The Journal of Prevention of Alzheimer's Disease 12 (2025) 100392

Contents lists available at ScienceDirect

The Journal of Prevention of Alzheimer's Disease

journal homepage: www.elsevier.com/locate/tjpad

Special Article

Statistical innovations in clinical trial design with a focus on drug combinations, factorials, and other multiple therapy issues

Donald A. Berry

Division of Discovery Science, Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

ARTICLE INFO

Keywords:
Factorial designs for drug combinations
Adaptive factorial designs
Innovative bayesian designs
Efficient adaptive dose-finding trials
Simulations in clinical trial design

ABSTRACT

Statistical methods in clinical research tend to become entrenched. Innovations threaten the status quo. The "right way" becomes frozen in lore. This is so even when the "right way" is not best. "Statistical significance" and the associated requirement of "high power" is an example. This attitude is an impediment to efficient design. Willingness to address some design issues with moderate power enables building highly informative and highly efficient clinical trials. This article considers several types of clinical trials, including dose-finding, combinations, and factorial designs. Bayesian adaptive methods are used to show that trials can be made more efficient and more informative. Surprisingly, the approach is consistent with many attitudes of the widely regarded "Father of Modern Statistics," R.A. Fisher. Fisher was anti-Bayesian in rejecting its subjective interpretations. But Fisher and Bayes come to the same conclusion in many applied matters. Fisher invented factorial design. Its principal attraction for him was enabling addressing two or more questions with a single experiment. He complained about attitudes that hindered progress: "No aphorism is more frequently repeated in connection with field trials [and clinical trials], than that we must ask Nature few questions, or, ideally, one question at a time… this view is wholly mistaken." Fisher's primary analysis required modeling and making assumptions. For example, his first analysis in a factorial setting assumed no interactions among the factors. He investigated possibilities of interactions but he did not see the need for doing so with high power.

1. Introduction

This article addresses recent modifications in statistical thinking regarding designs of clinical trials. The focus is on Alzheimer's disease (AD) but informed by learnings in other diseases. The final analyses can be Bayesian, frequentist, or a combination of the two.

Bayesian approaches are being increasingly used in clinical trials. They are enabling revolutionary modifications in the building and running of trials. Understanding today's clinical research requires some understanding of the differences in statistical philosophy and some understanding of the evolution of both approaches in influencing each other. The U.S. FDA has played a critical role in understanding innovation and in facilitating the introduction of Bayesian ideas in drug development. Health authorities outside of the U.S. are not opposed to such innovations but they have not taken leadership roles in effecting and perfecting them.

Building a trial that adapts to accruing information is natural within the Bayesian approach. Bayes' rule updates knowledge from one observation to the next. It is ideal for predicting future results, given the currently available results and the future course of the trial's prospective design. The frequentist approach can do neither.

The traditional statistical regulatory requirement of controlling type I error rate, usually to no greater than 2.5 %, is not part of the history or tradition of a Bayesian approach. But controlling type I error is always possible in a Bayesian trial by adjusting its design or analyses using simulations and trial and error. The FDA specifically addresses this approach for complicated trials with many types of adaptations in their Complex Innovative Designs (CID) initiative[1]. Programming clinical trial simulations requires detailed descriptions of the prospective design.

As a guideline, an automaton must be able to conduct the trial. Actually, an automaton does conduct the trial, electronically, by simulating it millions of times before it is actually run. When testing a null hypothesis of a treatment's effect, the proportion of simulations showing a positive effect is the design's type I error rate. The proportion of simulated trials that show a positive treatment effect in its primary analysis is the trial's statistical power when the simulations assume a positive treatment effect[2]. The focus of the present article is when there are many other questions addressed during the trial. For example,

E-mail address: don@berryconsultants.net.

https://doi.org/10.1016/j.tjpad.2025.100392
Received 9 July 2025; Received in revised form 10 September 2025; Accepted 22 September 2025
Available online 27 October 2025

many arms (perhaps some of which involving combination therapy) may have been considered by the design but were not selected for the primary analyses. The primary analysis recognizes the various adaptations along the way. These adaptations may affect the trial's overall type I error rate and are considered in the calculations. However, the individual adaptations do not require high power.

The Bayesian approach can be used to build efficient and accurate clinical trials. Considering a particular adaptive feature will be helpful and illustrative for several reasons. A suggestion by W.R. Thompson in 1933 was to use a Bayesian calculation to set randomization probabilities in a two-armed clinical trial[3]. The next patient is assigned to an arm with probability that is "some monotone increasing function of the current [Bayesian] probability that it is the better of the two arms."[3] Thompson was motivated by ethics: "If such a discipline were adopted, then though it were not the best possible [strategy], it seems apparent that a considerable saving of individuals otherwise sacrificed to the inferior treatment might be effected."

Using a Bayesian calculation does not make the trial Bayesian. It would be Bayesian if the final analysis of treatment effect is its posterior probability distribution[4,5].

Thompson was using Bayesian statistics as a tool. To underline the point, statistician Peter Armitage once addressed the role of R.A. Fisher in the first use of randomization in a clinical trial. The trial addressed whether streptomycin improved the symptoms of tuberculosis. It was designed and conducted by A.B. Hill and the UK Medical Research Council (MRC). While working at the Rothamsted Experimental Station from 1919 to 1933, Fisher made countless contributions to statistical theory and practice. One was his invention of and promotion of randomization[6]. Another was his invention of factorial design of experiments, a subject considered below.

