

TRANSLATIONAL TOOLBOX

Phase II Trials in Drug Development and Adaptive Trial Design



Gail A. Van Norman, MD

SUMMARY

Phase II clinical studies represent a critical point in determining drug costs, and phase II is a poor predictor of drug success: >30% of drugs entering phase II studies fail to progress, and >58% of drugs go on to fail in phase III. Adaptive clinical trial design has been proposed as a way to reduce the costs of phase II testing by providing earlier determination of futility and prediction of phase III success, reducing overall phase II and III trial sizes, and shortening overall drug development time. This review examines issues in phase II testing and adaptive trial design. (J Am Coll Cardiol Basic Trans Science 2019;4:428–37) © 2019 The Authors. Published by Elsevier on behalf of the American College of Cardiology Foundation. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Human clinical trials for drug development traditionally progress from small toxicity trials in healthy volunteers (phase I) to proof-of-concept and dose-finding trials in somewhat larger groups of patients with the target condition (phase II) (1), and finally to randomized trials to further delineate clinical efficacy, outcomes, and adverse events in large groups of patients (phase III). The timeframe for passage of a therapeutic agent through clinical testing for Food and Drug Administration (FDA) marketing approval is approximately 12 years (2), with costs now estimated from \$1 billion to \$1.8 billion dollars (2,3).

A recent study by the Biotechnology Innovation Organization of clinical success rates in advancing drugs to market between 2006 and 2015 found that only 9.6% of drugs entering phase I clinical testing will reach the market (4). Following phases II and III, 30.7% and 58.1% of drugs fail, respectively (4). The picture is even worse for cardiovascular (CV) agents; 6.6% of CV drugs entering phase I advance to market, 24% that enter phase II transition to phase III, and 45% that enter phase III result in a new drug

application filing (4). These late phase failure rates probably underestimate failures for first-in-class agents because the reported rates include trials that examine new indications for already-approved drugs and drugs that replicate the mechanism of another successful agent (5).

The flow of innovative agents to the marketplace is slowing significantly as the “low hanging fruit” of therapeutic targets appears to have been substantially harvested (6). Fewer blockbuster drugs and first-in-class drugs are being developed that are both effective in broad patient populations and would: 1) reap enough returns to pay for their own development costs; 2) substantially cover the cost of trials for failed drugs together; and 3) make up for patent expirations on existing drugs (3). New drugs increasingly target fewer indications, are more frequently used for second- and third-line therapy, apply to smaller patient populations, have smaller market interests, and produce a smaller margin in which to recoup costs. Drug companies are answering these challenges by focusing on therapies that are most likely to reach market approval rapidly, are less

From the Department of Anesthesiology and Pain Medicine, University of Washington, Seattle, Washington. Dr. Van Norman has received financial support from the *Journal of the American College of Cardiology*.

The author attests she is in compliance with human studies committees and animal welfare regulations of the author's institutions and Food and Drug Administration guidelines, including patient consent where appropriate. For more information, visit the JACC: Basic to Translational Science [author instructions page](#).

Manuscript received February 11, 2019; accepted February 14, 2019.

subject to pricing pressure once the market is reached, are less expensive to develop, and therefore, have higher potential to improve return on investment (7,8).

There are worries that drugs are being discarded too soon in clinical testing, either due to failure of trials to identify the right target patient population or due to commercial concerns rather than clinical concerns. Attention is turning to understanding why so few initially promising drugs fail to pass clinical testing and to reducing the time it takes for drugs that will ultimately be successful to pass through phases II and III. Phase II testing plays a pivotal role in drug development costs because so many agents “die” in phase II, and because successful completion of phase II is a poor predictor of whether a drug will complete phase III, the most expensive of clinical trials. As the CEO of GlaxoSmithKline pointed out, “if you stop failing so often, you massively reduce the cost of drug development” (7). This review examines some aspects of the phase II problem and discusses adaptive trials and statistical challenges.

WHY DRUGS FAIL IN CLINICAL TRIALS

Phase II represents the first time in which a drug is tested in actual patients, ranging from 50 to 200 patients in most heart failure (HF) studies (9). Failures in phase II testing overall usually occur because: 1) previously unknown toxic side effects occur (50%); 2) the trials show insufficient efficacy to treat the medical condition being tested (30%); or 3) commercial viability looks poor (15%) (10). For CV drugs, 44% of late trial failures are due to poor efficacy and 24% are due to safety concerns (11). Phase II trials face many challenges due to small sample size and choice of study design. In addition, the relatively short duration of phase II trials makes it difficult to identify long-term side effects and outcomes.

