Commentary on Hey and Kimmelman

Steven Joffe1 and Susan S Ellenberg2

In classical randomized controlled trials (RCTs), randomization probabilities remain fixed throughout the duration of the study. However, in any trial that ultimately shows one arm to be superior, data will at some point provide mounting though not yet definitive evidence in favor of the better performing arm. In trials with serious, irreversible outcomes, a data monitoring committee is typically charged with regular review of emerging data and with making recommendations about early termination if extreme effects are seen.1 Still, commentators struggle over the necessity of randomizing participants to the arm that appears inferior based on emerging data.2–6 Some investigators and ethicists have advocated response-adaptive over fixed randomization to maximize the proportion of participants assigned to the better performing arm.

In this issue of Clinical Trials, Hey and Kimmelman7 advocate a “rebuttable presumption” against adaptive randomization, at least in the setting of two-arm trials comparing a novel intervention to a control. They offer a range of arguments to buttress their conclusion. First, they contend that advocates of adaptive randomization overestimate the likelihood and magnitude of benefits that it offers participants. Studies may adjust randomization probabilities based on surrogate endpoints that don’t correlate with clinical benefit; clinical endpoints may be too delayed to allow timely adjustment; the magnitude of incremental benefit associated with novel therapies is often small; and the complexities of implementing adaptive designs may preclude their use in most settings. Second, they note that because 1:1 randomization represents the most efficient allocation ratio, adaptive trials are likely to take longer than conventional trials, require more participants, or both. Third, they claim that adaptive trials encourage rather than dispel participants’ unrealistic expectations about the benefits of experimental therapies, thereby failing to advance, or even undermining, informed consent. Finally, they note that changes over time in the proportions of individuals assigned to each arm might interact with temporal trends in background conditions or participant characteristics to threaten the validity of the trial.

In drawing their measured conclusion against the use of adaptive randomization, Hey and Kimmelman do not conclude that such trials are unethical. In this, they are surely correct. But adaptive trials do involve important tradeoffs among core scientific and ethical values that investigators must weigh as they consider alternative research designs.

A widely recognized framework holds that ethical research fulfills seven criteria: social value, scientific validity, fair subject selection, favorable benefit/risk ratio, independent review, informed consent and respect for participants.8 Three of these criteria—scientific validity, favorable benefit/risk ratio and informed consent—are particularly relevant to assessing adaptive designs.

Scientific validity is a basic condition for ethical research. Research aims to produce generalizable knowledge; studies that are unlikely to permit valid conclusions cannot justify exposing participants to risk. In addition, the prospect of valid conclusions is what legitimizes the use of resources in the research, including money, facilities, personnel and—most important—participants. The main challenge to validity noted by Hey and Kimmelman is the possibility that subjects who enter later in the trial may differ systematically—for example, by prognosis—from those who enter earlier. Adjustment for all known and unknown factors associated with prognosis would eliminate this problem, but such adjustment is generally impossible. Thus, imbalanced allocation could lead to systematic between-group differences in prognosis that would bias the analysis of the trial.

In addition to this and other challenges to validity that Hey and Kimmelman identify, open-label adaptive trials face two further threats. First, most phase III randomized trials strictly maintain the confidentiality of

1Department of Medical Ethics and Health Policy, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
2Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

Corresponding author:
Steven Joffe, Department of Medical Ethics and Health Policy, Perelman School of Medicine, University of Pennsylvania, 3401 Market St., Suite 320, Philadelphia, PA 19104, USA.
Email: joffes@upenn.edu
interim analyses in order to protect the integrity of the trial and to maximize the likelihood that it continues to completion. If double-blinding is not used, changes in allocation will be observable by study personnel and may generate ethical tensions, as well as selection biases, in approaching potential subjects for randomization. Site personnel may become reluctant to approach some or all eligible individuals if they learn that one arm is performing better, as was documented for cancer trials during the era when investigators routinely had access to interim results. 9 Also, if those already randomized become aware that they are assigned to the arm that appears inferior, they may be more likely than other participants to withdraw from the trial, leading to the possibility of bias due to differential loss to follow-up. These issues will be less problematic in double-blind trials.

The requirements for social value and scientific validity derive in part from the obligation to avoid wasting resources in research. This obligation also grounds a duty of efficiency. All else equal, investigators should prefer the design that minimizes the use of resources without compromising high standards of validity. The duty of efficiency is not absolute—considerations of beneficence, equity or respect for participants might warrant relaxing or overriding it. Nevertheless, the sacrifice of efficiency that adaptive trials entail requires justification based on other more weighty values.

The primary argument for adaptive randomization is that it improves the benefit/risk balance for study participants. The argument is most compelling in studies with clinical endpoints that occur with short latency and can be used to adjust randomization probabilities in a timely manner. In such trials, when strong trends emerge in favor of one arm, an increased probability of assignment to that arm likely improves the prospects of participants who are about to undergo randomization. Thus, from a consequentialist point of view, the average outcome among all participants in the trial will likely be better if it uses adaptive randomization than if randomization probabilities remain fixed.