This is Armitage in 2003: "I know of no written comment by [R.A. Fisher] on clinical trials, although Hill once remarked to me that Fisher had suggested to him that randomization proportions should be altered dynamically as a function of the *P*-value from a significance test, so that as the difference became more significant a smaller proportion of patients received the apparently inferior treatment. …It may be that in the 1930s Fisher thought that doctors would never accept controlled experimentation, and he may even have had ethical objections to the idea."[7] As indicated earlier, Fisher was not a Bayesian. But he was obviously aware of Thompson's work. Fisher simply replaced Thompson's Bayesian probability with a frequentist measure: *P*-value. Back in the 1930s both Thompson and Fisher would have struggled to control type I error for such an adaptive clinical trial without the availability of simulations using modern computers.

2. Bayesian clinical trials

Bayesian and frequentist approaches are inverses of each other (see Table 1). For example, *P*-values are conditional probabilities (of results as or more extreme than the observed results) assuming that the null hypothesis is true. The Bayesian analog is also a conditional probability, but the inverse. It is the probability of the null hypothesis conditioning on the data actually observed in the trial. Bayes' rule "flips the conditionals." It also differs by excluding the seemingly irrelevant "more extreme" data.

A trial's type I error rate is prospective. It is a characteristic of the trial's design but not of its outcome. A *P*-value is the frequentist outcome measure analogous to type I error rate. But calculations of type I error rate can include the effects of Bayesian-driven interim analyses and decisions such as Thompson's adaptive randomization procedure.

The Bayesian approach is inextricably tied to decision-making[8]. Implications of experiments vary with their goals. For example, clinical trial designs should depend on the disease, including its prevalence now and in the future. Consider AD versus

Tay-Sachs disease (TSD). AD affects more than 7 million people in the US and kills over 120,000 of them. On the other hand, infantile TSD is a fatal, autosomal recessive genetic disease that affects fewer than 50 newborns per year in the US.

Table 1
Comparison of bayesian and frequentist approaches in designing and analyzing clinical trials.

Characteristic	Bayesian	Frequentist
Inferential unit	Patient (nested within trial)	Trial
Inferential measure	Probability distributions of parameters and hypotheses	*P*-values and confidence intervals
Probability applies to what?	All uncertainty, including hypotheses and future results	Data: assuming a particular hypothesis
Update probabilities of parameters	Frequently; as needed, including after each observation	Not applicable
Predicting future data (including patients within trial)	Based on current data and outside-trial information	Assumes particular value of unknown parameters
Number of questions addressed by trial	Any	Preferably one
Modeling	Lots of modeling, including of individual patients	Minimal
Decision analysis	Utilities are fundamental and integral to the approach	Awkward

Testing hypotheses and developing generalizable knowledge may be reasonable for AD. Indeed, randomizing hundreds or even thousands of patients with AD in a double-blinded trial while controlling type I error may be an approximate solution to the decision analysis problem. For TSD, on the other hand, randomizing in a clinical trial would be not only logistically difficult and may be unethical. For rare diseases quite generally, randomizing hundreds of patients in a clinical trial is impossible and controlling type I error rate is irrelevant.

A better strategy for both AD and TSD—and for all diseases—is using adaptive designs. One adaptation is increasing the trial's sample size, if necessary—smaller to get a compelling answer sooner, larger to get a stronger answer in the final analysis. The design should accommodate to the information accumulating in the trial[9–14].

Accumulating information includes longitudinal outcomes of individual patients. In AD, a therapy that prolongs survival would be expected to slow progression as measured by clinical markers or biomarkers such as amyloid plaques. Addressing correlations among the various outcomes serves many purposes. One is to enable more accurate imputations for missing data[15]. As a special case of missing data, many patients in AD trials are censored. Correlations among longitudinal outcomes in the trial, depending on therapy, would help in making Bayesian probabilistic predictions of times of death for patients who are still alive at the final analysis of the trial. Symptoms of TSD involve diminution of the senses, including the ability to move. Similarly, a therapy that prolongs survival would be expected to show an earlier effect on the symptoms of the disease, but not always. Bayesian longitudinal modeling is all about reducing uncertainty in the primary outcome measures.

Possible adaptations for AD trials include dropping arms in combination trials with factorial designs (see Section IV). They also include re-estimating sample sizes and adjusting randomization probabilities. Such adjustments include the possibility of setting an arm's randomization probability to 0, which means pausing accrual for the arm and possibly stopping its accrual permanently[15]. For TSD, the optimal adaptive design is simple, and I hope it is obvious. When a new investigational therapy is proposed, it should be assigned to patients (with no randomization) until the death rate shows that it is not an improvement over the historical standard of care. "Standard of care" is itself updated over time based on several factors, possibly including availability of

therapies not currently considered in the trial.

3. Bayesian adaptive platform trials

I-SPY 2 was a phase 2 adaptive Bayesian platform clinical trial in neoadjuvant breast cancer. It ran from 2010 to 2022. It used many statistical innovations as listed in

Table 2. Details are in the references[1,4,5,12–14,16]. How a revolutionary trial such as I-SPY 2 came into being is arguably more important than what the trial was.

Laura Esserman, MD, MBA, was the PI of I-SPY 2. My role was statistical design and serving as the trial's co-PI. Dr. Esserman was an enthusiastic champion and fellow disrupter. The trial would not have been successful without Dr. Esserman's advocacy and strong push for innovation. Others who were instrumental in the formative stages of I-SPY 2 were Janet Woodcock, MD, director of the FDA's Center for Drug Evaluation and Research (CDER), and Anna Barker, PhD, former deputy director of the National Cancer Institute.

I-SPY 2′s design has served as a prototype for platform trials in a variety of diseases, including registration trials in oncology. It was also the model for IMI's EPAD—see below. The design of I-SPY 2 and its descendants are complicated and were made possible by modern improvements in Bayesian statistical software and computer hardware. As indicated above, the FDA recognized complicated designs for registration under the rubric CIDs[1]. I-SPY 2 was a prototype for CIDs, even though it was not itself a registration trial.