BIOMARKERS AS SURROGATE ENDPOINTS. Single-arm phase II (SA-II) studies are usually insufficient to test long-term outcomes because clear indications of the success or failure of a treatment can take months or years, and would extend trial times and cost (12). Phase II studies instead increasingly rely on surrogate clinical or biochemical markers to provide “interim” data about safety and efficacy, allowing faster drug approval conditioned on continued post-marketing safety and efficacy studies (9). In HF phase II trials, increasing emphasis has been placed on clinical biomarkers such as functional capacity, left ventricular ejection fraction, and chemical biomarkers of HF (e.g., B-type natriuretic peptide levels). However, validating that a given biomarker is an appropriate

surrogate study endpoint is complex and requires compelling evidence (13).

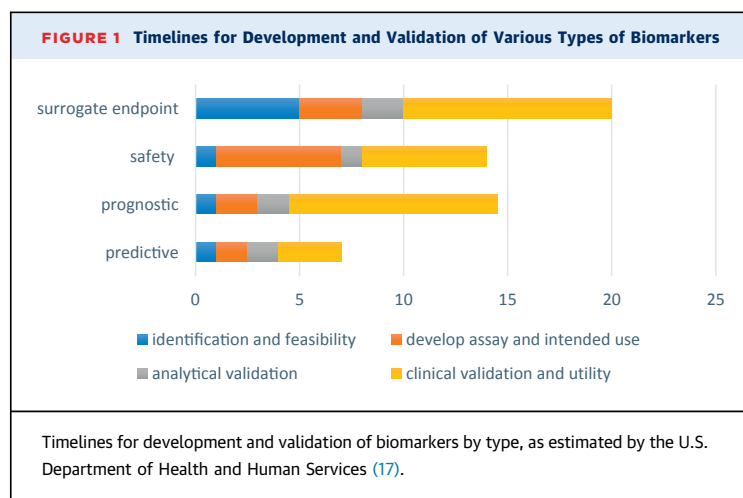
The use of biomarkers has been encouraged by the FDA, which has now established means by which to qualify biomarkers for use in drug development, but there is a lack of validated biomarkers at this time (14–16). Commercial development and validation of biomarkers is slow, because development of robust and meaningful biomarkers is both time-consuming and extremely expensive (Figures 1 and 2) (17). Biomarkers can also fail in development, just as drugs fail in clinical trials. As with drugs, biomarker failures carry both scientific and commercial consequences—if the risks of failure are too high, research to yield successful biomarkers and therapeutics simply will not be undertaken or may be terminated prematurely.

It is unclear how well biomarkers accurately predict positive patient (and trial) outcomes. Wong et al. (18) examined the probability of success of various clinical trial phases between 2000 and 2015 by examining 185,994 unique trials from the Informa Pharma Intelligence’s Trialtrove and Pharmaprojects databases. During their study period, probability of success for all indications declined until 2013, but then began to rise. The increase occurred as the use of biomarkers rose, although the authors admitted that there were many other potential causes for this trend. Even so, no significant difference was found in probability of success when phase II studies in which biomarkers were used for patient selection and/or as an indicator of therapeutic efficacy or toxicity were compared with trials that did not use biomarkers (18,19).

Examples of remarkable biomarker failures in CV therapies abound. Floseqinan increased exercise capacity of HF patients in phase II trials, but increased patient mortality in phase III (20,21). The phase III VISTA-16 (Vascular Inflammation Suppression to Treat Acute Coronary Syndrome for 16 Weeks) trial showed promising biomarker response in early phase trials, but had to be terminated early due to higher mortality in phase III (22,23). Darapladib failed to show clinical efficacy in a trial involving >15,000 patients, despite the biomarker (high-density lipoprotein) response (24). Despite evidence of reduction in B-type natriuretic peptide with aliskiren, the ASTRONAUT (Effect of Aliskiren on Postdischarge Mortality and Heart Failure Readmissions Among Patients Hospitalized for Heart Failure) trial failed to show improvement in major clinical outcomes (25,26). Tredaptive, a European therapy that combined niacin

ABBREVIATIONS AND ACRONYMS

AD	= adaptive (trial) design
CV	= cardiovascular
EMA	= European Medicines Agency
FDA	= Food and Drug Administration
FDR	= false discovery rate
HF	= heart failure
RP-II	= randomized phase II studies
SA-II	= single-arm phase II studies



and laropiprant, which was demonstrated to raise high-density lipoprotein, not only found no clinical effect on CV outcomes in a trial for U.S. approval that involved >25,000 patients, but had significant toxic side effects. The drug was not approved in the United States, and the drug maker began warning overseas doctors to stop prescribing it (27,28).

PHASE II AND PHASE III ATTRITION. Once a drug or biological has passed phase II efficacy trials, less than one-half of them show sufficient efficacy in phase III to make it to market. There are many reasons this might be so; phase II and III studies often examine different endpoints. A phase II study may examine surrogate endpoints for disease, but longer term outcomes and toxicities are usually the focus of most phase III studies. Phase III studies are performed in a less homogeneous patient population than phase II trials, to better reflect real-world results. In addition, failures of phase II studies to predict phase III success may be in part due to statistical phenomena (5).

