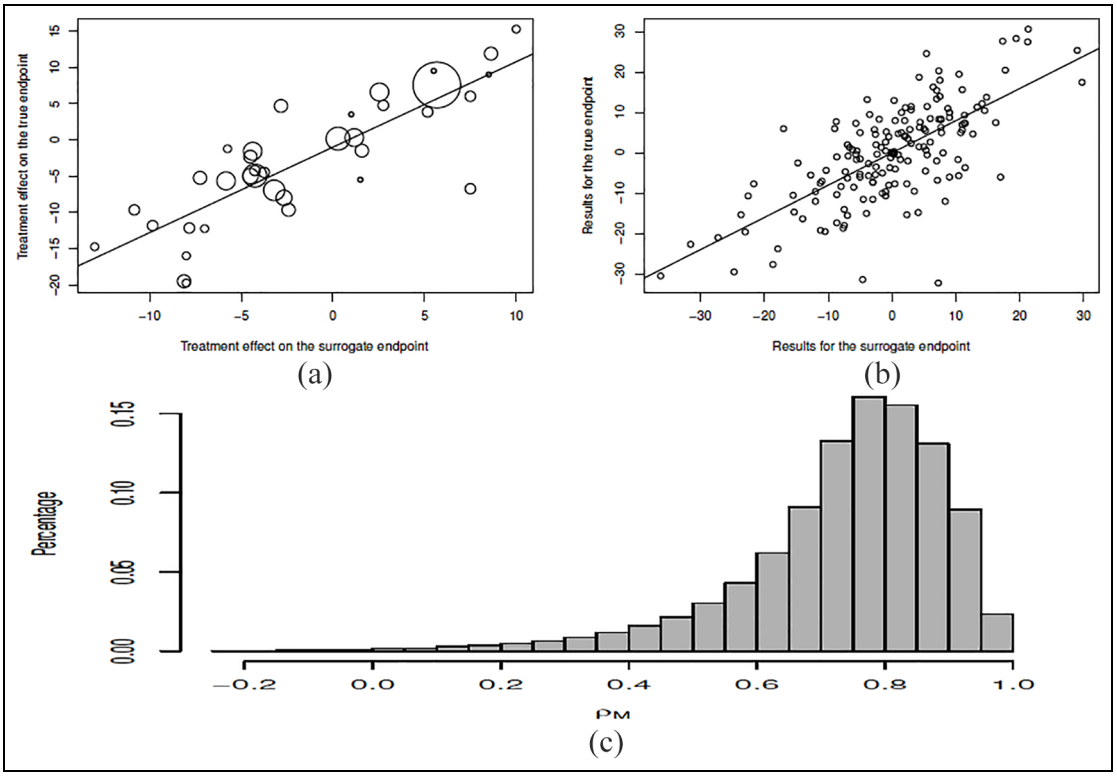

Figure 1. Age-related macular degeneration trial (with permission): (a) treatment effects on the true end point versus treatment effects on the surrogate end point in all centers; the size of each point is proportional to the number of patients in the corresponding center (b) True end point versus surrogate end point for all individual patients, after correction for treatment effect (individual-level surrogacy). (c) Histogram of the meta-analytic individual causal association (MICA).[11]

end point under placebo and the true end point under active treatment is non-estimable, but it is needed for the quantity of interest, the correlation between an individual's treatment effect and their surrogate value. So, if you want to identify this estimator—you have the data, you have the design, but that falls short of identification—you can make identifying assumptions and put them on the table, but that's unfortunately not always done. Or you can perform a sensitivity analysis, ranging over all possibilities uniformly or by favoring certain values of the unidentifiable parameter, that is, correlations that are not identified, and take it from there. So what would you get in a case for example, like an age-related macular degeneration trial? You will get a whole range of possibilities. That may sound awfully imprecise at first. But in the specific example of age-related macular degeneration, we obtain a correlation distribution that concentrates on the smaller values, which will allow us to draw useful conclusions.