The design of I-SPY 2 was itself an experiment, in effect it was designing a trial starting from scratch. We bypassed standard approaches. Drs. Esserman, Woodcock, Barker, and initial funding from the Foundation for the National Institutes of Health were essential. The design was based on multiarmed bandit problems[17,18]. The goal was to treat trial participants effectively while learning efficiently and accurately about the effects of the various therapies [17,18], especially therapies that are effective. One proposal for such a design was Thompson's[3]. The experience of using response-adaptive randomization in scores of trials conducted in the early 2000x at MD Anderson Cancer Center was also essential[19,20].

The FDA's support for I-SPY 2 and beyond I-SPY 2 was instrumental, especially Dr. Woodcock's support. In the case of I-SPY 2 and GBM AGILE, the FDA has encouraged and indeed led innovations. In an article on master protocols, Woodcock and LaVange advertised I-SPY 2 as a prototypic trial[16]. For example: "Innovative aspects of the I-SPY 2 trial design include response-adaptive randomization to assign patients to the most promising treatment or combination of treatments in their respective molecular breast-cancer subgroups…while maintaining a sufficient number of patients assigned to the standard of care, shared use of control patients across treatment comparisons, and Bayesian decision rules to determine whether or when therapies with low probabilities of success or side effects should be discontinued and therapies with high probabilities of future success…should advance for further study."[16]

As indicated above, I-SPY 2′s innovations were driven and facilitated by a Bayesian decision-analytic approach. The trial was successful in using its innovations for the benefit of patients both in the trial and after the trial. Nine of the 23 investigational therapies graduated ready for phase 3, representing most of the trial's 10 prospectively defined, molecular marker-defined signatures. Four of the nine graduates have since received marketing approval within molecular subtypes of adjuvant or neoadjuvant breast cancer. But there's something for everyone (see Table 2). For example, investigational therapies that did not graduate had a thorough phase 2 randomized evaluation in the various molecular subtypes of the disease along with a large-sample-size common control arm.

I-SPY 2 ended in 2022. It was replaced by a very different trial called I-SPY 2.2 that I did not support[21].

The I-SPY 2 trial set a standard for Bayesian platform trials, including in Alzheimer's disease. According to the Innovative Medicines Initiative (IMI) in 2015:

> "Innovative clinical trial designs. The definition and implementation of innovative trials to accelerate access to efficient and safe medicines is of major interest to industry, regulators and patient organizations. Inspired by the I-SPY initiative, the EPAD consortium will develop an adaptive design in a proof-of-concept trial for early intervention in Alzheimer's disease."[22]

Despite the optimism expressed by the IMI, the European Prevention of Alzheimer's Dementia (EPAD) story was mainly negative. However, it

Table 2
Innovative features of I-SPY 2. These features apply regardless of disease and endpoints.

Adding and dropping arms. I-SPY 2 was designed to be potentially never-ending, although it did end in 2022. Arms in the trial would stop accruing patients when they graduated to phase 3, stopped for futility, or reached a predetermined maximum sample size. In all cases, we followed patients and kept confidential the fact that the arm had stopped accruing patients until all the arm's patients were through surgery.
Collection of basket trials. Patients were assigned to one of eight subtypes defined by three molecular tumor markers: hormone receptor status (HR), HER2 receptor status (HER2), and MammaPrint (MP). Investigational arms were evaluated in up to 10 signatures that are combination patient subtypes and that are possible clinical indications.
Definition of Type I error. The basket aspect of the investigational arms meant that there are several possible types of error for each arm. For example, an arm may graduate in a signature for which some subtypes are correctly positive, but others are incorrectly positive. That's not a type I error. In I-SPY 2, I defined a false-positive conclusion as the case when the concluding signature contained no patients who would benefit from the therapy.
Bayesian predictive probability of success in a future phase 3 trial. Adaptive randomization was based on these predictive probabilities. Stopping accrual because of graduation to phase 3 or futility was based on the Bayesian predictive probability of future success.
Continuous learning and Bayesian updating. Each month we calculated the current distributions of pCR rates for all subtypes and all possible signatures for all arms in the trial. We used these distributions to calculate predictive probabilities and, in turn, the assessment of each investigational arm's status in the trial. This included monthly updates of each arm's randomization probabilities and decisions regarding graduation to phase 3, dropping for futility, and stopping accrual at the arm's maximum sample size.
Common control arm (by patient subtype). Investigational arms in I-SPY 2 were compared against a common set of controls, depending on patient subtype. The final analysis of each investigational arm was its Bayesian probability of superiority to control for the primary endpoint of pCR for each signature.
Time machine[34,35]. Patients were assigned to the control arm with 20 % probability unless there was only one investigational arm in the subtype, in which case randomization was 1:1. The control cohort for an investigational arm includes all concurrently randomized controls plus all previously randomized controls. However, the outcomes of the previous non-concurrently randomized controls were adjusted and partially discounted for time trends in the trial. The time trends were assessed for all the arms in the trial, including the other investigational arms. The result was that we built a large data bank of controls that was typically an order of magnitude greater than the total number of patients in the investigational arms, with statistical power much, much greater than typical phase 2 cancer trials.
Longitudinal model of disease burden as auxiliary endpoint[5]. We built a model that related reductions in tumor volume during neoadjuvant treatment, depending on patient subtype. We used historical data from I-SPY 1 but updated the model using I-SPY 2 patients who had experienced surgery. At each analysis epoch, we calculated the probability of achieving pCR for each patient in the trial who had not yet had surgery. We used multiple imputation to find the probability distributions of all the pCR rates for the arms in the trial [4,5].
Nested partial factorials. The continuing control arms of I-SPY 2 enabled addressing combination therapies. One of the initial arms in the trial was AbbVie-sponsored veliparib + carboplatin (VC)[5]. Its successful graduation led to a trial that isolated the effects of V and C[36].
Use of computer simulation to assess design operating characteristics. Traditional statistical designs of clinical trials include type I error rate and statistical power. They should also be addressed in Bayesian trials, especially type I error, if only to preserve continuity with traditional trials.

made important contributions and helped the community understand the difficulties in running a platform trial in AD. The IMI (now the Innovative Health Initiative, or IHI) arranged for generous funding for the EPAD consortium (that included Berry Consultants whom the IMI selected in a competitive process). From 2015 to 2020, the EU contributed 26 M € and the European Federation of Pharmaceutical Industries Associations (EFPIA) contributed another 27 M € of the total 59M € in EPAD funding[22]. The development of EPAD floundered until October 2020 when it failed.