The false discovery rate. The false discovery rate (FDR) is a concept that arises from relatively simple statistical characteristics of type I (false positive) and type II (false negative) errors in clinical testing. The FDR is the ratio of false positive results to total positive results (29).

Consider a study finding in favor of a drug effect for which the *p* value is 0.05, a common significance level set in clinical trials. The *p* value is not the probability of mistakenly rejecting the null hypothesis (i.e., the odds of wrongfully attributing an effect to the drug when there is none), but rather is the probability that a random sampling error could lead to the difference that is observed. A low *p* value indicates that the data are unlikely to have occurred in the presence of a true null, but the *p* value does not

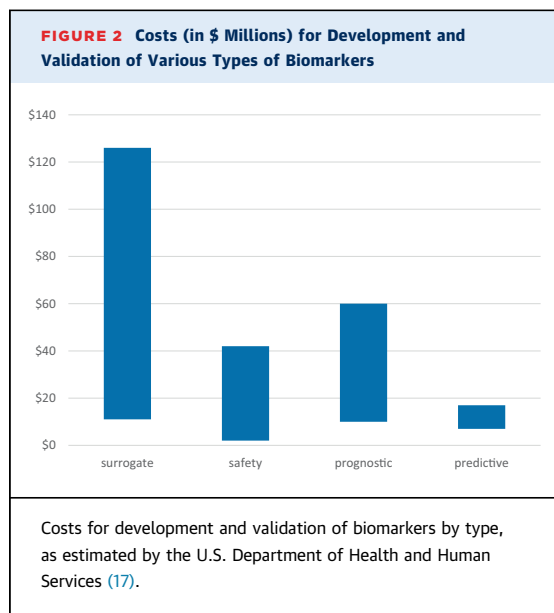
evaluate: 1) whether the null is actually true, that is, there actually is no drug effect, but the study sample was unusual or flawed; and 2) whether the null hypothesis is false. In the case of our drug, the correct conclusion is that there is a 5% chance that the apparent effect shown in the study is due to a random sampling error, regardless of whether other features of the study are flawed. The odds of incorrectly rejecting the null hypothesis (and therefore incorrectly concluding the drug has an effect when it does not) in a study with a *p* value of 0.05 is at least to 23% and typically closer to 50%. Even at a *p* value of 0.01, the odds of incorrectly deciding the drug has an effect are still 7% to 15% (30).

The multiplicity problem. Another issue in statistical analysis is the “multiplicity problem”: the odds of a false positive result rise as the number of tests rises (31). If a test has a 5% significance level (*p* = 0.05), and the test is run only 20 times, the odds of seeing a false positive result are >64% (32).^{*} Repeating the test multiple times increases the FDR by increasing the number of false positive tests (Table 1). Put another way, repeat a test enough times, and you are guaranteed to find an effect, but the effect you see is increasingly likely to be a false positive as the number of repetitions grows. Most phase II clinical studies have enrollments of 100 to 200 patients and multiple tests; therefore, the odds of multiple false positive results are high. An obvious answer seems to be to decrease the accepted *p* value, but then the number of tests must be increased to reach significance. Controlling for type I errors (false positives) furthermore causes the complementary errors (type II, false negatives) to increase. The study starts to risk not identifying the true effects that are present, and therefore, misidentifies a drug as being ineffective when it is actually effective, which is an equally undesirable false conclusion.

Challenges in clinical design. Traditionally, phase II clinical trials have been SA-II studies, many of which use historical controls. Selection bias is common, because these trials are often carried out in a single institution or a small group of academic institutions where the patient population differs

^{*}This number is arrived at using the binomial formula to calculate the probability of having a zero (0) FDR at a 5% (*p* = 0.05) significance level for 20 tests. The odds of a zero FDR is approximately 36%, and therefore, the odds of a false positive is 64%.

$$P(0) = \frac{20!}{(20-0)!0!}(0.05)^0 \times (0.95)^{20-0} = 0.358$$



significantly from the at-large phase III target population. Multicenter community-based studies partially overcome this issue.

The historical controls are also problematic. They do not consider the fluidity of patient outcomes over time due to changes in supportive care, changes in the quality of diagnostic “trial entry” studies, or changes in other standards of care over time. They also do not account for interim advances in understanding of disease biology that permit subclassifications of diseases (which might respond optimally to different therapies), as well as advances in surgical, radiological, and other nondrug interventions that are also constantly evolving. Even the criteria used to grade disease response are evolving, and this adds complexities to interpreting time-related endpoints (33).