On the other hand, because patterns of outcomes can be unstable, particularly early in a trial, the allocation ratio in an adaptive design may oscillate over time. Such oscillations might result in a need to enroll more subjects, including in the poorer performing arm, than would have been the case had the trial used fixed randomization. If the goal is to minimize the number of people randomized to the poorer performing arm, especially when differences in outcomes turn out to be large, the optimal approach may be fixed randomization with a sequential design.

When adaptive trials begin to tilt randomization probabilities in favor of the better performing arm, they abandon any pretense of indifference between the treatments under study. Although doing so is not necessarily ethically problematic, 10 the acknowledgement that mounting evidence favors one treatment over the other raises the question of what investigators should disclose to prospective participants in the context of consent. Because knowledge of adaptive randomization will likely enhance the trial’s appeal, investigators will be motivated to share this information to encourage enrollment. At the same time, in open-label trials, investigators may be reluctant to disclose actual randomization probabilities in order to avoid difficult conversations with those who are randomized to the worse-performing arm.

Three approaches to disclosure are plausible. First, investigators might disclose only the fact of random assignment, without any mention of the plan to adjust randomization in light of emerging data. However, given the strong presumption of transparency and the fact that disclosing the plan for adaptation is unlikely to threaten study validity, withholding this information is difficult to justify. Second, investigators might disclose the fact that randomization probabilities may change in response to interim data, while withholding the actual probability that prospective participants will be randomized to each arm of the trial. (In blinded trials, actual randomization probabilities would be withheld from investigators as well as from participants.) Finally, investigators might disclose both the use of adaptive randomization and the actual randomization probabilities at the point at which the individual is considering participation in the trial.

To protect their integrity, classical RCTs commonly limit knowledge of the results of interim analyses to the study statistician and members of the data monitoring committee. Withholding this information from investigators and participants involves no ethical breach, so long as the fact that interim data are withheld is disclosed to prospective participants in the context of consent. By analogy, withholding information regarding the actual probability that an individual will be randomized to each arm of the trial is also ethically acceptable, so long as consent discussions convey that randomization probabilities may vary in light of emerging data. Withholding this information, however, is likely feasible only in blinded trials or in trials that enroll few participants at each site.

Full disclosure regarding adaptive randomization, including sharing of the actual randomization probabilities at the time the individual is considering joining the trial, presents a difficult communication challenge. Participants who are aware that randomization favors a particular treatment will naturally hope for assignment to that group. If, upon enrollment in an open-label trial, they learn that they have been assigned to the worse-performing treatment, they may withdraw immediately, thereby endangering the validity of trial conclusions. Even if they only later learn their
assignment, the knowledge that they were assigned a treatment that was expected at the time of enrollment to be inferior will likely lead to distress.

In conclusion, in carefully defined circumstances, adaptive randomization may enhance the prospect of benefit for individuals offered participation in a trial. Sponsors and investigators considering its use must consider whether the increased potential for benefit justifies the loss of efficiency and the threats to validity—especially but not only in open-label trials—that adaptive randomization poses. Finally, they must recognize that adaptive randomization has both benefits and risks in the context of informed consent. Adaptive randomization will likely increase enthusiasm among those considering participation in a trial, and will please those who are assigned to the favored treatment. At the same time, investigators must be prepared to address the concerns of participants who learn of their assignment to the worse-performing arm.

Declaration of conflicting interests
The authors declare that there is no conflict of interest.

Funding
This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

References
Uncertainty about the risks and benefits of medical treatments is burdensome and costly: burdensome because it leads to some patients being treated with ineffective drugs and others being deprived of effective drugs and costly because healthcare systems pick up the tab for such errors.

Just societies have evolved a division of labor that—if it functions properly—minimizes the costs and burdens associated with medical uncertainty. It works like this: research systems first reduce uncertainty to a level that is tolerable for patients and worth the price. They do this by implementing a suite of techniques—like blinding, randomization, study registration, and pre-specified hypotheses—that promote valid causal inference from observations in limited populations. Then, healthcare systems take up these inferences and apply them to large, less well-defined populations.

Because uncertainty is costly for patients and healthcare systems, a test of whether a research system is working properly is whether it reduces uncertainty quickly, sustainably, and as frugally as possible. Speed is important because the longer uncertainties persist, the more their costs accrue. Sustainability is important because the demand for medical evidence is constant. And frugality is important because if uncertainties can be reduced with fewer resources, surplus can be redirected toward other, costly uncertainties. Any researcher who has ever watched his or her peers receive undeservedly large grants will have a sense of what we mean.

This is the conceptual foundation for our analysis of outcome adaptive randomized (OAR) trial designs. We offered four main arguments. The first three focused on two-armed OAR trials. First, we suggest that they do not improve risk–benefit for subjects, as proponents suggest. Instead, they tend to increase total burden for both patients and research systems by demanding larger sample sizes. Second, we suggest that they redistribute tensions in informed consent, rather than dissolving them as proponents have suggested. Indeed, we suggested that, if anything, OAR discourages frank discussions of risk–benefit. Third, we argued that in many circumstances, OAR designs introduce validity threats that erode the value of whatever is achieved by this method. Finally, we entertain the possibility that OAR enables more efficient designs where multiple hypotheses are tested—for example, where several doses, diagnostic eligibility criteria, or drugs are put to early phase testing. This last point was based not on the view that patients benefit therapeutically, but rather that fewer patients may be needed to whittle a set of hypotheses down to the one worth advancing toward confirmation.