Now consider a third situation, but in addition think about the surrogate in its two appearances, under placebo and under active treatment. It's a little more symmetric, a little more satisfying maybe, a little bit more causal, but there are also a few more unidentifiable parameters. We have four outcomes, two treatment outcomes and the two surrogates, for every patient. If you're lucky you observe two of these four. If you're unlucky you observe nothing, but two of the correlations in that 4×4 matrix, defined by the surrogate end true outcomes, both under placebo and active treatment, are identifiable from the data. The others are not. So, again, you can take it as it is or put in some unverifiable assumptions if you really want a causal interpretation, or you could use some causal diagrams to exclude certain relationships, but the sensitivity issue remains on the table. I illustrate these points with the results of the Age-Related Macular Degeneration Trial. In this case you would get graphics as shown in Figure 1. Figure 1(a) and (b) offers a comparison with the classical approach where identifiability would not be a problem. Figure 1(c) shows the impact of the unidentified correlations on the assessment of surrogacy. Clearly, the individual-level surrogacy (Figure 1(b)) suggests a high correlation. This is indeed nicely paralleled by Figure 1(c), which suggests that the correlation of ρ_M, that is, the causal version of individual-level surrogacy, is very likely to be high.

The settings we have been talking about today involve enrichment of data that is less than perfect (Table 1). By that I mean on one hand incomplete data, censored data and several other situations that may arise, non-compliance very importantly, is coarse data.

Table 1. Enriched data.

Coarse data	Augmented data
Incomplete data	Randomized studies
Censored data	Random effects
Joint models	Latent classes
Grouped data	Latent variables
Non-compliance	Mixtures

"Coarsened" studies are used as opposed to augmented studies; what I've been doing in the examples as really augmentation.[15] Even if nothing is missing haphazardly, by design there are many things you cannot observe, and that's not dissimilar from latent classes or random effects in mixed models, and so on, but it's in some sense clean. It is design-based. You know beforehand what you're going to observe and what you're going to miss, that is, what is unobservable. In other words, one either observes the treatment effect under placebo, but then not under active treatment, or vice versa. So, the right-hand side of Table 1 is more design-based and therefore cleaner. On the left-hand side, it's more patient-driven.

You just have to see what happens in the field while collecting data. So, we have an increasing complexity in what we have now gone through. If I go back to the beginning, in the standard clinical trial, with no surrogacy and no missingness, the design compensates for what is unobserved without further assumptions. Or if you wish, you can say all the assumptions are born out of the use of the powerful randomization concept, and you preserve causality. We know that in practice, of course, we have to discuss a number of pragmatic issues that get in the way of this clean situation, but that's pretty much what it is. In the surrogacy case, it's still augmentation. It's still clean design-based, but the data plus the design fall short of identifying all the quantities that you have, so the worst you can do is make assumptions, and bury them, and hope that nobody notices. If you make assumptions, you should put them on the table, or, if you don't want to make assumptions, you just want to span the entire sensitivity space, that's fine also. I'm not taking a position here on which is better, but these are your options. But in the incomplete data and non-compliance case, there's coarsening. So, sensitivity is not observation-based and subjective choices are absolutely unavoidable. There are many things that we can do. There is the interference of intercurrent events we have to deal with, but at the end of the day, you're going to need scenarios about the so-called predictive distribution.

It is important to consider the distribution of what you don't have given what you do have, and in the context of missing data, it is important to understand the missing data given the observed data and given the observed covariates. A very important point, that is now, I think, blessedly taken for granted, is if you consider various alternative scenarios, they should preserve the estimand. If you think they shouldn't, you have to make a pretty good case.

Thomas Permutt: Dr Rockhold asked—What if the effect in compliers is good but nothing else is good and the ITT effect in particular isn't good? What might I do as a regulator about such a study? The answer is that it depends on the clinical setting. In a setting where the drug kills most of the patients, but the ones who survive do pretty well, the effect restricted to patients who comply by surviving is not very important. That's a bad drug. In a setting where the drug does nothing much for most patients but doesn't hurt them, and it helps a few, it helps compliant patients a lot, how's the drug going to be used? People are going to try it, and a lot of people are going to stop using it, and the people who don't stop using it are going to be helped, and that's a good drug. As a regulator, I have no reason not to like a drug like that.

