RESEARCH · RESEARCH DESIGN

Adaptive Research Designs That Evolve Based on Interim Findings

How adaptive research designs use interim results to change a study while it runs, covering group sequential stopping rules, sample size re-estimation, dropping and adding arms, platform trials, response-adaptive allocation and bandits, responsive survey designs, emergent qualitative designs, the risks of adaptation and the safeguards that keep conclusions valid.

ModQu research article · 2,554 words · 23 references (APA 7th edition)

A conventional study is fixed in advance. The design, the sample size and the analysis are set before data collection begins, and nothing changes until the end. This has a clear advantage: the conclusions cannot be shaped by what the researcher saw along the way. It also has a cost. If early data show that one treatment is clearly harmful, that the sample size was based on a wrong assumption or that one group of respondents is not being reached, a fixed design continues regardless.

An **adaptive design** allows planned changes to a study based on data collected while it is running. Pallmann et al. (2018) defined it as a design that offers pre-planned opportunities to modify aspects of an ongoing study in light of the accumulating data, while preserving its validity and integrity. The key words are "pre-planned" and "validity". Adaptation is not improvisation. This article explains the main types of adaptive design, how they are used in clinical trials, surveys, online experiments and qualitative research, and the safeguards that keep their conclusions trustworthy.

Why looking at interim data is dangerous

The case for careful rules starts with a simple problem. If a researcher tests for a significant difference after every batch of data and stops as soon as one appears, the chance of a false positive rises well above the nominal 5%. Armitage et al. (1969) showed that repeated significance testing on accumulating data inflates the error rate substantially: with five tests at the conventional 5% level, the overall chance of at least one false positive is roughly 14%, and with unlimited looks it approaches certainty.

The same problem appears in online experiments, where dashboards show results continuously. Johari et al. (2017) showed that the common practice of watching an A/B test and stopping when it reaches significance produces far more false positives than the stated level, and proposed methods that remain valid however often results are checked. Every adaptive design must solve this problem: it must decide in advance when and how the data will be examined, and what will happen as a result.

Group sequential designs

The oldest and most widely used adaptive approach is the **group sequential design**. The study plans a small number of interim analyses, for example after each fifth of the planned participants, and sets a **stopping boundary** for each one. If the result crosses the boundary, the study stops early; otherwise it continues. The boundaries are chosen so that the overall false positive rate stays at the intended level. The idea goes back to work on sequential testing by Wald (1945), which showed that decisions could be made as data accumulated rather than only at the end.

Two classic boundaries differ in how they spend the error rate. Pocock (1977) proposed using the same strict threshold at every analysis. O'Brien and Fleming (1979) proposed very strict thresholds early on, relaxing towards the conventional level at the final analysis, so that a study stops early only for overwhelming evidence and its final analysis is almost the same as in a fixed design.