We learned a lot from reading the responses to our commentary. Don Berry’s response is, in places, vinegary and strange. First, we are told that ethics is a matter of taste, not authority. Then we are taken to task for failing to appeal to an authoritative document, the Belmont Report. Dr. Berry acknowledges that our skepticism is directed primarily at two-armed studies, not multi-armed studies. He then mounts a vigorous defense of multi-armed studies. Dr. Berry also leaves our arguments about validity and informed consent unattended.

Dr. Berry makes one important point that we did not adequately address (we did in a companion piece): trials test multiple hypotheses—and that moral evaluation of risk should consider a trial’s entire bandwidth. Expanding a better performing arm will deliver greater precision on secondary hypotheses like safety endpoints or pharmacodynamics. We are open to arguments that this can defray some of the moral penalties of OAR. But one circumstance where we are not persuaded is with two-armed studies comparing novel agents against standard of care. Should the standard of care prove competitive or superior—which is at least as likely to occur as the opposite—the greater precision on secondary hypotheses has costs but little value.

Much like his colleague, J. Jack Lee questions whether we’ve given OAR a fair shake. He situates the
debate as a conflict between “individual ethics” (which OAR purportedly favors) and “collective ethics.” Dr Lee offers several defenses of OAR. First, he says that OAR performs better than fixed randomization “when the efficacy difference between treatments is large.” The problem with this argument—which is also made by Scott Saxman—is that drug developers don’t know a drug’s effect size before testing it. If they did, either they should skip the study, or perhaps design smaller studies powered for larger effects, so that interventions can be delivered to healthcare systems more rapidly. Second, Dr Lee disputes our claim that benefit on surrogate endpoints is an unreliable proxy for actual clinical benefit. “Poor choice of a surrogate endpoint is a study design flaw.” However, trials use “poor surrogates” because the alternative—no surrogate at all—is worse. Dr Lee’s third argument is that failure of more than 50% of phase 3 trials reflects inefficiency of current drug development and not a lack of merit for adaptive randomization. However, we never suggested that failure in drug development is attributable to application of this technique. Moreover, to suggest that negative phase 3 results represent an inefficiency misses the point of trials altogether. If we knew in advance that all phase 3 studies would deliver positive results, there would be no point in running them.²

Scott Saxman accepts our arguments about validity and the weaker of our consent claims. His criticisms are directed toward our benefit claims. Recall, our commentary stated that in a best-case scenario where a new drug performs better than a comparator in an OAR, the chances that patients in the better performing arm will have received a drug that is competitive with or superior to standard of care are only 50%. This estimate is based on two published reports—not our own figures.³,⁴ So, if the chances of receiving care that is competitive with or advantageous to standard are 50% in a best-case scenario, how will patients fare in more realistic scenarios? Given that only a third of candidates tested in phase 2 trials even advance to phase 3, we’d say the prospects for an advantageous risk–benefit by exposure to the new drug in phase 2 are very low indeed.

Steve Joffé and Susan Ellenberg offer an analysis of OAR that is grounded on seven criteria offered by Emanuel et al.⁵ They elaborate on other aspects of validity and consent—in particular, management of communication about allocation ratios. Joffé and Ellenberg reach conclusions that are very similar to ours. But there is daylight between our positions. The framework utilized by Joffé and Ellenberg implies a fungibility of the various ethical criteria—a quality the authors seem to endorse when they say “considerations of beneficence, equity or respect for participants might warrant relaxing or overriding [a duty of efficiency],” and “sponsors and investigators considering [OAR] must consider whether the increased potential for benefit justifies the loss of efficiency and the threats to validity.” We believe that this is a recipe for arbitrary and opaque decision making, and it potentially frustrates the rationale for dividing labor between research and healthcare systems. On our view, you shouldn’t be able to buy off an avoidable deficit in valid inference or efficiency by upping trial benefits or demanding better informed consent.

Marc Buyse agrees with our point about division of labor and the way costs, uncertainty, and care trade off with each other. He then offers an eloquent critique of OAR, illustrated with examples. He also highlights other problems for OAR—transparency issues and the sensitivity of OAR to misimplementation. The ability of a technique to perform in the field ought—in our opinion and in Dr Buyse’s—to figure in its moral evaluation.

Edward Korn and Boris Friedlin begin their analysis by highlighting validity issues that arise from time trends in OAR studies. They then go on to explain how—to paraphrase Annie Oakley—anything OAR can do, fixed randomization with interim analysis can do better. Korn and Friedlin—and Buyse too—suggest that the efficiency problems of two-armed OAR studies also dog multi-armed studies. Their concerns also give us pause and suggest that the ethics of OAR may be more problematic than we argued.