I'll also address Dr Lindblad's comments about where the word *estimand* comes from. After we started using it again, some people found it in a book by Lehmann and Casella,[16] but the usage of the word *estimand* there is not really very relevant to what we're talking about today. Rather, the word emerged from obscurity after the report of the National Research Council[2] panel on missing data. I was in some of the meetings of the panel as the representative of the FDA, and at one point I started talking about how we requested them to discuss missing data, but there was this other question about how we were actually defining the effects of the drug. I fumbled some over how to express that, and one person on the panel finally interrupted me and said, "Yes, the causal estimand." That was the rebirth of the term in the statistics literature. It's unfortunate that *causal* got lost because I think what he really meant was the "causal" thing. *Estimand* is not the relevant part of that phrase, *causal* is.

Jay P Siegel: Both the addendum and Dr Mehrotra's talk place a lot of emphasis on the sequential approach, that is, that you should identify the question you want to answer first, and then decide upon the estimand, and then the estimator. Maybe Dr Permutt's remark reflected that when he spoke critically of investigators who say something like, "Well, here's the analysis we want to do, and we'll figure out that estimand stuff later." While the emphasis on a sequential approach is certainly understandable in that it creates a construct to prevent people from doing the wrong thing, an iterative approach is often more appropriate. As Dr Permutt concluded, sometimes it's better to get a reliable estimate of something that isn't exactly what you want than to get an estimate based on lots of unverifiable assumptions. So in reality, the optimal approach is

often iterative. The question of interest should inform selection of the estimand and then the estimator, but sometimes you should modify the question to a relevant and important variant to make sure it can be associated with a reliable and measurable estimand.

Devan V Mehrotra: That's an excellent point. In general, one should understand the question of interest, and in a natural way you're then led to the corresponding estimand using the sequential framework in the E9 (R1) addendum. But you have to think ahead in terms of the estimation as well in that sequence, and indeed it is at that point if you appreciate that it will be virtually impossible to come up with a reasonable estimator of the estimand, then you have to rethink. Indeed, our working group has appreciated that this can be an iterative process. However, what we're trying to discourage is the selection of an estimand simply on the basis of a more convenient or more powerful analysis that is misaligned with the clinical question of interest. You would agree with that, but the point you make is a very good one, that one has to think ahead about estimation.

Frank W Rockhold: To expand on Dr Siegel's comment, in many diseases that are represented in trials, toxicity, which might lead to an intercurrent event, happens long before the benefit you might see. Oncology is a good example as is diabetes. So, therefore, having that principal stratification analysis for the physician to be able to present to the patient and say, "I know these adverse events are annoying, but actually if you push through them, this is what's on the other side." If you don't have that information, it could be very difficult.

Anne Lindblad: When I review results and it looks like the compliers are benefitting, that's the impetus to do the next study; maybe an enrichment design, if it really looks like there's a population of patients that can tolerate and may benefit from the therapy. I worry that implementation of principal stratification will lead people to think that statistics can mitigate bias and provide an answer to a question, when what is needed is the next study to answer the question.

Geert Molenberghs: Still following up on this, I agree with Dr Siegel that it may be necessary sometimes to deviate from the estimator because the event actually is occurring, the perspective's not good, and actually you would go for another estimator that corresponds to a slightly different estimand. Still, you should be very clear what the original estimand is. It's fine to approximate (and if you disagree with that, there would be no Taylor series expansions), but we have to make sure we have the language and the culture to clearly say what we are doing and allow people to criticize it and come up with their own opinion.

Kristofer Jennings: One of the things that I haven't heard mentioned, which was somewhat surprising considering that this conference has addressed it in the past, is adaptive trials and other new trial techniques. In adaptive trials, the inference and estimation become prickly and difficult. Does this discussion on estimands come up when you are thinking about an adaptive trial? Or a sequential trial in general?