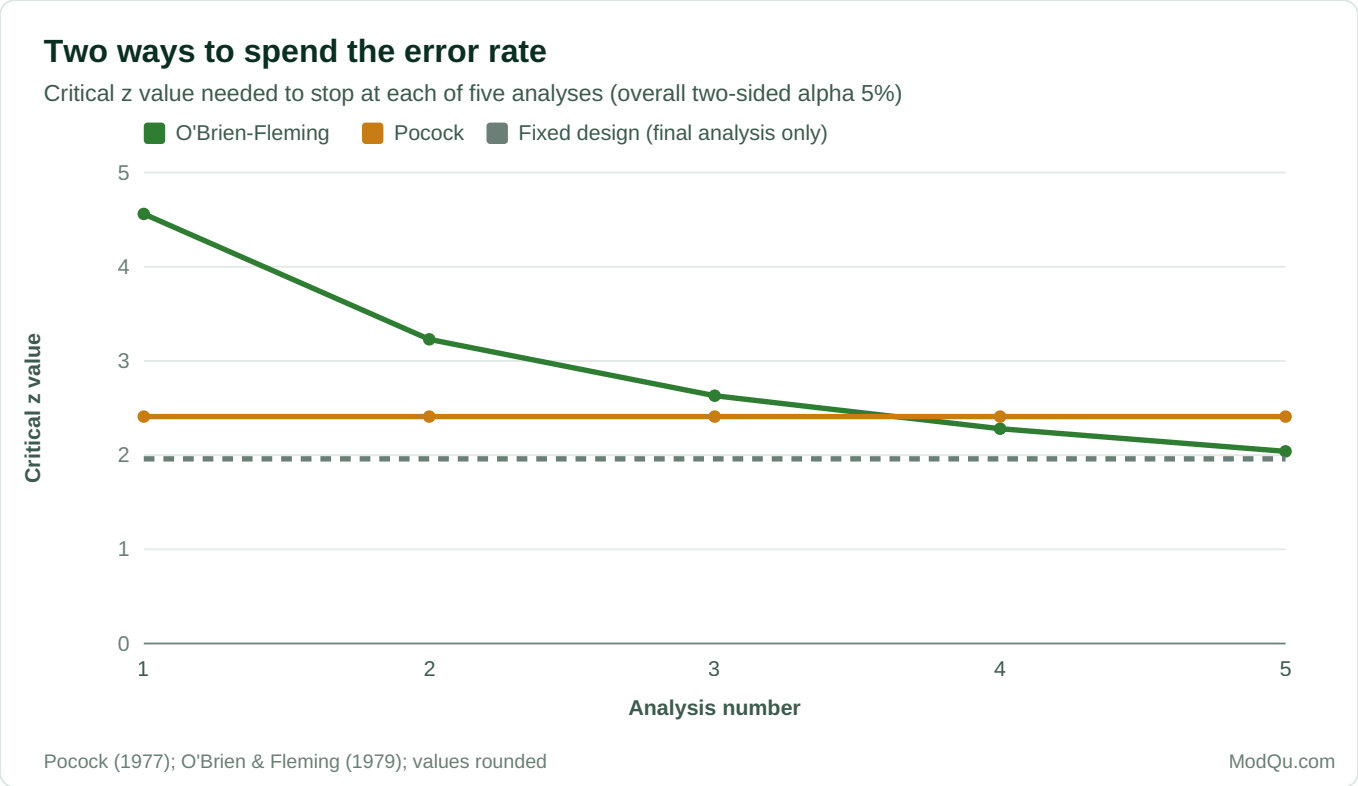


Figure 1. Stopping boundaries for a study with five equally spaced analyses and an overall two-sided false positive rate of 5%. Standard published values for the designs of Pocock (1977) and O'Brien and Fleming (1979), rounded; chart by ModQu.

Lan and DeMets (1983) made the approach more flexible with the **alpha spending function**, which specifies how much of the total error rate is used up as the study progresses rather than fixing the exact timing of each look. This allows interim analyses to be held when convenient, for example at scheduled committee meetings, without breaking the error control.

Group sequential designs can stop for **efficacy**, when the evidence of benefit is overwhelming; for **harm**, when the treatment is clearly worse; or for **futility**, when the study is unlikely to show a benefit even if it continues to the end. Stopping for futility saves resources and spares participants an intervention that is not going to work.

Other types of adaptation

Group sequential designs change only when the study ends. Other adaptive designs change the study itself. Bhatt and Mehta (2016) and Pallmann et al. (2018) describe the main types, summarised below.

Types of adaptation

Each is planned in advance and triggered by interim data

Adaptation	What changes	Example
Early stopping	When the study ends	Stop for clear benefit, harm or futility
Sample size re-estimation	How many participants	Increase the sample if variability is higher than assumed
Dropping or adding arms	Which options are tested	Stop testing a concept that is clearly weaker
Adaptive enrichment	Who is recruited	Focus on the subgroup that benefits
Response-adaptive allocation	Share assigned to each group	Send more participants to the better option
Seamless phases	How stages connect	Select and confirm in one study

Based on Bhatt & Mehta (2016); Pallmann et al. (2018) ModQu.com

Figure 2. Common types of adaptation, what each changes and an example. Based on Bhatt and Mehta (2016) and Pallmann et al. (2018); table by ModQu.

Sample size re-estimation revisits the planned sample size at an interim point. Sample size calculations depend on assumptions, such as the variability of the outcome, that may prove wrong. If the interim data show more variability than expected, the sample can be increased so that the study keeps its intended power. Bauer and Köhne (1994) showed how results from the stages before and after such a change can be combined in a way that preserves the overall error rate.

Dropping or adding arms allows a study that compares several treatments to stop testing those that are clearly not working and concentrate participants on the more promising ones. **Adaptive enrichment** narrows recruitment to a subgroup, such as patients with a particular characteristic, if interim results suggest that the benefit is limited to them. **Seamless designs** combine two phases of research, such as selecting a dose and confirming its effect, in one study without the gap that usually separates them.

Response-adaptive randomisation changes the proportion of participants assigned to each group as results accumulate, so that more are allocated to the treatments that are performing better. Berry (2006) argued that Bayesian methods, which update the probability that each treatment is best as data arrive, are well suited to this kind of design.

Platform trials

The most ambitious adaptive designs are **platform trials**, which test several treatments under a single **master protocol**, add new treatments as they become available and drop those that fail, all within one continuing infrastructure. Woodcock and LaVange (2017) described master protocols, including umbrella trials (several treatments for one disease), basket trials (one treatment for several diseases) and platform trials.