According to its website, "EPAD was a unique collaborative research effort. Our 39 partners across Europe were committed to transforming our understanding of Alzheimer's disease."[22] There were identifiable reasons for EPAD's ultimate failure, but the details are not public. However, the impact of the COVID-19 pandemic did not help, nor did Brexit. Also, having 39 partners in a trial may be wonderful from several perspectives, but a developmental process can be dysfunctional if some of the partners are competitors. This is so even if they have equal shares in making the project successful. There can be too many cooks.

Platform trials have many stakeholders. A lesson from EPAD is that it is difficult to build a platform trial that promises rewards for every stakeholder. Adequate funding from philanthropy, government, and industry is essential but not sufficient. A corollary is that the design must be flexible in terms of patient population and patient subpopulations for the individual treatment arms. Moreover, the design should be flexible in terms of issues such as sample size and statistical power. The trial could be registrational for some arms and pre-registrational for other arms and post-registrational for still other arms. Although the trial must have a master protocol, each treatment arm that enters the trial must have its own appendix, and the master protocol should state explicitly which aspects of an arm's appendix can supersede those aspects of the master protocol.

4. Investigating multiple therapeutic issues

A typical empirical researcher frequently asks this question: "Is your observation statistically significant?" The questioner is asking if the associated *P*-value is less than 0.05. If so, then the questioner and others in the audience accept the observation as "real." Thus, they conclude that the experimental treatment is superior to control, or that a biomarker subgroup has a greater risk for a particular disease than the larger population, etc. If the *P*-value is greater than 0.05, then the reaction is usually that the observation is spurious.

There are many flaws with such reasoning. One is that, regardless of what interpretation one gives to a *P*-value, there is nothing magical about the threshold value of 0.05. Simply put, statistical significance cannot substitute for truth. Moreover, in a decision-making approach, there can be no single threshold for judging significance.

Asking many questions in a clinical trial raises statistical issues that are widely regarded to be problematic. They are almost always less problematic than researchers think. Some feel that merely asking many questions of data is a problem. That feeling becomes part of the culture, to the detriment of science[23].

> R.A. Fisher weighed in regarding these issues: "No aphorism is more frequently repeated in connection with field trials [and clinical trials], than that we must ask Nature few questions, or, ideally, one question at a time. The writer [Fisher] is convinced this view is wholly mistaken. Nature, he suggests, will best respond to a logical and well-thought-out questionnaire. Indeed, if we ask her a single question, she will often refuse to answer until some other topic has been discussed."[24]

Resolving one of the multiplicity issues helps in understanding and potentially solving other issues. Consider dose-response because its solution reveals solutions for other multiplicity issues. Getting the dose right has always been a difficult problem in drug development, and it is an expensive problem—unnecessarily expensive. In AD, identifying the wrong dose in phase 2 likely accounts for many of the failures in phase 3 trials. However, it is difficult to know for sure because it is impossible to distinguish between a phase 3 failure caused by an ineffective dose and one caused by an ineffective drug.

Traditional dose-response clinical trials are woefully inefficient. The conventional approach is to compare doses with placebo in a pairwise fashion and then to pay statistical penalties in type I error rate. Half of the problem is cured by recognizing that the true response for a dose between two other doses is almost always between the true responses at the other doses. Researchers should model the dose-response relationship with a family of functions that allow for monotonicity. Then, if the data are consistent with the assumption of monotonicity, the modeling process borrows outcome information across doses, greatly increasing the power of the trial.

Such modeling applies regardless of how dose assignments are made, even if they stay the same throughout the trial. But it is possible to do a lot better by adapting randomization. That is the other half of the solution. Assigning to some doses will be relatively uninformative about what doses are better. Non-informative doses will be revealed by interim results as the trial proceeds. Researchers should build design algorithms that assign few if any patients to such doses.

Imagine a robot that has been programmed to assign doses to patients—artificial intelligence. The robot is unblinded to accumulating data and "observes" what doses are performing well and which are not. It learns. Program the robot to avoid underperforming doses, including dropping such doses…with the possibility of restarting doses that are becoming more promising with longer follow-up. The robot is very specifically programmed to learn about doses that are likely doses for phase 3, say. Have it stop the trial when it knows enough to make a go/no-go decision regarding phase 3, and when it determines what dose or doses to use in phase 3 if the drug is a "go."

Go/no-go decisions should depend on Bayesian predictive probabilities of a successful phase 3 assuming longer follow-up of patients already enrolled in the trial. Phase 3 can be a second trial or a second stage of the present trial. In the latter case, the trial can shift seamlessly into the second stage with the intention that the combination of the two stages will count as a single phase 3 trial.

Once the robot has modeled dose-response and adapted to results accruing in the trial, it has other important tasks. One involves the endpoint. Suppose the primary endpoint is some combination of cognition and function at 12 months. Earlier information about the patients' responses to treatment may be available (including the same cognition/function measure, just assessed earlier). You, the designer, build a model that has unknown longitudinal parameters for the robot so that it can predict the primary endpoints for those patients who do not yet have 12-month visits. This longitudinal model should depend on treatment arm (ie, dose) as well as patient covariates. Patients with fewer than 12 months follow-up have a probability distribution in place of primary endpoint. Therefore the robot's statistical analyses, including interim analyses, can include all patients accrued so far in the trial via multiple imputation.