Geert Molenberghs: The inferences that we usually draw about treatment effects do not naturally apply in the face of very dynamic and sometimes volatile designs. People who have worked for decades in sequential trials know how difficult it is to do estimation after a standard group sequential trial. It's actually not so standard if you look at the details, and that's why there has been a debate about what you should do in terms of estimation. Testing is a bit of a different situation because that is what the paradigm has been invented for in the first place, so all developments have started from ascertaining proper operational characteristics of the tests done. But in terms of the properties of estimators, let alone the nature of the possible estimands, this is a complex but interesting area for research. Much more work is needed in that area for sure.

Devan V Mehrotra: If you stop an ongoing clinical trial early because the results are very promising, there is an upward bias in the estimated effect size, that is, the interim results tend to exaggerate the drug effect. What I find intriguing is that a drug label will show what the effect size was when that trial was stopped without any formal acknowledgement that maybe the actual effect size is not as optimistic. But interestingly enough, the same type of argument applies if you do a phase II trial and you wish to decide to invest in a phase III trial. Even if the end point in phase III is the same end point as in phase II, the presumption is that it's the same estimand when you go from phase II to phase III, so we should expect a similar effect size in phase III. But keep in mind that the first component of the estimand is the population, and when you go from phase II to phase III, oftentimes there is a tendency to expand the population. You typically have a more pristine population with fewer restrictions in phase II. You might have what you think is the same estimand in mind, but strictly speaking you've changed the population. So, in that sense you have to think very carefully about why, as is often the case, the estimate in phase III doesn't look as promising as it did in phase II. Many things contribute to that, but the key point is that the phase III estimator is for a different estimand because your population is now broader than the one you had in phase II.

Anne Lindblad: I think it's not just the population, but also the duration of the trial. Phase II often tends to be

of shorter duration, so your chance for intercurrent events is probably less both in terms of incidence of individual events as well as the types of potential events. In a phase III setting you're going to have a much wider range of events that could be affecting your estimator.

Scott S Emerson: In an adaptive clinical trial, you've changed your indication based on response across populations that we don't understand, and that leads to not understanding what the bias is. Did you curtail your population because of just random observations, or was it something that really matters? This does make it very, very difficult to appreciate what you are estimating. The only thing that helps us a little bit in this regard is probably that the damage that's done by the adaptation is less than the damage that's done by including in the label exactly what was observed in the clinical trial that led to approval of the drug, which, as Dr Mehrotra noted, is inevitably biased. There's always going to be regression to the mean. I actually worry far more about the clinical estimand than I do the statistical estimand; if you're talking about a mean or a trimmed mean or a quantile or a geometric mean or something like that, it's a clinical question, as to what's important. We frequently get asked on FDA panels, do you believe there's an effect, and we answer that question, and then say is it clinically important? I'm not going to use the same estimand to decide whether it's clinically important for exactly the reason that Dr Siegel brought up: a treatment that only works in 10% of the population, but the population is taking it chronically and they can safely self-identify whether it works or not, we call that a good drug. So I'll judge the clinical importance by what's the magnitude among putative tolerators/responders/compliers, the people who will really take it chronically, and that may take multiple clinical trials to do. Sometimes we mess up by trying to answer every question in the same clinical trial.

Mat Davis: As we are pre-specifying sensitivity analyses, should we be designing trials that are powered for only the primary estimate, or should we design trials that are also powered for sensitivity analyses? Because such duality has a real potential to increase the sample size of our trials to the point where feasibility would be a problem.

Jay P Siegel: In addressing that question, I want to make a clear distinction between issues with sensitivity analyses versus issues with secondary analyses or secondary estimands. A properly constructed sensitivity analysis for missing data should generally use the same estimand as the primary analysis and test the sensitivity to missing data assumptions that underlie the primary analysis, by exploring alternative assumptions. Hence, it should generally involve mostly the same trial design. However, sample size and power considerations will vary, and the potential need for a larger sample will depend on many factors including the reliability of the assumptions, the robustness of the findings, and the ability to minimize missing data.

Another approach to exploring the robustness of a finding with missing data is to test different but related estimands that are less dependent on missing data. While such an approach does not directly test missing data assumptions, findings on estimands that are less sensitive to missing data may be helpful. The question of the extent to which you should design a trial to measure more than one estimand is closely related to the question of whether you should design a trial to measure more than one end point. A trial optimized for the primary estimand may not be well designed to measure others.