SA-II trials using historical controls do not account for the heterogeneity of real-world patients for whom the treatment will be targeted and assume that the study patients are identical to historical patients. The true success rates of a drug tested in SA-II design against historical controls can only be incompletely known. This is because even a small variance (e.g., 5%) of the actual control success rate away from the historical success rate rockets the false positive rate in single-arm studies by 2- to 3-fold (34). Thus, the drug may falsely appear in a phase II study to demonstrate the level of efficacy that is required to survive phase III studies. Trying to compensate for this issue by increasing the sample size in the SA-II study actually inflates the false positive error rate further, by as much as 50% (34).

TABLE 1 Adaptive Designs*

Adaptive Design	Description
Dose-finding	After interim analysis, a randomized trial with multiple dosing arms assigns more patients to dose groups of higher interest.
Hypothesis	After interim analysis, the study hypothesis is altered (e.g., a pre-specified swap of primary and secondary endpoints).
Sequential or group sequential	After interim analysis, adaptations include pre-specified options of changing sample size, modification of existing treatment arms, elimination or addition of treatment arms, changes in endpoints, changes in randomization schedules.
Randomization	Randomization is adjusted after interim analysis so that patients enrolled later in the study have a higher probability of assignment to a treatment arm that appeared successful earlier.
Seamless phase II/III	Study moves from phase I to phase II without stopping the patient enrollment process.
Treatment switching	Investigator is allowed to switch patients to a different treatment arm based on lack of efficacy, disease progression, or safety issues.
Biomarker adaptive	Interim analysis of treatment responses of biomarkers allows pre-specified adaptations to trial design.
Pick-the-winner and drop-the-loser designs	After interim analysis, treatment arms are modified, added, or eliminated.
Sample size re-estimation	Interim analysis allow sample size adjustment or re-estimation.
Multiple adaptive	Multiple adaptive design characteristics applied in a single study.

*See references 36,42,73.

One way to theoretically manage this problem is to actually include a control arm of contemporary patients in the phase II study in place of historical controls (a so-called randomized phase [RP-II] study) (35). RP-II trials guarantee better matching of patients and control patients, and that similar assessment methods and supportive medical care occur in the trial arms contemporaneously. Although RP-II trials increase the size, complexity, and costs of phase II studies, they mitigate the problem of inflating the false positive rate (34,35). However, the increased size randomization makes RP-II studies less suitable as screening trials. RP-II trials can require 4-fold more patients than SA-II trials (33).

The concept of an RP-II trial also raises the question: if a RP-II trial generates sufficiently robust clinical data, why could it not be used alone to achieve full drug approval and avoid a phase III trial altogether? If sufficiently compelling data are obtained in RP-II, would not the performance of a phase III trial, which includes possibly randomizing patients to a nontreatment arm, be unethical? A definitive RP-

TABLE 2 Magnification of Type I Error With Multiple Tests*

No. of Tests	Type I Error (approximate %)
1	5
2	10
3	14
10	40
20	64

For $p = 0.05$, the odds of type I error (false identification of an effect where none exists) as the number of tests increases. *See reference 31.

II trial might be appropriate for approval of a drug that treats a sufficiently rare disease for which a conventional phase III trial might be difficult or impossible to conduct. An important challenge rests in determining what would define a “sufficiently robust” data set for a given drug to bypass phase III. The concept is not without precedent. The FDA has based initial approval of a few oncology drugs on the outcome of SA-II trials (33).

RP-II trials are larger and more complex than traditional phase II trials; more than one-quarter have been shown to require major amendments during the trial (33). This has led to calls for the development and acceptance of “seamless” trials—strategies that in effect allow a phase II trial to fold seamlessly into a phase III trial by combining elements and data analysis of both in a single innovative trial design (36,37).

INNOVATIVE TRIAL DESIGN

In March 2004, the FDA issued a report recognizing that the approval of innovative medical therapies had slowed over the preceding years (38). The estimated phase II failure rate in 2006 was 50% versus 20% 10 years earlier (39).

A result of the critical pathway initiative of the FDA was increased interest in innovative trial designs. In December 2016, the U.S. Congress passed the 21st Century Cures Act, allotting \$500 million to the FDA to establish an “innovation account” for National Institutes of Health funding to speed regulatory approval of medical therapies (40). Since then, the FDA has devoted efforts to exploring modern trial design and evidence development, including the use of adaptive trial designs (ADs) and real-world evidence (41).

A number of different designs fall under the category of innovative trial design, all of which allow interim data analysis and modification of the trial (Table 2) (42). Examples include enrichment trials, adaptive trials, and flexible trials.