Coming this far means you have learned about the lecanemab phase 2 trial 201 in AD [15,25], the Pfizer ASTIN trial in stroke [26], and Eli Lilly's AWARD-5 phase 2/3 trial of dulaglutide in type 2 diabetes[27]. These three trials had some similarities, including that they were wholly Bayesian.

One aspect of the lecanemab trial deserves special mention: modeling missing data. The Bayesian longitudinal model of AD in this phase 2 trial[15] was essential in view of the substantial missing data. Namely, the model applied whether the missingness was due to missing visits or right censoring. Some of the right censoring was caused by an external regulator. In all cases the modeling worked perfectly. The protocol specified Bayesian analysis of the phase 2 trial exactly predicted the outcome of the lecanemab phase 3 trial[15,28].

Finally, Lilly's AWARD-5 trial had three additional jobs in its robot's workload[27]. One was to include safety in the primary analysis via a

clinical utility index (CUI) that would be used for all the robot's decisions. Each of the CUI's components required longitudinal modeling. The second job was making the phase 3 go/no-go decision and picking the two doses that would move seamlessly to a fixed-randomization stage 2 if the decision was "go," which it was. Third, the robot determined the sample size for the fixed-randomization stage 2 (which would be combined with stage 1 dose-finding results in carrying out the trial's final analysis).

The FDA has been enormously supportive of such innovations. Its support includes granting accelerated approval of lecanemab-irmb (Leqembi®) based on phase 2 trial 201 and full approval of the two doses of dulaglutide (Trulicity®) based on the AWARD-5 trial and other phase 3 trials that compared dulaglutide with active control arms. An FDA initiative that may have been affected by these trials is the CID[1].

An FDA initiative of its Oncology Center of Excellence (OCE) called Project Optimus regards getting the dose right[29]. Their guidance dated August 2024 addresses the possibility that two or more doses are carried into a phase 3 trial with the intention of dropping all but one dose during the trial. According to the guidance's Section III, Dosage Optimization Recommendations, subsection B, Trial Designs to Compare Multiple Dosages: "The trial does not need to be powered to demonstrate statistical superiority of a dosage or statistical non-inferiority among the dosages using type I error rates which would be used in registrational trials." So there is no need for statistical significance ($P < 0.05$) of one of the doses over another dose in order to drop the latter for inferiority. An interpretation of this directive is, "Give us some randomized evidence comparing the doses with each other but not necessarily highly powered evidence."

In recognition that there is an impact on a trial's overall type I error rate associated with dropping an arm, the guidance goes on to say that "the trial design should provide strong control of Type I error." Indeed, this and other adaptations are accommodated in the trial simulations to demonstrate control of type I error.

This attitude is enormously important and may set a positive regulatory precedent. The term *moderate power* for adaptive decisions applies to those for which "The trial does not need to be powered to demonstrate statistical superiority … or statistical non-inferiority." It is not clear whether moderate powering applies in other circumstances (such as adaptively dropping single agents in a factorial trial) or for other FDA divisions (such as Neurology). However, in my experience the OCE is serious about moderate powering regarding adaptively dropping doses during registration trials in oncology. In March 2023, the OCE agreed with my proposal to drop one of two doses in an ongoing trial based on a moderate Bayesian probability of inferiority.

Factorial designs

In developing combination therapies, "moderate powering" should be proposed in what the FDA called "adaptive design elements of factorial and partial factorial designs." Consider two drugs, A and B, at least one of which is investigational. Combination AB is compared in a clinical trial with no therapy, A^cB^c, where superscript ^{c} stands for complement. If AB fares significantly better than A^cB^c in the trial, should the combination be approved? No, not unless there's good outside-the-trial evidence that both A and B are contributing to the outcomes of combination AB.

At a July 25, 2024 meeting of an FDA advisory panel, the Oncologic Drugs Advisory Committee (ODAC) was presented results of a trial that had compared AB with A^cB^c. The FDA's briefing document for ODAC stated the FDA's position, including that "in a Type B meeting held on November 01, 2018, FDA stated the design of AEGEAN [the trial in question] would not isolate the effect of the treatment phases [A and B] and recommended that the Applicant should consider a factorial study design, potentially with adaptive design elements. The Applicant opted to proceed with a two-arm trial."

The specifics of AEGEAN were a bit different from what the above description. Therapies A and B were not different drugs but different periods (phases) of using the same drug, neoadjuvant and then adjuvant. So in essence AB was longer therapy, called "perioperative." It can be thought of as a dose regimen. The FDA asked ODAC to vote on this question: "Should FDA require that new trial design proposals for perioperative regimens for resectable NSCLC include adequate within trial assessment of contribution of treatment phase?" ODAC voted "Yes," unanimously.

FDA and ODAC questioned the applicant regarding why they had not conducted a factorial trial. A 2×2 factorial trial would have four treatment groups: AB, AB^c, A^cB, A^cB^c. The applicant's rationale in answering a direct question from ODAC was that a factorial design or even a 3-armed (partial factorial) design would require a much larger sample size and would take much longer to conduct—in their discussion, ODAC used the estimate of two years longer.

As explained below, this response and its rationale misses out on the fundamental benefit of factorial design. My explanation will also suggest what the FDA likely meant by "potential adaptive design elements." Further, it raises the issue of whether "moderate powering" can be applied to factorial designs and other similar innovations.

The applicant's presentation and the ensuing discussion at the ODAC meeting was prefaced by the statistical comparisons being across pairs of the four patient treatment arms. That is not what R.A. Fisher had in mind when he invented factorial designs while working at Rothamsted Experimental Station from 1919 to 1933. In 1935, he wrote the book on factorial designs, *The Design of Experiments*[30]. All 10 editions of the book have stressed that statistical analyses of factorial designs should focus on the factors (main effects) and not on individual arms. The difference is critical, with implications on statistical power and sample sizes.