Korn and Friedlin close with some reflections about why OAR has such a hold on nonstatisticians. To their speculations, we would add optimism bias. We all want to believe that major advances against dread disease are just around the corner. But medical advance is—with rare exception—hard won, messy, and incremental. Consider the melanoma treatment vemurafenib—the “poster drug” for personalized cancer medicine and which inspired a New York Times article depicting the drama of randomizing patients to comparator arms.⁶ The median survival advantage for this drug was reported as 3.9 months against standard of care.⁷ Even making allowances for crossover, this counts as an incremental—although certainly meaningful—advance. Optimism serves research well insofar as it encourages us to persevere against long odds. Better, however, to rely on evidence when evaluating trial techniques.

Before closing, we wish to recapitulate a foundational point. Berry, Lee, and Saxman seem puzzled by our preoccupation with efficiency, feasibility, and costs. But there are no free lunches. The higher overhead of OAR is paid down using resources that might have been directed toward other uncertainties or care. Any time the welfare of one population is traded off against another’s, you have an ethics issue. In this sense, even the minutest technical and operational dimensions of trials are laden with ethical content.

In the end, our arguments concerning informed consent and OAR go uncontested. Our suggestion that OAR erodes study validity in many contexts also goes...
uncontested. Three commentators take issue with our claims about two-armed OAR, efficiency, and patient burden. We have suggested that these arguments are flawed and perhaps anchored in a dash of Pollyanna-ism. Finally, two commentators urge greater skepticism about multi-armed OAR trials. We continue to maintain that OARs in the two-armed setting should be considered unethical unless investigators can satisfy the criteria we lay out. As for multi-armed studies, the points raised by Buyse, Korn, and Friedlin raise concerns, but we suspend judgment until we can make a more careful study of the approach.

Acknowledgements
The authors thank referees for helpful comments.

Declaration of conflicting interests
The authors declare that they have no conflicts of interest.

Funding
This research received support under the PACEOMICS project, funded by Genome Canada, Genome Quebec, Genome Alberta and the Canadian Institute for Health Research (CIHR).

References
I am pleased to see ethicists addressing clinical trial design. Clinical trials represent a fundamental ethical conflict between clinical research and practice. I am less pleased with the contribution of Drs Hey and Kimmelman. They characterize adaptive randomization in two-armed trials as unethical. I do not object to the implication that I am unethical because I take their premise to be unethical. My dictionary defines ethics as “concerned with or derived from the code of interpersonal behavior that is considered right or acceptable in a particular society.” My “society” is different from theirs. Like beauty, ethics is in the eye of the beholder.

The authors seem to imply that it is unethical to deviate from the status quo. After all, most people accept the status quo. One can make no more headway arguing ethics than arguing religion or politics. But I venture that a society in which deviations from the societal norm are unethical is destined for the trash heap of evolution. In my society, challenging norms is ethical and blindly accepting the status quo is unethical.

Drs Hey and Kimmelman are naive regarding clinical trials, science, and medical ethics. Moreover, they provide little or no basis for their ethical principles. In particular, they do not refer to the Belmont Report of 1979, which is “a statement of basic ethical principles and guidelines that should assist in resolving the ethical problems that surround the conduct of research with human subjects.” Nor do they provide foundation for such statements as “A foundational ethical principle is that trials should minimize research burdens.” The Belmont Report discusses in depth the question of who should bear the burdens of research, but it takes the existence of such burdens as given. It does not address how or whether research burdens should be minimized. The only trial that minimizes research burdens is no trial. The ethical question is whether a trial’s contributions to society have utility, more than making up for its burdens. Assessments of utility are fundamentally ethical questions that reflect societal norms and must address the fundamental conflict between individual and collective ethics, including both extremes. I have space to address only a few of the points raised by Hey/Kimmelman. Suffice to say that I disagree with essentially all their points, sometimes because they are factually wrong.

Hey/Kimmelman focus on two-armed trials. I come closest to agreeing with them when they leave open the possibility that outcome-adaptive randomization is appropriate for multi-armed trials. Indeed, multi-armed trials investigating many therapies and combinations are the future of medical research. All two-armed trials may one day be regarded to be unethical! Simply having a common control arm can nearly double the efficiency of clinical trials, with additional efficiencies stemming from adaptively randomizing to experimental arms. However, even in the context of multi-armed trials, the authors fail to understand the goals of adaptive randomization. For example, these trials do test hypotheses, but many more than a single hypothesis as in many two-armed trials. And the authors qualify their enthusiasm: “multi-armed designs in early phase settings seem no more likely to enhance therapeutic benefit than two armed adaptive allocation schemes.”