ENRICHMENT TRIALS. Enrichment trials allow patient enrollment by clinical criteria, and each is then assayed for a pre-specified drug target (1). After that, several different trial strategies can be pursued: 1) randomize all enrolled patients and analyze the patients carrying the target in a subgroup analysis; 2) continue the trial with patients who only express the target; or 3) split the trial into 2 groups (those with the target and those without) and randomize and analyze each group separately. Enrichment trials may hasten to market therapeutics that benefit a specific patient subpopulation rather than a more heterogeneous population with a broad disease designation, but they depend in part on knowing in advance what factors may contribute to disease progression, and then constructing trial populations that contain the various factors. A downside of enrichment trials are that they identify agents that work in enriched populations but may show less efficacy in unselected populations. Such trials may also inadvertently exclude patient subpopulations that are responsive to the drug, because a characteristic common to that subpopulation was not recognized in trial design and patient selection. A therapy that might be effective in an untested patient subpopulation would then be inadvertently discarded from further development for lack of efficacy (1).

ADAPTIVE TRIALS. ADs have been discussed for 30 years (43–45). The FDA defines an AD as a clinical trial design that allows for prospectively planned modifications to ≥ 1 aspects of the design based on accumulating data from subjects in the trial (46,47).

In ADs, the goal is to learn from accumulating data in the trial and apply what is learned as quickly as possible in a prospectively specified way during the trial itself to hone flexible aspects of the study while it is still ongoing. ADs can be classified as prospective, continuously adjusted or concurrent (ad hoc), and retrospective (9,42,48). In prospective ADs, there is a pre-specified protocol to alter aspects of the study, such as size, follow-up period, and clinical endpoints following interim data analysis. This might lead to early termination of a study based on futility or unacceptable toxicity, or, alternatively, might require a change in sample size. A platform study or master protocol design is a type of adaptive trial in which multiple treatment arms are simultaneous studies, and interim analysis allows early termination of various arms due to futility or lack of efficacy (49). Concurrent or ad hoc study designs allow flexibility to alter multiple parameters in a study in a pre-specified way based on interim results. In ad hoc design, investigators are allowed to hone their hypothesis

based on interim results and re-steer the study accordingly. Both retrospective and prospective data following changes are used in analysis. Retrospective ADs allow the investigators to change the primary study endpoint or analysis methodology in a pre-specified way after a study is closed.

ADs must be approached cautiously. Seamless progression of an AD study from phase II proof-of-concept and dose-finding stages into phase III studies of efficacy and safety in large populations implies that an efficacious outcome was seen in phase II. This “unblinding” can bias both investigators and caregivers going forward. Also, because the patient population is adjusting throughout the trial, patients enrolled earlier in the trial are likely to show a different magnitude of outcomes than those enrolled later, and this effect must be carefully accounted for in a more complex statistical analysis.

Adaptation in ADs is a design feature and not a cure for poor trial design and inadequate planning (50,51). The PhRMA Working Group on Adaptive Design in Clinical Drug Development identified statistical, logistical, and procedural issues that can arise in ADs and made recommendations for meeting these challenges (50). For ADs to take full advantage of the efficiencies they supposedly offer, they recommended that: 1) study endpoints in ADs should have short follow-up time relative to the overall duration of the trial; 2) data accrual should occur in real time or rapidly, making electronic data collection a priority; and 3) use of databases must incorporate case-by-case decisions regarding how well the data must be “cleaned” (a potentially time-consuming enterprise) versus making adaptation decisions based on all available data. They and other authors also emphasize that data monitoring committees that carry out interim analyses must be constituted in such a way as to minimize bias, including commercial bias (50,52). This can be challenging, because such committees usually need representation from the commercial sponsor itself, to provide input on practical aspects of trial design and commercialization. They strongly recommend that such data monitoring committees be “firewalled” from project personnel, that sponsor representatives on the committee be isolated from all trial activities, and that sponsor access to interim data be minimized.

ADs are slowly achieving increasing regulatory acceptance. Both the European Medicines Agency (EMA) and the FDA have published papers and guidance documents addressing the limitations of ADs in the regulatory context (47,53,54). Regulatory guidances all agree that strict control of the type I error rate is a regulatory prerequisite for acceptance

of a clinical trial (55). In a review of >5 years of EMA and FDA advice letters regarding proposed AD phase II or phase II/III studies, 20% of the 59 studies were not accepted, and the most frequently cited concerns raised by the agencies were, not unexpectedly, insufficient justifications for the adaptation, inadequate type I error rate control, and study bias (53).

FLEXIBLE DESIGN TRIALS

The term flexible design (FD) is not entirely synonymous with adaptive design, and there is some confusion of these terms in the literature (56-58). FDs are a subset of ADs that allows both planned and unplanned changes. Flexible aspects of such trials might include inclusion and/or exclusion criteria, sample size, randomization ratios, analytic methods, drug dose, treatment schedule, and endpoints. For example, if the incidence of a primary endpoint is much lower than expected, a FD would allow a mid-trial increase of sample size. The primary endpoint itself could be altered by including additional outcomes in a composite primary outcome. Protocol changes might be made based on unblinded interim results.