The RECOVERY trial in the United Kingdom showed the value of this approach during the COVID-19 pandemic. Using a simple shared protocol across many hospitals, it tested several treatments in parallel and added or dropped them as evidence accumulated. Among patients receiving oxygen or mechanical ventilation, it found that dexamethasone reduced deaths within 28 days: 22.9% of patients in the dexamethasone group died, compared with 25.7% in the usual care group (RECOVERY Collaborative Group, 2021). Several other treatments were shown to have no benefit in the same platform, which was itself valuable evidence.

Adaptive treatment strategies

A different kind of adaptation concerns not the study but the treatment. In many areas, such as mental health, addiction or education, the right next step depends on how a person has responded so far. Murphy (2005) proposed the **sequential multiple assignment randomised trial** (SMART), in which participants are randomised at the start and then randomised again at later decision points, depending on their response. The design allows researchers to compare whole sequences of decisions, for example "start with A; if there is no response by week eight, switch to B", rather than single treatments. Collins et al. (2007) placed SMART within a wider framework for building and optimising interventions in stages.

Responsive and adaptive survey designs

Surveys have developed their own forms of adaptation. Groves and Heeringa (2006) proposed **responsive design**, in which a survey is run in phases and data collected during fieldwork, such as contact attempts, response rates by subgroup and interviewer workload, are used to change the protocol in later phases. A survey that finds younger men responding at low rates, for example, might shift effort to that group, change the mode of contact or offer a larger incentive in the next phase. The aim is to control both cost and error, particularly nonresponse bias.

Schouten et al. (2017) developed this into **adaptive survey design**, in which different groups of sample members receive different data collection strategies from the start, based on information known in advance about how they are likely to respond. Both approaches require clear rules about what indicators will be monitored and what changes they will trigger, set out before fieldwork begins.

Bandits and adaptive online experiments

Online businesses often face a choice between learning and earning. A standard A/B test divides traffic evenly between versions until the test ends, which means that half the users see the weaker version for the whole test. **Multi-armed bandit** designs shift traffic towards the better-performing version as evidence accumulates. The idea dates back to Thompson (1933), who proposed assigning each option with a probability equal to the probability that it is the best. Scott (2010) showed how this approach, now known as **Thompson sampling**, can be applied to online experiments, reducing the cost of experimentation when there are many versions to compare.

Bandits are well suited to short-term optimisation, such as choosing among headlines or offers, where the aim is to maximise results during the test itself. They are less suited to precise estimation of how large an effect is, or to decisions with long-term consequences, because the uneven allocation makes standard analysis more difficult and early random fluctuations can steer traffic away from an option too soon. Kohavi et al. (2020) recommend that most product decisions still rest on well-powered, fixed-allocation experiments with pre-specified decision rules.

Emergent design in qualitative research

Qualitative research has long allowed designs to change as the study proceeds. Lincoln and Guba (1985) described **emergent design** as a defining feature of naturalistic inquiry: because the researcher does not know in advance what will matter most, the questions, participants and settings are refined as understanding grows. Theoretical sampling in grounded theory, in which each new interview or site is chosen to develop the concepts emerging from the analysis so far, is a structured version of the same idea.

The safeguards here are different from those in statistical designs, but the principle is similar. The researcher should record each change, the reason for it and the point at which it was made, so that readers can follow how the findings developed. An audit trail of this kind allows emergent design to be flexible without becoming arbitrary.

Risks and safeguards

Adaptive designs bring specific risks. The first is the inflation of false positives already described, which is controlled through statistical methods that account for interim looks. The second is **operational bias**. If the people running the study learn interim results, their behaviour may change: recruiters may enrol different participants, clinicians may treat groups differently, and analysts may make different choices. Regulatory guidance therefore emphasises that interim results should be seen only by an independent committee, and that the rules for adaptation should be specified in advance (U.S. Food and Drug Administration, 2019).

The third risk is that estimates of effect size become biased. A study that stops early because results look very good is more likely to have stopped at a random high point, so the estimated effect tends to be larger than the true one. Pallmann et al. (2018) recommend using adjusted estimation methods and reporting the adaptation clearly so that readers can interpret the results.

The fourth risk is complexity. Adaptive designs need more planning, more simulation to check their properties, more sophisticated data systems and faster data flow than fixed designs. Wason et al. (2019) argued that adaptive designs are not always useful. When the outcome takes a long time to observe, when recruitment is fast enough that most participants are enrolled before interim data are available, or when the team lacks the infrastructure to act on interim results quickly, a simpler fixed design may be better.

The common safeguards are:

- specify every possible adaptation, its trigger and its timing before the study begins;
- check by simulation that the design controls error rates and has adequate power under a range of scenarios;
- restrict access to interim results to an independent group;
- document every adaptation actually made, with the data that triggered it;
- report the design, the adaptations and the methods used to adjust the analysis.