Anne Lindblad: Sponsors should explore the parameter space around the assumptions that go into an estimand to make sure the trial is adequately powered before it starts. How robust is your design to those assumptions? Under which scenarios is it well powered and under which is it poorly powered and what risk tolerance do you have based on that information? There will be unknown, unplanned, intercurrent event rates as well that can influence your outcome. At the end of the day, you are more likely to be successful if you consider these aspects and make design and plans for implementation decisions accordingly.

Geert Molenberghs: You can actually think of the question a little differently. As a thought experiment, say I want to perform a tipping point analysis, maybe not as a sensitivity analysis, but as my primary analysis. You can try to work out the power for that, but I think you're going to be pretty disappointed because the required sample size will be huge. So you have to treat that cautiously, and of course, there may be some evidence for efficacy coming, for example, from a surrogate marker. You could say, let's use that surrogate marker, and we are then going to see what sample size we need to use the surrogate marker and then translate the effect that we get from the surrogate onto the true end point. If you want proper power for that, you may not get there. Infinity may not be enough. Even if it's finite, it's maybe not feasible. Then, it may be better to say, if we think this is a good surrogate it will become our new true end point the next time, maybe just out of pragmatic considerations.

Devan V Mehrotra: Our ICH assembly includes the senior executives on the pharma side and on the regulatory side, and a typical inquiry is something like "we don't understand the technical details of what you statisticians have produced, but we are wondering if your work is going to make drug development more expensive." Our response is that we don't automatically have to plan for larger studies. In fact, it could be true that in some cases the trials indeed will be larger, but larger because you're

doing the right thing by not using an analysis that exaggerates the drug benefit. However, methods are being developed, like the one I just presented, which can simultaneously deliver honest estimates while preserving good power so that an increase in sample size is not necessary.

Dr Lindblad said earlier to plan ahead in terms of your sensitivity analysis, which of course is needed to test key assumptions in the main analysis. To state the obvious, it's best to not make unrealistic assumptions. Interestingly, we have examples where, by adopting the E9 (R1) framework and planning ahead for the sensitivity analysis along with minimizing unnecessary assumptions, the sample size relative to current practice has actually come down. So, I just want to reassure you that it will not always go up. Dr Permutt has made this point very eloquently in our working group meetings as well based on his trimmed mean analysis idea.

Marc Walton: We are talking chiefly about regulated drugs, and that there are two big decision-making events that happen in this process. One is that the regulatory agency has to decide about whether or not to approve a drug because they think it does something worthwhile for patients. The second is that patients and physicians need to decide whether or not to use the drug. Those are really two questions that differ in a very important way, and the choice of estimands for those two purposes may very well be different in many cases. Perhaps, for many trials we need to create two different estimands that will serve the two different purposes. Traditionally, we have had one estimand, and it's made to serve both purposes, but may serve one less well than the other. So perhaps what would be valuable is one estimand that is designed to aid the regulatory agency in answering the question "does the drug do something?" That's a binary question. A different question, though, is what the patients and physicians need to know: "how much does the drug do?" That's a quantitative question.

Perhaps many clinical trials, particularly like those for diabetes that have rescue medications need to have two different estimands of that primary end point, one intended to aid the regulators in the yes–no initial step of their decision-making, and the other (which actually winds up in the labeling) that is deemed the fairest way to help patients and physicians understand how much benefit there is to the drug. I would be interested in hearing comments on that.

Frank W Rockhold: I think if you design your program strictly around targeting what's approvable, I agree with you that the estimand chosen may not address the issue of what proves to be useful. I would say that if you can't actually answer both questions then start by only looking at the approval question. I think you need to design your program around what's useful and then go backwards.