AD is complex, must be undertaken carefully to minimize bias, and tends to draw greater regulatory scrutiny. In 2006, the FDA strongly recommended ADs to address the decline in innovative medical products being submitted for approval (47). FDs have been criticized as being subject to both more perceived and more actual bias, and present more complex challenges to regulators (58,59). However, such designs could theoretically speed study efficiency, reduce the number of subjects needed (thus, saving time and money), and expose fewer patients to ineffective or even harmful treatment by allowing intra-trial adjustment of pre-determined parameters.

ARE AD TRIALS MORE ETHICAL? Randomization is believed to enhance the validity of clinical research. However, many researchers now question whether traditional randomization is actually ethical. Asking a patient with a serious medical condition to submit to randomization is only ethical if the investigator is truly uncertain about the efficacy of the 2 study arms. Zelen (60) originated the idea of the “play the winner” rule, an example of an AD in which if a patient in a comparative study responded to treatment, the next patient would be assigned to that study arm. Zelen pointed out the ethical implications of this method: 1) patients would tend to be increasingly assigned to the successful arm during the course of the trial; 2) fewer patients would theoretically receive

ineffective or less effective treatment; and 3) trial duration would theoretically be lessened, thus minimizing the time to get effective drugs to market.

Despite arguments that adaptive trials are more ethical (42,60), strong arguments are advanced that they are not. It is not clear that actual patient burden is reduced, because contemporary adaptive trials must often rely on intermediate markers of response that will not be validated or repudiated until late in phase II or even later in phase III, if at all. Unless phase II studies actually lead to faster approval of more effective therapies (a hypothesis yet to be proven), adaptive trials may not improve patient burdens (61). At this point, no studies have shown that adaptive trials improve patient outcomes overall.

Ethical controversies regarding adaptive trials are numerous, and include problems of informed consent, the importance of maintaining the validity of research, and many other issues (49,62-69). However, determining whether adaptive phase II trials accomplish their goals by actually improving patient outcomes and/or shortening trial duration and costs are key to answering the ethical concerns.

DO ADs ACCOMPLISH THEIR GOALS? The goals of ADs include reduction of overall costs of drug development through several mechanisms: 1) potentially shorter phase II trial durations due to early termination for futility or efficacy; 2) elimination of phase III trials through adequate proof of efficacy and safety via RP-II trials; 3) less expensive studies due to smaller total sample sizes for phase II and phase III studies; and 4) more accurate prediction of success in phase III trials. All of this must be accomplished while holding type I error rates at bay. But do ADs accomplish these goals?

Statistical issues in ADs are complex. Tsong et al. (70) demonstrated that dropping a treatment arm could increase the type I error rate, unless certain precautions are taken in design. Hung et al. (57) pointed out problems in interpreting noninferiority might be problematic in ADs that compare a new treatment with established treatments (nonplacebo comparison trials).

In a review of 60 AD medical trials published between 1989 and 2004, Bauer and Einfalt (71) found that 60% ended after interim analysis, with 20% stopped for futility. However, it is difficult to know if that is actually an improvement from non-AD trials; for example, only 24% of CV drugs pass traditional phase II testing. Moreover, 44% of CV drugs fail in phase III trials due to lack of efficacy (i.e., “futility”). The authors did not detail how many of the trials

were stopped because of concerns over commercial viability (71).

Lin et al. (72) from the Center for Biologics Evaluation and Research at the FDA, reviewed AD features of investigational device exemptions applications to the FDA between 2008 and 2013. ADs consisted of only approximately 11% of all studies. The number of AD trials in submissions fluctuated during that period between 10 and 40 per year, which did not represent a clear increase in applications using ADs despite release of FDA draft guidance. Multiple authors agreed that real or even perceived excess regulatory scrutiny might be discouraging commercial sponsors from embracing nontraditional ADs (52). Lin et al. (72) pointed out several problematic features of AD studies that might adversely affect costs. AD studies might not necessarily be of short duration: they must be long enough to allow the adaptation called for in the design. Complex ADs might be neither time- nor cost-efficient for sponsors because they often require detailed justification, extensive simulation studies, and multiple review cycles, all of which can significantly delay the start of a study. Studies stopped too early for success might not have accumulated sufficient safety information, which is an important issue, because for early phase studies, regulators are more concerned with safety than efficacy. International trials pose special problems. Local regulators may require significant in-country trial data; substantial differences in standard of care between countries and regions and substantial differences in populations may exist.

There is conflicting evidence about whether ADs shorten study duration or enroll fewer patients. One survey published in 2018 examined 245 AD clinical trials between 2012 and 2015 (46). ADs in this study resulted in shorter study durations and smaller numbers of subjects. More than 80% of the adaptive phase II trials resulted in early termination at interim analysis, with more than one-half ending up with fewer randomized patients than initially planned (46). However, early termination of some trials occurred for reasons unrelated to study findings, such as poor study enrollment or commercial considerations.