This comment is emblematic of their misunderstanding. Better treatment for trial participants is seldom the principal focus of an adaptive trial. The focus is usually scientific efficiency, with better treatment of participants being a possible consequence. For example, the goal of adaptively randomizing to doses of an experimental agent is to rapidly identify important aspects of the dose–response curve, including doses to be carried into further development. Optimally learning about a dose–response relationship usually requires assigning patients to dose 0. Patients are likely to be assigned to better performing doses (dose 0 in some trials!), but that is not the goal.
I-SPY 2 is a more complicated example.\textsuperscript{8–10} It is a phase II drug screening trial in neoadjuvant breast cancer. Its goal is to identify each therapy’s signature, if it has one. A signature is a molecularly defined subset of tumors that is sufficiently responsive to merit the therapy’s graduation to phase III. I-SPY 2 has randomized over 700 patients, and it has evaluated or is in the process of evaluating eight experimental therapies in comparison with standard therapy. Two therapies have graduated to phase III, with particular signatures. Adaptive randomization is imperative for efficiently identifying signatures. Therapies that demonstrate little evidence of benefit in a subset of patients are assigned to such patients with low probabilities, possibly 0. Therapies that are performing well in some subsets move through the process fast. A therapy may graduate in one subset even though it has been dropped from consideration in another subset, and indeed this has happened in the trial. Just as for dose–response, there is a tendency to assign patients to better performing therapies, but that is not the trial’s goal. In both examples, no fixed randomization scheme could begin to achieve the trial’s goals as efficiently as does adaptive randomization.

The authors focus on two-armed trials. They offer no argument in favor of fixed randomization except that it is the status quo, and they say 1:1 randomization is “most efficient.” But they don’t say what this means (an annoying characteristic of this article generally). Perhaps it means minimizing the standard error of the difference between two sample means assuming that the two population variances are equal. But even under this restrictive assumption, equal randomization may be most inefficient. One reason is that two-armed trials can have multiple objectives. For example, registration trials must address safety as well as efficacy. Suppose regulators require 400 patients randomized to experimental therapy to demonstrate safety. Then 1:1 randomization of 800 patients would suffice, but this may be unnecessarily large. If 4:1 randomization is also acceptable, then 500 patients would suffice. But having only 100 patients assigned to control may or may not be sufficient to show the therapy’s efficacy. One can guess which randomization ratio between 4:1 and 1:1 would be best. Adaptively randomizing between these two extremes based on accumulating trial results eliminates guesswork. The final sample size might be 550, say, with 150 controls. This would minimize research burdens, including making the experiment therapy available to patients much sooner than when using 1:1 randomization. Or the final sample size might be 800, but if so then that was necessary to achieve the trial’s goals. (As a technical matter, the adaptive randomization algorithm would be prospectively defined. This enables calculating type I error rate and statistical power, perhaps analytically and perhaps via simulation, depending on the complexity of the adaptive algorithm.). Generally, it is more efficient to have an assignment algorithm that is tuned to results accumulating in the trial. This is so even accommodating the associated type I error “penalty.”

The most important ethical issue in clinical trials is the inevitable conflict between clinical practice and science. The Belmont Report provides an unambiguous resolution: clinical trials are for science and not clinical practice, stressing that they are “designed to test an hypothesis … and thereby to … contribute to generalizable knowledge.”\textsuperscript{1} Adaptive randomization balances these two desiderata, treating the next patient as effectively as possible while recognizing the need to learn so as to benefit future patients.

The Belmont Report and the associated ethical arguments should be updated. Traditionally, clinical trials have focused on large homogeneous populations. All diseases are heterogeneous, and we are increasingly understanding their heterogeneity. Consider oncology, which is a harbinger for other diseases. Biologists are slicing and dicing cancer into ever smaller subsets. Soon every cancer patient will have an ultra-rare disease. Traditional approaches to designing clinical trials are helpless and hopeless when evaluating therapies in such settings. Requiring a focus on hypothesis testing will soon make clinical research impossible. Trial participants will make up an increasing portion of the patient population, in some cases becoming the whole patient population. (Already in the United States, more than 60% of cancer patients younger than 15 years old are in clinical trials.\textsuperscript{11})

The Belmont Report focuses on “generalizable knowledge.” But advances in cancer biology are so transformative and so rapid that any scientific contributions from a clinical trial may be irrelevant when the trial is over. Whatever is one’s ethical perspective, it would be inane to randomize 50% of trial participants to each of A and B when both A and B will be passe at the trial’s end. Hypothesis testing in clinical trials will be replaced by decision analysis in which the explicit objective is delivering effective therapy to as many patients as possible.\textsuperscript{12–16} Recognizing the burgeoning number of possible therapies and the rapid turnover in what is standard therapy. Randomization will have to be adaptive, with emphasis on effectively treating trial participants.

In conclusion, these authors are naive. They don’t understand clinical research, neither as it is conducted nor as it should be conducted. I am keenly aware that this very charge was once leveled against me when I was critical of equal randomization in contrast to bandit strategies, which are adaptive.\textsuperscript{12–15,17} My critics were right. In response, I learned about conventional clinical research—including developing respect for it—and I designed and helped carry out hundreds of standard clinical trials. Over time, some of my radical ideas became accepted by others. This long experience
convinced me that although I had been naive in all ways and wrong in many ways, my initial overall view was correct. The authors may one day come to the same conclusion—which I say recognizing the obvious ambiguity!

Declaration of conflicting interests

Donald Berry is a co-owner of Berry Consultants, LLC, a company that specializes in designing Bayesian adaptive trials for pharmaceutical and medical device companies and NIH cooperative groups.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

Are outcome-adaptive allocation trials ethical?