Eric J Tchetgen Tchetgen: Ultimately, the question is what are you going to do? What's the decision you're going to make? This is a causal decision. You want to be able to intervene. Now, there are some estimands that are more controversial than others in the causal inference literature. I come from that world in epidemiology. The principal strata causal effect is more of an exploratory causal effect estimand, and gives me some concern, because you have no idea who the compliers are. You might get an estimate of the proportion of compliers under certain assumptions, which is a useful quantity, but in a new population, you have no idea how to identify individuals who would comply irrespective of what they're assigned to. In addition to that, in a survivor average causal effect context, it's even worse. You have no idea of who the "always survivors" will be. I think it's a much taller order to say we should approve a drug if we have no idea who the drug would work for or what proportion of the population might respond to it.

Chunqin Deng: My first question is about the transition from the concept of ITT to estimands. Years ago, we were talking about the strict definition of the ITT principle. There are a lot of "regardlesses" in the ITT definition—for example, regardless of what happened after the patient was randomized. Our discussions this morning focused on protocol compliance. What about the rest of the "regardlesses" in the ITT definition? For example, if a patient is dispensed the drug in error, based on the ITT principle we would need to analyze 'as randomized'; but now with the concept of estimands, if the patients have received the wrong drug, I'm assuming we will do the analysis based on "as treated" instead of "as randomized." So we have moved away from the strict ITT principle to "practical" ITT, then to "modified" ITT. Over the years, clinicians started to understand and accept the ITT concept. Now as we introduce this new concept of estimands, when we communicate with clinicians we may have some difficulties in explaining to them about the transition from the traditional ITT principle to the concept of estimands. My other question is regulatory. How much will the regulatory agency (such as FDA) actually embrace this new concept of estimands in the drug approval process?

Frank W Rockhold: If you read the Guidance all the way through, it actually stresses retaining the principles of randomization. ITT, as the document points out, is a difficult term to interpret in a lot of trial designs, but I don't think anybody is suggesting we move away from the principles of a trial design and use of randomization. It's just a matter of how to interpret ITT based on randomization alongside the desired

estimand. So I don't think there's a risk of abandoning ITT. The Guidance is not suggesting that we move away from something that served us well for 50 years.

Anne Lindblad: The language I grew up with was "once randomized always analyzed," and that language fits right into this discussion. How you plan to analyze randomized participants and the potential bias resulting must be specified in advance based on your assumptions regarding the different case scenarios that could occur throughout the trial. Excellence in implementation can prevent or at least minimize avoidable missing data or improper intervention administration. We must follow all randomized participants through to the end of the trial regardless of the intervention actually received, provided doing so does not increase risk to the participant. The intended course should be predetermined in the protocol based on the potential intercurrent events. We do not want to be left with informative missing that we ignore and as a result introduce bias, both known and unknown.

Roderick J Little: Definitions of ITT often confuse the estimand and the method of estimation. I think one good feature of this estimand terminology is to make clear the difference between what it is you're trying to estimate in the population and how you're estimating it. I think that is a really useful distinction.

Janet Wittes: I want to return to the issue of sensitivity analysis. I'm fully on Dr Lindblad's page in this regard: we should be thinking about the likely, or even the possible, failures of assumptions and then designing our trials around those potential failures. What I don't understand and I'm uncomfortable with is the notion that one should pre-specify your entire set of sensitivity analysis and not give credence to sensitivity analyses that are not pre-specified. That seems to me backwards—sometimes things happen that we don't expect and we ought to tailor sensitivity analyses to those surprises.

Anne Lindblad: The steps needed to minimize bias in inference are to pre-specify your primary analysis and sensitivity analyses. Pre-specification should include all assumptions and the analyses carried out accordingly. It would not preclude, in a secondary analysis, doing other sensitivity analysis that may explore conditions and events you did not foresee before the trial. The question is, should the results of those analyses inform your decisions regarding next steps for the intervention under study? I would say yes. Our decisions should be based on the totality of the information, appropriately weighted for strength of evidence and potential bias. Analyses not pre-specified should always be given less weight. Trialists need to consider and define the primary outcome, the assumptions and the sensitivity analysis around that primary outcome, and yet there remains a place for exploratory analyses that were not predefined.

Frank W Rockhold: I would think part of the answer to that question depends on what you mean by pre-specification, because there's pre-specification at the start of the trial, and then there's specification before unblinding. Sensitivity analysis is probably an area that you want to look at and tune up in the period of time before you unblind, not once you know what types of intercurrent events have occurred, with what frequency they've occurred and in which treatment arms.