Another analysis included 31 adaptive phase II or III clinical trials at the EMA (55). In contrast to the study by Sato et al. (46), planned and actual study sizes ended up being similar, and ADs did not result in significantly smaller numbers of subjects. Whether ADs led to shorter durations of study was not clear. Only 4 of 23 completed trials (17.4%) were terminated early, 2 of which were stopped due to difficulty recruiting subjects, and 2 of which (8.7%) were stopped according to a pre-planned stopping-for-

futility analysis—a much smaller percentage of trials than those in the Sato et al. (46) study. It is unknown if this means that more drugs were successful in passing to market, or if more drugs were passed through, only to fail in later phase trials.

Bothwell et al. (73) reviewed 142 AD clinical trials. In 9% of cases, adaptive trials were used for FDA product approval consideration and in 12% for EMA product approval consideration. In 8% and 5% of cases, ADs were the final or pivotal trials used for FDA and EMA approvals, respectively. Review times for the FDA and EMA for adaptive trials was a median of 12.2 and 14 months, respectively, which exceeded estimates of review times of non-AD studies by the FDA and EMA by 6 to 7 weeks. Frequently cited problems in reviews were lack of sufficient statistical power, risk of ineffectively evaluating doses, risk of falsely detecting treatment effects (type I errors), and inadequate blinding. Some sample sizes were deemed at times to be too small to draw robust conclusions by both the FDA and EMA. Both agencies cited inadequate sample sizes in adaptive trials to gather sufficient subpopulation effects, such as outcomes on race and sex. Lengthy review correspondence was noted at both agencies in many adaptive trials. At the FDA, 9% (n = 13) and at the EMA 7% (n = 10) of AD trials tested drugs that were later approved at least in part, but not solely through reliance on the adaptive trial data. However, most of these trials involved orphan drugs at both these agencies (9 and 6, respectively), in which the challenges of traditional trial design and population sizes might have mitigated some of the sample size concerns of the ADs.

The question of whether the success of AD trials in phase II better predict that all-important success in subsequent phase III trials is entirely unanswered. One 2016 review of 143 adaptive design studies found

that 30% and 50% of early terminations of phase II and III AD studies, respectively, were due to findings of futility, but did not provide information of what preceded these failures (i.e., was a termination in an AD phase III preceded by a successful AD in phase II?) (74). To date, the average costs of bringing a drug to market via AD trials relative to traditional trials are also unknown.

SUMMARY

Adaptive trials are a proposed way to shorten clinical trial phases, reduce the number of patients needed for enrollment, better predict later drug success, and reduce drug development costs. Criticisms of ADs have included increased risks of falsely detecting treatment effects (type I errors), premature dismissal of promising therapies as falsely ineffective (type II errors), statistical challenges and bias, and operational bias. Use of ADs has been limited due to lack of inadequate information regarding completed adaptive trials, a lack of practical understanding of how to implement an adaptive trial, and worries about excessive regulatory scrutiny and nonapproval. To date, analysis of AD trials gives conflicting results with regard to their effects on study size and duration. Data regarding whether phase II ADs permit more accurate prediction of successful completion of phase III and whether ADs reduce overall costs of drug development are needed.

ADDRESS FOR CORRESPONDENCE: Dr. Gail A. Van Norman, Department of Anesthesiology and Pain Medicine, University of Washington, 2141 8th Avenue West, Seattle, Washington 98119. E-mail: lbsparrow@yahoo.com.

REFERENCES

1. Cummings JL. Optimizing phase II of drug development for disease-modifying compounds. *Alzheimers Dement* 2008;4:1 Suppl 1:S15-20.
2. Paul SM, Myteka DS, Dunwiddie CT, et al. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature Rev Drug Discov* 2010;9:203-14.
3. Van Norman G. Drugs, Devices and the FDA: Part 1. An overview of approval processes for drugs. *J Am Coll Cardiol Basic Transl Sci* 2016;1:170-9.
4. Thomas DW, Burns J, Audette J, et al. Clinical development success rates 2006-2015. Biotechnology Innovation Organization, Washington DC. June 2016. Available at: <https://www.bio.org/sites/default/files/Clinical%20Development%20Success%20Rates%202006-2015%20-%20BIO,%20Biomedtracker,%20Amplion%202016.pdf>. Accessed December 7, 2018.
5. Grainger D. Why too many clinical trials fail—and a simple solution that could increase returns on Pharma R&D. *Pharma and Health Care*. Forbes. January 29, 2015. Available at: <https://www.forbes.com/sites/davidgrainger/2015/01/29/why-too-many-clinical-trials-fail-and-a-simple-solution-that-could-increase-returns-on-pharma-rd/#46e76ff1db8b>. Accessed December 23, 2018.
6. Van Norman G. Overcoming the declining trends in innovation and investment in cardiovascular therapies: beyond Eroom's law. *J Am Coll Cardiol Basic Transl Sci* 2017:613-25.
7. Huss R. The high price of failed clinical trials: time to rethink the model. *Clinical Leader* October 3, 2016. Available at: <https://www.clinicalleader.com/doc/the-high-price-of-failed-clinical-trials-time-to-rethink-the-model-0001>. Accessed December 7, 2018.
8. Honig P, Huang SM. Intelligent pharmaceuticals: beyond the tipping point. *Clin Pharm Ther* 2014;94:455-9.
9. Lavine KJ, Mann D. Rethinking phase II clinical trial design in heart failure. *Clin Investig (London)* 2013;3:57-68.
10. Tufts Center for the Study of Drug Development. Causes of clinical failures vary widely by therapeutic class, phase of study. Tufts CSDD study assessed compounds entering clinical testing in 2000-09. *Tufts CSDD Impact Rep* 2013; 15(5).