Spencer Phillips Hey and Jonathan Kimmelman

Abstract
Randomization is firmly established as a cornerstone of clinical trial methodology. Yet, the ethics of randomization continues to generate controversy. The default, and most efficient, allocation scheme randomizes patients equally (1:1) across all arms of study. However, many randomized trials are using outcome-adaptive allocation schemes, which dynamically adjust the allocation ratio in favor of the better performing treatment arm. Advocates of outcome-adaptive allocation contend that it better accommodates clinical equipoise and promotes informed consent, since such trials limit patient-subject exposure to sub-optimal care. In this essay, we argue that this purported ethical advantage of outcome-adaptive allocation does not stand up to careful scrutiny in the setting of two-armed studies and/or early-phase research.

Keywords
Adaptive randomization, ethics, equipoise, therapeutic misconception

Randomization is firmly established as a cornerstone of clinical trial methodology. Yet, the ethics of randomization continues to generate controversy. On its face, random allocation appears to conflict with both a patient's best interest and the ethics of clinical practice. Why would a patient ever agree to receive a randomly chosen treatment?

The concept of clinical equipoise provides an answer and helps to resolve this fundamental tension. Clinical equipoise demands that there exists a state of community uncertainty about the relative therapeutic merits across all arms of a trial. Insofar as this condition holds, all patient-subjects enrolled in a trial can be assured of receiving nothing less than competent medical care. Generally speaking, the randomized controlled trials (RCTs) for serious illnesses are ethically acceptable only where conditions of clinical equipoise hold (for a survey of debates over clinical equipoise, see London).

However, clinical equipoise is silent on the nature of the random allocation scheme. The default, and most efficient, allocation scheme randomizes patients equally (1:1) across all arms of study. But many RCTs use unequal allocation schemes (e.g. 2:1 or 3:1), which assign more patients to the experimental intervention, or outcome-adaptive allocation schemes, which dynamically adjust the allocation ratio in favor of the better performing treatment arm. These alternative allocation schemes are often justified by appeals to increasing patient-subject benefit. That is, by weighting the allocation ratio in favor of the experimental intervention or “better performing” arm, the trial increases the number of patients who receive the presumed superior treatment. Therefore, advocates contend that unequal allocation schemes can better accommodate clinical equipoise, since such trials limit patient-subject exposure to sub-optimal care.

While this argument is appealing, it deserves careful scrutiny. Elsewhere, we argued that unequal allocation ratios in confirmatory trials are ethically problematic, since they fail to minimize patient burden, leverage patients’ therapeutic misestimations, and potentially introduce new internal validity threats. What about trials that use outcome-adaptive allocation?

In what follows, we consider whether criticisms of unequal allocation apply to two-armed studies using adaptive allocation (see, for example, Maki et al. and Garcia-Manero et al.). These studies—which are the most contentious adaptive allocation study types—generally begin by randomizing patients on a 1:1 basis.
to treatment or control. As preliminary patient-response data emerge, the allocation ratio is adjusted in favor of the better performing arm. Once the probability of accepting or rejecting the null hypothesis has dropped below a pre-specified threshold, the trial is closed. After reviewing the ethics of two-armed adaptive allocation trials, we briefly consider the ethics of a less contentious study design: multi-arm adaptive allocation (see, for example, Giles et al.\textsuperscript{11} and Kim et al.\textsuperscript{12}).

**Minimizing patient burden**

A foundational ethical principle is that trials should minimize research burdens. Proponents of outcome-adaptive allocation designs argue that they advance this charge, since they maximize the number of patients receiving superior treatments (and consequently minimize the number receiving inferior ones).\textsuperscript{5,6}

This argument holds little water when considered against the background of drug development. First, the primary objective of early-phase clinical testing is to establish the evidentiary grounds for conducting confirmatory trials. Whereas confirmatory trials employ clinical endpoints (e.g. survival) and large sample sizes, early-phase trials typically use surrogate endpoints (e.g. tumor response) that can be collected in a shorter time frame, and smaller sample sizes. The fact that surrogate benefit and underpowered studies are fickle guides for clinical benefit is indicated by the frequent discordance between effect sizes in phase 3 studies and those in phase 2.\textsuperscript{13} Indeed, the average probability that an intervention reaching phase 3 testing will advance to regulatory licensure—a good proxy for clinical utility—is approximately 50%. This probability is considerably lower in realms like neurology and oncology, where adaptive allocation is often employed; not infrequently, agents that look very promising in early-phase trials actually make patients worse off against comparators in late phase trials. This means that in a best case scenario where one drug outperforms another in an outcome-adaptive study, patients receiving the better performing drug still only have a 50% probability of receiving a drug that is competitive with or better than standard of care.\textsuperscript{14}

Second, new treatments tend to deliver only small improvements over standard ones. Several systematic reviews of RCTs in cancer, for example, show that the odds ratios cluster close to unity.\textsuperscript{15} However, adaptive randomization tends to offer advantages only when effect differences between two interventions are large.\textsuperscript{5} As a consequence, the purported therapeutic advantages of adaptive randomization will rarely be realized, yet their disadvantages in terms of requiring larger sample sizes will often be encountered.