Anne Lindblad: I always worry when people say I'm looking at all the data in aggregate, not by treatment, and I have no information. I don't think that is true. When a person looks at all the data, even though there is no "by treatment" analysis, that person is subconsciously and consciously influenced by that look and will make decisions that may influence subsequent inference, thus introducing bias. Here is an example. When intercurrent event rates are viewed in their totality, not by treatment, during the trial but before unblinding, and they are higher than expected, it will influence you. You may consciously or subconsciously believe that treatment is not working as well as you would hope and adjust your analysis plan to optimize detecting an effect. Regardless of whether you do or not that potential influence has been introduced by that look.

Deborah Donnell: I have never worked with a clinician who didn't just want to know "whether the drug works or not," and they have a difficult time understanding the concept of the ITT analysis. My question is about this idea of the hypothetical estimate of "how this drug would work were everybody to use it perfectly." Many people would say that is their ideal estimand, and I want to understand how the Guidance is going to deal with trials that propose this as their planned estimand. When you plan the trial, you won't know the extent to which dropout, poor adherence or safety discontinuations will affect the observed data. Is there anything the Guidance can say about when this hypothetical estimand is really not reasonable?

Anne Lindblad: This is a good point. When our assumptions do not make sense or the reality of implementation degrades those assumptions, then what? We have more work to do in that area to provide appropriate guidances and clarifications. The stratification strategy is particularly susceptible to misuse.

Thomas Permutt: Yes, I think that's a really good reason to separate estimands from estimators. I happen to

think that your clinician colleagues are wrong. If they think carefully about it, they will realize that what they want is the effect in tolerators rather than the effect if everybody tolerated. But that's not important. What is important is, I don't think anybody knows how to estimate either effect, but there are people who purport to be estimating the effect if everyone tolerated, and I can't live with those purported methods. But yes, if you're going in the direction of saying we need this other estimand and FDA is only looking at ITT, and that's really not what our clinicians want to know, then we want to say, Yes, we get it, only there's a lot of work to do to understand how to estimate those things, and it's big data work, and there are going to be lots of variables in it and lots of theory, and I'm not telling anybody that they can go out and do that right now. I'm telling people that if you want to do that, these are the kinds of things you have to think about.

Sammy Yuan: It seems that we have not talked about the estimand for safety events. We have many tools and methodologies to get a better estimate of the treatment effect, but not so much for safety, which is as important as efficacy. Actually, we have more intercurrent events for a safety end point then for efficacy. So should we have a definition of a safety estimand?

Frank W Rockhold: I agree with you. But you have to walk before you can run. If you think of the history of statistics in clinical trials, safety has lagged behind efficacy by many years. Now there's a lot more emphasis on safety, so how do you translate what's in this Guidance to a safety question? End points are obviously different. As I said previously, we should not go down the road of saying, I'm going to look at efficacy in this or that group of patients. Dr Scott Evans and others have done some nice work saying, what we should be doing to help healthcare practitioners is to look at the population of patients who respond and see what their safety data look like. Look at the population of patients who have had serious adverse events and see what their response rates looked like. If there's no overlap that gives me a very different insight then observing that all those that responded are also the people who had adverse events. I definitely think somebody should work on translating this Guidance into the safety domain, but be careful that you don't do it in a vacuum. You need to think about benefit-to-risk or benefit-to-risk "estimands" guidances because otherwise I think what'll happen is that the concepts of safety and efficacy will diverge, and you will end up with two guidances that don't operate in harmony.