Consider AB – AB^c and A^cB – A^cB^c. These two differences are independent, and both estimate the benefit of factor B. The best single estimate of that benefit is the average of the two. Similarly, AB – A^cB and AB^c – A^cB^c are independent estimates of factor A. The primary analyses in a classical 2×2 factorial experiment are estimates of the "main effects of A and B" and not, for example, on comparison A^cB versus A^cB^c for the effect of B. A secondary analysis is estimating the interaction between A and B. For example, the two estimates of B's effect, AB – AB^c and A^cB – A^cB^c, may be different. Assuming that there's no interaction means the second factor is included in the trial for free.

Fisher argued that there are two advantages of factorial design. One occurs when there is no interaction because you then get two trials for the price of one. The other occurs when there is an interaction, for in that case, a factorial design is the only way to get information about the interaction. But to repeat, addressing interactions does not require high power.

A reviewer suggested that my enthusiasm for factorial designs may be excessive. One concern was the possibility of ambiguous interactions. Similarly, an NIH cooperative group statistician once told me they had designed a factorial design and it was a disaster; they'd never do it again. So what happened? It turned out that both single agent arms were as effective as their combination; a classical negative interaction. My reaction: "That's great. You learned something very important. Suppose you had run a two-armed trial instead, comparing the combination with control. Think of all the people who would have been overtreated before the error was discovering." There is no outcome that would lead me to regret using a factorial design. There are many outcomes of non-factorial trials that would lead me to regret not using a factorial design, including a case in I-SPY 2.

Example of two trials for the price of one

In the 1990s the breast cancer committee of the Cancer and Leukemia Group B (CALGB), a national oncology group, carried out many trials having factorial designs at my urging. None of these trials suggested even a hint of an interaction either before or after the results

became known.

As an example, in 1993–1994, CALGB was charged by the US. Breast Cancer Intergroup with designing and running the next adjuvant breast cancer trial in lymph node–positive disease. There were two main competing proposals for the scientific question to be addressed. One was to add four 3-week cycles of paclitaxel (Taxol®) to the standard four 3-week cycles of doxorubicin and cyclophosphamide.

The other proposal from the breast cancer community was the dose-response of doxorubicin. The standard dose was 60 mg/m^2. A previous CALGB study had shown that this dose was better than two lower doses. When this trial was being designed in the 1990s, there was a great deal of hype and some negative opinions in the community regarding high-dose therapy. The proposers of this factor wanted to address three doses of doxorubicin: 60, 75, and 90 mg/m^2.

My recommendation was to conduct a 3 × 2 factorial that would address both questions. There had been very little experience with factorial trials in cancer research, and there was some reluctance to undertaking such an innovation. But after several presentations to the CALGB, they agreed. Most important was the strong support of Craig Henderson, MD, the chairperson of the CALGB Breast Cancer Committee. We met with some resistance at the US Breast Cancer Intergroup level. They worried that presenting a complicated trial to patients would affect accrual. Thus, we incorporated the option to drop two of the six treatment arms if the trial did not accrue well, a simple type of adaptation that required no statistical adjustment because it depended only on information about the trial that was openly available.

There was a bottleneck at the sponsoring agency, the National Cancer Institute. Their statisticians expressed concern that the proposed 3000 patients to address the paclitaxel question would not be enough to highly power the ability to address the interaction between paclitaxel and dose of doxorubicin. But they eventually agreed that not all questions in a clinical trial need to be answered with high power. And they agreed that the two factors in this trial were unlikely to interact. Moreover, we had "moderate power" to address interactions, and some information is better than no information.

We ran the trial, CALGB 9344 [31,32]. It accrued surprisingly well, with 3121 patients across the six treatment arms. The trial was practice-changing in two ways, based on the two main effects[32]. Paclitaxel reduced the risk of recurrence or death by 17 %. In 1999 the drug received FDA approval based on CALGB 9344. As regards the other factor, increasing the dose of doxorubicin in the trial showed no difference in clinical outcomes. Further, there was not even a hint of an interaction between the two factors. The results and conclusions were so clear that the dose question and the fact that paclitaxel was only one of the two main effects in a factorial trial were not even mentioned at the ODAC meeting that addressed the approval of adding four cycles of paclitaxel. Both the use of paclitaxel and the doxorubicin dose of 60 mg/m^2 continue to be standard in clinical practice today.

Somewhat serendipitously, the circumstance of CALGB 9344 offered a perfect opportunity to check claims about the benefits of factorial designs. An almost identical trial, B-28 [33], was conducted at the same time by the National Surgical Adjuvant Breast and Bowel Project (NSABP). The difference between the two trials was that B-28 did not have a factorial design. It addressed only the addition of paclitaxel. All patients received the 60-mg/m^2 dose of doxorubicin. The sample size of B-28 was 3060, only 61 patients fewer than CALGB 9344. Moreover, B-28′s outcome was the identical 17 % reduction in disease recurrence associated with adding paclitaxel. Regarding trial length, the time between first patient accrued and announcement of results was essentially the same in the two trials. Thus, the factorial design was truly two trials—two practice-changing trials—for the price of one. And there were no negative consequences of doubling the value of the trial, including essentially no extra cost.

Sizing factorial clinical trials and adaptations

The unfortunate and undeserving reputation of factorial designs regarding increased sample size stems from the tradition of comparing arms rather than main effects. A clinical question is, "What do A and B contribute to the combination AB?" This question cannot be addressed adequately by any two-arm clinical trial. Analyses of factorial designs that ignore the relationships among the four arms are scientifically flawed. They leave critical information on the table.

If one assumes there is no interaction between factors, then there is *no increase in sample size* over a two-armed trial that addresses a single factor. Of course, one should test for interaction at the end of the trial. If there is a suggestion of interaction, then that would be important and useful information.