These arguments take on greater force in the context of phase 3 trials. In favor of adaptive allocation in this context is the fact that allocation adjustment will be based not on surrogate, but rather clinical endpoints. This means that better performing arms are much more likely to be clinically superior than they would be in phase 2 studies. This advantage, however, will often—though not always—be nullified by the lengthy period between enrollment and observation of clinical outcomes. These trials must therefore be designed to either recruit very slowly, in order to gain the benefits of adaptive allocation, or else they may have already completed enrollment by the time sufficient outcome evidence is available to adjust the allocation ratio.

Advocates of adaptive allocation might offer two rejoinders. First, they might argue that inefficiencies described above work to the advantage of patients, since larger sample sizes afford more opportunities for patients to enter trials and access superior treatment (e.g. Berry\textsuperscript{16} seems to suggest this point). Leaving aside the fact that this advantage only holds if standard of care proves inferior, this reasoning neglects the economies that arise in dividing labor. Research systems are set up to resolve scientific uncertainties; care systems are established to deliver therapies based on that resolution. Running larger trials because they offer patients better care options asks research systems to shoulder tasks that care systems are designed to carry. It is far better to resolve scientific questions as efficiently as possible, so that on the one hand research resources can be channeled towards resolving other uncertainties, while on the other, the therapeutic “baton” can be expediently passed to systems that are designed to deliver care.

Concerning the efficiency of the research enterprise as a whole, all commentators agree that outcome-adaptive trials are more complicated and expensive to plan and coordinate. While some research centers\textsuperscript{4} may be able to implement outcome-adaptive allocation as a rule, the requisite funding and stakeholder support cannot be assumed to hold across the research enterprise. To be sure, the current unavailability of support does not mitigate the force of proponents who claim that trials ought to be done this way, but it does undermine the claim that we can just “slot in” adaptive allocation into the usual course of research. Absent this support, and in light of the Food and Drug Administration’s (FDA) ambivalence toward the evidence produced in adaptive trials,\textsuperscript{17} outcome-adaptive trials seem more likely to make research less efficient on the whole.

A second argument in defense of two-armed adaptive randomization might be made for trials employing placebo comparators in spite of the existence of standard care options. In such cases, adaptive randomization will indeed decrease the number of patients receiving placebo, should the novel agent show activity. However, we regard such departures from clinical equipoise as unethical—at least in the context of serious illnesses. The appropriate way to resolve the ethics here
is to not use placebo—or to advise patients against entering such trials.

In the end, it seems doubtful that adaptive allocation generally improves risk/benefit for patients. To the contrary, the larger sample sizes translate to more patients enduring more research procedures and extra clinic visits. Since costs scale with sample sizes, it also means more resources are consumed in answering a single research question than would have been the case with a fixed 1:1 design.

**Informed consent**

Another cornerstone principle in medical research is that trials should safeguard the autonomy of research subjects. This can pose challenges in trials involving life-threatening illness because patients can lack a realistic understanding of risk/benefit (therapeutic misestimation) or they can fail to understand the ways that study protocols antagonize treatment objectives (“therapeutic misconception”). Random allocation with fixed proportion is a good example of a protocol-driven element that often frustrates a patient’s treatment objectives. Some have suggested that by using all available information to adjust a protocol, adaptive allocation offers a “partial remedy” for the therapeutic misconception.

For reasons canvassed in the previous section, this argument is not convincing in settings where allocation adjustment is driven by surrogate and short-term outcomes. Indeed, given the marginality of advantage and the clinical uncertainty that should exist after a phase 2 study is ended, adaptive allocation seems instead to invite confusion about risk/benefit rather than dispel it.

But there are also other problems. Studies show that many subjects struggle with the concept of randomization, believing erroneously that investigators will allocate them to the treatment arm they believe is most favorable. Imagine that all patients enter trials incorrectly believing that study teams will allocate them to the best possible treatment. That is, they enter the trial with 100% confidence that they will receive the drug believed to work best by the study team, a highly plausible misconception in the context of a trial with adaptive randomization. Imagine further that as the trial progresses, the study drug trends toward superiority against standard of care and the allocation ratio is adjusted to 4:1. Now, all patients harbor less wrong beliefs on trial entry, since indeed their probability of being allocated to the arm that is believed to be superior is 80% rather than 50%. This is the remedy for therapeutic misconception that proponents of adaptive allocation point out. However, valid adaptive allocation designs dictate allocation of some patients to the flagging arm—in our example, every fifth patient on average. Those patients who also entered the study believing they will be allocated to the better arm are now allocated to a treatment that is actually believed to be inferior. Contrary to dissolving the research–treatment distinction, outcome-adaptive trials simply redistribute its tensions, concentrating misunderstandings for patient-subjects in the flagging arm.

Previously, we argued that using uneven, fixed allocation ratios works against informed consent by leveraging unrealistic patient therapeutic expectations to the advantage of clinical investigators. It also provides no incentive for investigators to work with patients to dispel unrealistic expectations. These arguments almost certainly hold for early-phase trials using adaptive allocation. They probably also hold for most late phase trials, where the advantages of outcome-adaptive allocation are even less compelling.