Participants

Mat Davis	Teva Pharmaceuticals	
Chunqin Deng	United Therapeutics	
Deborah Donnell	Fred Hutchinson Cancer Research Center	
Susan Ellenberg	University of Pennsylvania	
Scott S. Emerson	University of Washington	
Kristofer Jennings	University of Texas Medical Branch	
Anne Lindblad	Emmes	Panelist
Roderick J. Little	University of Michigan School of Public Health	
Devan Mehrotra	Merck Research Laboratories	
Geert Molenberghs	Universiteit Hasselt and KU Leuven	Panelist
Thomas J Permutt	Food and Drug Administration (FDA)	
Yongming Qu	Eli Lilly and Company	
Frank W. Rockhold	Duke University	Panelist
Jay P Siegel	Formerly Janssen, FDA	Panelist
Eric Tchetgen Tchetgen	University of Pennsylvania	
Marc Walton	Janssen Research & Development	
Janet Wittes	Statistics Collaborative, Inc.	
Sammy Yuan	Merck	

Acknowledgements

We thank the Center for Clinical Epidemiology and Biostatistics in the Perelman School of Medicine at the University of Pennsylvania (CCEB), Genentech (A Member of the Roche Group), Glaxo Smith Kline, Johnson & Johnson (Janssen R&D) and Merck and our professional society sponsors the Society for Clinical Trials and the American Statistical Association whose structural and/or financial support has been invaluable for the success of this conference. Members of the program committee include Drs Susan Ellenberg, Jonas Ellenberg, Mary Putt, Pamela Shaw and James Lewis in the Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

External funding for this conference was provided in part by GlaxoSmithKline, Merck (Inventing for Life), Amgen, Genentech (A Member of the Roche Group) and Janssen (Pharmaceutical Companies of Johnson & Johnson).

References

1. Guidance for Industry E9 Statistical Principles for Clinical Trials. *U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER)* Geneva: ICH, 1998.
2. National Research Council. *The prevention and treatment of missing data in clinical trials. Panel on handling missing data in clinical trials. committee on national statistics, division of behavioral and social sciences and education*. Washington, DC: The National Academies Press, 2010.
3. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. ICH Harmonised Guideline. *Estimands and sensitivity analysis in clinical trials*, http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/E9-R1EWG_Step2_Guideline_2017_0616.pdf (2017; accessed 6 March 2019).
4. Rockhold FW and Ruberg SJ. ICH-E9 reflections and considerations. *Pharm Stat* 2008; 7(4): 233–235.
5. CONSORT. Transparent reporting of trials. http://www.consort-statement.org/ (2010; accessed 17 October 2018).
6. Tamblyn R, Eguale T, Huang A, et al. The incidence and determinants of primary nonadherence with prescribed medication in primary care: a cohort study. *Ann Intern Med* 2014; 160(7): 441–450.
7. Span P. The clinical trial is open, the elderly need not apply health. *The New York Times*, 13 April 2018, https://www.nytimes.com/2018/04/13/health/elderly-clinical-trials.html
8. Mosteller F and Tukey JW. Data analysis, including statistics. In: Jonas LV (ed.) *The collected works of John W. Tukey: philosophy and principles of data analysis 1965–1986*, Vol. IV. Boca Raton, FL: CRC Press, 1987, pp. 601–720.
9. Balas EA and Boren SA. Managing clinical knowledge for health care improvement. In: Bemmel J and McCray AT (eds) *Yearbook of medical informatics 2000: patient-centered systems*. Stuttgart: Schattauer Verlagsgesellschaft mbH, 2000, pp. 65–70.
10. Permutt T and Li F. Trimmed means for symptom trials with dropouts. *Pharm Stat* 2017; 16(1): 20–28.
11. Alonso A, Van der Elst W, Molenberghs G, et al. On the relationship between the causal-inference and meta-analytic paradigms for the validation of surrogate endpoints. *Biometrics* 2015; 71(1): 15–24.
12. Oxford English Dictionary. http://www.oed.com/ (2019; accessed 9 January 2019).
13. Fisher RA. *The design of experiments*. 6th ed. Edinburgh: Oliver and Boyd, 1951, p. 26.
14. Armitage P. Fisher, Bradford Hill, and randomization. *Int J Epidemiol* 2003; 32(6): 925–948; discussion 945.
15. Molenberghs G. Incomplete data in clinical studies: analysis, sensitivity, and sensitivity analysis. *Drug Inf J* 2009; 43(4): 409–429.
16. Lehmann EL and Casella G. *Theory of point estimation*. New York: Springer Science & Business Media, 2006.