Adaptive factorial clinical trials

In accordance with moderate powering for an adaptation, a sponsor could begin a clinical trial with a factorial design having equal randomization probabilities to the four arms but then adapt the randomization to the accumulating data. There are many possibilities. For example, suppose the interim results and predictive probabilities of future results are suggesting that factors A and B are both contributing to the benefit of AB. Then, single-agent arms, which in the above notation are arms A^{c}B and AB^c, could be assigned lower probabilities and even eventually dropped completely (but with follow-up continuing for all patients in the trial). The two arms remaining would be AB and A^{c}B^c, which would maximize the precision regarding the combination versus no therapy. In any case, the sample sizes for the individual arms should depend on predictions of the future of the trial with potential effects on sample sizes. The final sample size will be random, but as the FDA suggested at the aforementioned ODAC meeting, it would almost certainly be substantially smaller and the trial substantially shorter than the sample sizes that were presented by the applicant and that were cited by the ODAC panel members.

There are various adaptations to consider. A tack that is an opposite of that in the previous paragraph also works. Emphasize AB and A^{c}B^c early in the trial, dropping or lowering its randomization probability when there is substantial predictive probability that the trial will demonstrate a benefit for AB over A^{c}B^c. The primary focus of the trial would then switch to isolating the contributions of both A and B to AB. Regardless of tack, the design algorithm can be built to decide which if any of the candidate therapies are better than which other therapies.

All adaptations must be set in stone prospectively and applied in accordance with the trial's protocol and Statistical Analysis plan (SAP). Prospective design is necessary to enable what the FDA calls "strong control of type I error" for the primary final analyses. Calculating and controlling Type I error rate will no doubt require simulation. Such a trial could qualify for FDA's CID initiative as a registration trial[1].

5. Discussion and conclusions

In the spirit of R.A. Fisher, it is always possible and it is always efficient to ask more than one question in a clinical trial. Adaptive factorial designs are always available in trials that address combination therapies. Sometimes, as in the case of interactions in factorial trials, achieving efficiency requires modeling and making assumptions and perhaps synthesizing with evidence from outside the trial in question. For NIH and patient advocacy-sponsored trials there is no excuse for running one-question clinical trials. On the other hand, industry-sponsored drug trials will no doubt continue to be predominately two-armed trials. However, industry trials addressing combination therapy should use adaptive factorial trials as described above.

The Bayesian approach has always had theoretical appeal—see Table 1. Over the past 30 years, the theory has been applied in designing actual clinical trials. Bayesian trials still represent a small minority of all

trials, but their existence is changing the way investigators, regulators, and government and industry sponsors view innovation in clinical trials.

By far the most important Bayesian contribution to clinical trials is the ability to observe the accumulating results and modify the future course of the trial on their bases. The approach in this article is that the "observer" is an automaton armed with a prospective algorithm that spells out and dictates adaptations that were determined in advance of the trial. Any such design can then be simulated to calculate its type I error rate and statistical power.

The Bayesian adaptive approach applies to addressing and answering multiple questions in a single clinical trial. An example includes evaluating many doses. Another is evaluating the effectiveness of combination therapy versus single-agent therapies while also considering questions regarding which therapies are best in which patient biomarker-defined subpopulations.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Donald A. Berry reports financial support was provided by Berry Consultants LLC. Donald A. Berry reports financial support and article publishing charges were provided by Alzheimer's Drug Discovery Foundation. Donald A. Berry reports writing assistance was provided by Precision AQ with financial support from the Alzheimer's Drug Discovery Foundation, and in compliance with International Good Publication Practice guidelines.

Acknowledgements

Precision AQ in Bethesda, Maryland, provided editorial support under the direction of the author, with financial support from the Alzheimer's Drug Discovery Foundation, and in compliance with International Good Publication Practice guidelines.