51. Weichung JS. Plan to be flexible: a commentary on adaptive designs. *Biom J* 2006;48:656–9.
52. Mehta CR, Jemai Y. A consultant's perspective on the regulatory hurdles to adaptive trials. *Biom J* 2006;48:604–8.
53. Elsaßer A, Regnstrom J, Vetter T. Adaptive clinical trial designs for European marketing authorization: a survey of scientific advice letters from the European Medicines Agency. *Trials* 2014;15:383.
54. European Medicines Agency. Reflection paper on methodological issues in confirmatory clinical trials for drugs and biologics planned with an adaptive design. 2007. Available at: https://www.ema.europa.eu/en/documents/scientific-guideline/reflection-paper-methodological-issues-confirmatory-clinical-trials-planned-adaptive-design_en.pdf. Accessed March 18, 2019.
55. Collignon O, Koenig F, Koch A, et al. Adaptive designs in clinical trials: from scientific advice to marketing authorisation to the European Medicine Agency. *Trials* 2018;19:642.
56. Gallo P, Chuang-Stein C, Dragalin V, et al. Rejoinder. *J Biopharm Stats* 2006;16:311–2.
57. Hung HMJ, O'Neill RT, Want SJ, Lawrence J. A regulatory view on adaptive/flexible clinical trial design. *Biom J* 2006;48:565–73.
58. Jennison C, Turnbull BW. Discussion of “executive summary of the PhRMA working group on adaptive designs in clinical drug development. *J Biopharm Stats* 2006;16:293–8.
59. Gould AL. How practical are adaptive designs likely to be for confirmatory trials? *Biom J* 2006;48:644–9.
60. Zelen M. Play the winner rule and the controlled clinical trial. *J Am Stat Assoc* 1969;64:131–46.
61. Hey SP, Kimmelman J. Are outcome-adaptive allocations trials ethical? *Clin Trials* 2015;12:102–6.
62. Hey SP, Kimmelman J. Rejoinder. *Clin Trials* 2015;12:125–7.
63. Korn EL, Friedlin B. Commentary on Hey and Kimmelman. *Clin Trials* 2015;12:122–4.
64. Buyse M. Commentary on Hey and Kimmelman. *Clin Trials* 2015;12:119–21.
65. Saxman SB. Commentary on Hey and Kimmelman. *Clin Trials* 2015;12:113–5.
66. Berry DA. Commentary on Hey and Kimmelman. *Clin Trials* 2015;12:107–9.
67. Li Y, Mick R, Heitjan DF. Suspension of accrual in phase II cancer clinical trials. *Clin Trials* 2015;12:128–38.
68. Joffe S, Ellenber SS. Commentary on Hey and Kimmelman. *Clin Trials* 2015;12:16–8.
69. Begg CB. Ethical concerns about adaptive randomization. *Clin Trials* 2015;12:101.
70. Tsong Y, Hung HMJ, Wong SJ, Dui L, Nuri WA. Dropping a treatment arm in clinical trial with multiple arms. In: *Proceedings of Biopharmaceutical Section*. Alexandria, VA: American Statistics Association, 1997:58–63.
71. Bauer P, Enfalt J. Application of adaptive designs—a review. *Biom J* 2006;48:493–506.
72. Lin M, Lee S, Zhen B, et al. CBER's experience with adaptive design clinical trials. *Therap Innov Reg Sci* 2016;50:195–203.
73. Bothwell L. Adaptive design clinical trials; a review of the literature and clinicalTrials.gov. *BMJ Open* 2018;8:e018320.
74. Hatfield I, Allison A, Flight L, Julious SA, Munyaradzi D. Adaptive designs undertaken in clinical research: a review of registered clinical trials. *Trials* 2016;17:150.

KEY WORDS adaptive design, biomarker studies, false discovery rate, multiplicity problem, phase II clinical trials