**Validity**

The overarching aim of trials is to deliver evidence that can support clinical decision-making. First and foremost, this requires that studies support valid inferences about the causal relationship between an intervention and a clinical effect. The very purpose of randomization is to limit the confounding effects of population variability, since randomization all but ensures that patients in all arms of a study are drawn from the same population. It is therefore critical to consider whether adaptive allocation introduces validity threats that are avoidable using other means.

Both critics and proponents alike acknowledge that adaptive allocation introduces some threats to internal validity. In the first place, trials are dynamic entities. Populations and treatments change over the course of testing, new research sites drawing from different populations might be added or protocols might be modified, and investigators become more skilled with delivery techniques. The trial environment is dynamic as well: care standards can change over the course of a study, as can supportive therapies and reagent suppliers. These factors all have the potential to introduce bias into the treatment comparisons when allocation proportions are changed. Moreover, adaptive allocation should have an influence on physician-investigator behavior: Given that the odds of receiving the better treatment will continually improve over the course of the trial, it is in the best interests of patient-subjects (and the physicians advocating on their behalf) to wait and enroll as late as possible.

Of course, not all eligible patient-subjects can afford to wait. For some potential patient-subjects—particularly those who are more seriously ill or treatment-refractory—enrollment in research may be their only chance for an effective treatment. This introduces statistical challenges due to the systematic difference in the early- versus late-enrolling patient populations.
The later-enrolling population—more likely to receive the better treatment—is likely to be healthier. As explicitly acknowledged by the FDA in their guidelines concerning adaptive methods, this predictable time-trend in the study population increases the risk of a biased effect estimate. Block design methods can reduce some of these validity threats. However, these stymie the very advantages of using adaptive randomization.

The multi-arm context

Let us now briefly consider outcome-adaptive allocation in the context of multi-arm trials. These studies evaluate many different research questions at once. For example, they might compare the efficacy or toxicity of three or more different drug-dosage regimens, or they might evaluate response rates of three or more biomarker populations. Again, as patient-subject data become available, the probability of being allocated to a particular treatment arm is adjusted in accordance with the response rate (or sometimes a poorly performing arm may be dropped altogether). In contrast to the above two-arm trial, whose goal is often to reject the null hypothesis and demonstrate efficacy, the goal in these trials is often to identify the most promising regimen or patient population for further investigation.

Given their ability to evaluate many different hypotheses at once, outcome-adaptive multi-arm trials may have advantages over rigid allocation schemes. Gains in efficiency are ethically attractive, and by eliminating flagging arms relatively quickly, multi-armed adaptive studies may also diminish total subject burden associated with testing multiple hypotheses. However, for the same reasons canvassed in the previous sections, multi-armed designs in early-phase settings seem no more likely to enhance therapeutic benefit than two-armed adaptive allocation schemes. Nor are they likely to alleviate moral concerns about therapeutic misconception.

Conclusion

We are dubious of the suggestion that adaptive allocation studies offer generic ethical advantages. At least in the two-arm setting, they seem to worsen total burden by increasing patient exposure to research procedures and to drugs that—even at the end of testing—remain unproven. We are also skeptical that adaptive allocation alleviates the therapeutic misconception. Instead, telling patients that allocation will be adjusted according to evolving (but still highly fallible) evidence seems to invite therapeutic overestimation, while leaving unresolved tensions between the scientific and care objectives in research. Adaptive allocation introduces new sorts of validity threats—although these will vary from study to study. Finally, at least in the two-arm setting, adaptive allocation asks research systems to shoulder some of the weight that care systems are designed to carry. So can we say that adaptive allocation is unethical?

We can contemplate several scenarios where adaptive allocation might have a compelling ethical basis. First, we remain open to the application of multi-armed adaptive allocation studies—not because they increase therapeutic advantage or ameliorate consent problems, but rather because they may enable more efficient resolution of uncertainty where many hypotheses are entertained. Second and related, validity threats need not deter adaptive allocation where outcomes are collected over short periods, or where there are sound reasons to anticipate stability in the trial environment. Third, consent discussions should at least explain that adaptive allocation involves allocating some patients to treatments that accumulating evidence disfavors, and that at least in the case of studies using surrogate endpoints and small sample sizes, even if one arm performs better than another, there remains substantial uncertainty about its actual clinical utility. Indeed, were this disclosed to patients, one suspects purported recruitment advantages for adaptive allocation studies would diminish.

In the end, the principal question that should guide evaluation of novel trial designs is whether they can resolve medical uncertainties with the least patient burden and fewest resources. We favor a rebuttable presumption against adaptive randomization in the setting of two-armed trials.

Acknowledgements

The authors thank referees for helpful comments.

Declaration of conflicting interests

The authors declare that they have no conflicts of interest.

Funding

This work was supported under the PACEOMICS project, funded by Genome Canada, Genome Quebec, Genome Alberta and the Canadian Institute for Health Research (CIHR).

References