References

[1] Interacting with the FDA on complex innovative trial designs for drugs and biological products: guidance for industry. 2020. https://www.fda.gov/media/130897/download. Accessed 12 April 2025.
[2] Adaptive designs for clinical trials of drugs and biologics: guidance for industry. https://www.fda.gov/media/78495/download. 2019. Accessed 22 December 2024.
[3] Thompson WR. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika 1933;25:285–94. https://doi.org/10.2307/2332286.
[4] Park JW, Liu MC, Yee D, et al. Adaptive randomization of neratinib in early breast cancer. N Engl J Med 2016;375:11–22. https://doi.org/10.1056/NEJMoa1513750.
[5] Rugo HS, Olopade OI, DeMichele A, et al. Adaptive Randomization of veliparib-carboplatin treatment in breast cancer. N Engl J Med 2016;375:23–34. https://doi.org/10.1056/NEJMoa1513749.
[6] Savage LJ. On rereading R.A. Fisher. Ann Statist 1976;4:441–500. https://doi.org/10.1214/aos/1176343456.
[7] Armitage PFisher. Bradford Hill, and randomization. Int J Epidemiol 2003;32:925–8.
[8] Savage LJ. The foundations of statistics. New York: Dover Publications, Inc.; 1954.
[9] Berry DA. Bayesian statistics and the efficiency and ethics of clinical trials. Stat Sci 2004;19:175–87. https://doi.org/10.1214/088342304000000044.
[10] Berry DA. Introduction to Bayesian methods III: use and interpretation of Bayesian tools in design and analysis. Clin Trials 2005;2:295–300. https://doi.org/10.1191/1740774505cn100oa.
[11] Meurer WJ, Lewis RJ, Berry DA. Adaptive clinical trials: a partial remedy for the therapeutic misconception? JAMA 2012;307:2377–8. https://doi.org/10.1001/jama.2012.4174.
[12] Berry DA. Adaptive clinical trials in oncology. Nat Rev Clin Oncol 2012;9:199–207. https://doi.org/10.1038/nrclinonc.2011.165.
[13] Berry DA. The brave new world of clinical cancer research: adaptive biomarker-driven trials integrating clinical practice with clinical research. Mol Oncol 2015;9:951–9. https://doi.org/10.1016/j.molonc.2015.02.011.
[14] Berry DA. State of the art: emerging innovations in clinical trial design. Clin Pharmacol Ther 2016;99:82–91. https://doi.org/10.1002/cpt.285.
[15] Berry DA, Dhadda S, Kanekiyo M, et al. Lecanemab for patients with early Alzheimer disease: bayesian analysis of a phase 2b dose-finding randomized clinical trial. JAMA Netw Open 2023;6:e237230. https://doi.org/10.1001/jamanetworkopen.2023.7230.
[16] Woodcock J, LaVange LM. Master protocols to study multiple therapies, multiple diseases, or both. N Engl J Med 2017;377:62–70. https://doi.org/10.1056/NEJMra1510062.
[17] Berry DA. Modified two-armed bandit strategies for certain clinical trials. J Am Stat Assoc 1978;73:339–45. https://doi.org/10.2307/2286662.
[18] Berry DA, Fristedt B. Bandit problems: sequential allocation of experiments. Netherlands: Springer; 1985.
[19] Biswas S, Liu DD, Lee JJ, Berry DA. Bayesian clinical trials at the University of Texas M. D. Anderson Cancer Center. ClinTrials 2009;6:205–16. https://doi.org/10.1177/1740774509104.
[20] Gönen M. Bayesian clinical trials: no more excuses. Clin Trials 2009;6:203–4. https://doi.org/10.1177/1740774509105374.
[21] Essermen L., De Michele A., Symmans W.F., et al. I-SPY 2.2: evolving the I-SPY 2 trial to adapt therapies for each patient and optimize outcome. 2021. https://globalforum.diaglobal.org/issue/july-2021/i-spy-2-2-evolving-the-i-spy-2-trial-to-adapt-therapies-for-each-patient-and-optimize-outcome/. Accessed 12 April 2025.
[22] EPAD: european prevention of Alzheimer's dementia consortium https://www.ihi.europa.eu/projects-results/project-factsheets/epad. Accessed 12 April 2025.
[23] Berry DA. Multiplicities in cancer research: ubiquitous and necessary evils. J Natl Cancer Inst 2012;104:1125–33. https://doi.org/10.1093/jnci/djs301.
[24] Fisher RA. The arrangement of field experiments. J Minist Agricult 1926;33:503–15. https://doi.org/10.23637/rothamsted.8v61q.
[25] Dhadda S, Kanekiyo M, Li D, et al. Consistency of efficacy results across various clinical measures and statistical methods in the lecanemab phase 2 trial of early Alzheimer's disease. Alzheimers Res Ther 2022;14:182. https://doi.org/10.1186/s13195-022-01129-x.
[26] Berry DA, Müller P, Grieve AP, et al. Adaptive Bayesian designs for dose-ranging drug trials. In: Gatsonis C, Kass RE, Carlin B, et al., editors. Case studies in bayesian statistics v. New York: Springer-Verlag; 2002. p. 99–181.
[27] Geiger MJ, Skrivanek Z, Gaydos BL, Chien JY, Berry SM, Berry DA. An adaptive, dose-finding, seamless phase 2/3 study of a long-acting glucagon-like peptide-1 analog (dulaglutide): trial design and baseline characteristics. J Diabetes Sci Tech 2012;6:1319–27. https://doi.org/10.1177/193229681200600610.
[28] van Dyck CH, Swanson CJ, Aisen P, Bateman RJ, Chen C, Gee M, Kanekiyo M, Iwatsubo T. Lecanemab in early Alzheimer's disease. N Engl J Med 2023;388:9–21. https://doi.org/10.1056/NEJMoa2212948.
[29] The Oncology Center of Excellence. Project optimus: reforming the dose optimization and dose selection paradigm in oncology. 2024. https://www.fda.gov/about-fda/oncology-center-excellence/project-optimus. Accessed 12 April 2025.
[30] Fisher RA. The design of experiments. Edinburgh: Oliver and Boyd; 1935.
[31] Couzin J. The new math of clinical trials. Science 2004;303:784–6. https://doi.org/10.1126/science.303.5659.784.
[32] Henderson IC, Berry DA, Demetri GD, et al. Improved outcomes from adding sequential paclitaxel but not from escalating doxorubicin dose in an adjuvant chemotherapy regimen for patients with node-positive primary breast cancer. J Clin Oncol 2003;21:976–83. https://doi.org/10.1200/JCO.2003.02.063.
[33] Mamounas EP, Bryant J, Lembersky B, et al. Paclitaxel after doxorubicin plus cyclophosphamide as adjuvant chemotherapy for node-positive breast cancer: results from NSABP B-28. J Clin Oncol 2005;16:3686–96. https://doi.org/10.1200/JCO.2005.10.517.
[34] Saville B, Berry DA, Berry NS, Viele K, Berry SM. The Bayesian time machine: accounting for temporal drift in multi-arm platform trials. Clin Trials 2022;19:490–501. https://doi.org/10.1177/17407745221112013.
[35] Berry SM, Reese CS, Larkey PD. Bridging different eras in sports. J Am Stat Assoc 1999;94:661–76. https://doi.org/10.1080/01621459.1999.10474163.
[36] Loibl S, O'Shaughnessy J, Untch M, et al. Addition of the PARP inhibitor veliparib plus carboplatin or carboplatin alone to standard neoadjuvant chemotherapy in triple-negative breast cancer (BrighTNess): a randomised, phase 3 trial. Lancet Oncol 2018;19:497–509. https://doi.org/10.1016/S1470-2045(18)30111-6.