Interpreting results from an adaptive study

Results from adaptive studies need careful reading. Chow and Chang (2008) noted that when a study changes its eligibility criteria, its doses or its outcome measures part-way through, the population and treatment being studied at the end may differ from those at the start, so the final result may not answer exactly the question that was originally posed. A study that enriches its sample towards one subgroup, for example, provides strong evidence about that subgroup but weaker evidence about the population originally targeted.

Three questions help a reader judge an adaptive study. First, were the adaptations and their triggers specified before the data were seen, and is the original plan available? Second, did the analysis account for the interim looks and the changes, using methods that keep the error rate at the stated level? Third, who knew the interim results, and could that knowledge have influenced recruitment, delivery or measurement? A study that answers all three clearly can be trusted as much as a fixed design, and often reaches its answer sooner. A study that cannot answer them should be treated with caution, however striking its results.

The same questions apply to commercial work. A report that says "we changed the sample halfway through because the early results looked interesting" describes an unplanned adaptation, and its conclusions are weaker than they appear. A report that says "as planned, we dropped two concepts at the interim review, under the rule agreed at the outset" describes a sound adaptive design.

Adaptive designs in applied research

The ideas behind adaptive design apply well beyond clinical trials. Consider a hypothetical customer research programme testing four new service concepts. The example is illustrative. A fixed design would show all four to equal samples until the end. An adaptive design might plan one interim review after a third of the sample, with pre-specified rules: any concept whose appeal score is more than ten points below the leader at that stage is dropped, and its share of the remaining sample is reallocated to the others. The team gains precision on the concepts that matter, at no extra cost. The rule is written down in advance, the interim analysis is run by someone outside the product team, and the report states that the design was adaptive and what changed.

Similarly, a tracking survey that monitors response rates by region during fieldwork, and has agreed in advance which steps it will take if a region falls behind, is using a responsive design. The value comes not from the flexibility alone but from the combination of flexibility and pre-commitment.

Conclusion

Adaptive designs let a study learn from its own data while it runs. They can stop ineffective or harmful treatments early, correct mistaken assumptions about sample size, concentrate resources on promising options, and direct effort towards groups that are hard to reach. Their validity depends on discipline: every adaptation is planned in advance, error rates are controlled for interim looks, interim results are protected from those running the study, and every change is documented and reported. Used in this way, adaptive designs make research faster and more efficient without weakening the evidence it produces.

References

Armitage, P., McPherson, C. K., & Rowe, B. C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society: Series A (General)*, *132*(2), 235–244. https://doi.org/10.2307/2343787

Bauer, P., & Köhne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics*, *50*(4), 1029–1041. https://doi.org/10.2307/2533441

Berry, D. A. (2006). Bayesian clinical trials. *Nature Reviews Drug Discovery*, *5*(1), 27–36. https://doi.org/10.1038/nrd1927

Bhatt, D. L., & Mehta, C. (2016). Adaptive designs for clinical trials. *New England Journal of Medicine*, *375*(1), 65–74. https://doi.org/10.1056/NEJMra1510061

Chow, S.-C., & Chang, M. (2008). Adaptive design methods in clinical trials: A review. *Orphanet Journal of Rare Diseases*, *3*, Article 11. https://doi.org/10.1186/1750-1172-3-11

Collins, L. M., Murphy, S. A., & Strecher, V. (2007). The multiphase optimization strategy (MOST) and the sequential multiple assignment randomized trial (SMART): New methods for more potent eHealth interventions. *American Journal of Preventive Medicine*, *32*(5, Suppl.), S112–S118. https://doi.org/10.1016/j.amepre.2007.01.022

Groves, R. M., & Heeringa, S. G. (2006). Responsive design for household surveys: Tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *169*(3), 439–457. https://doi.org/10.1111/j.1467-985X.2006.00423.x

Johari, R., Koomen, P., Pekelis, L., & Walsh, D. (2017). Peeking at A/B tests: Why it matters, and what to do about it. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1517–1525). Association for Computing Machinery. https://doi.org/10.1145/3097983.3097992

Kohavi, R., Tang, D., & Xu, Y. (2020). *Trustworthy online controlled experiments: A practical guide to A/B testing*. Cambridge University Press. https://doi.org/10.1017/9781108653985

Lan, K. K. G., & DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika, 70*(3), 659–663. https://doi.org/10.1093/biomet/70.3.659

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. SAGE.

Murphy, S. A. (2005). An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine, 24*(10), 1455–1481. https://doi.org/10.1002/sim.2022

O'Brien, P. C., & Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics, 35*(3), 549–556. https://doi.org/10.2307/2530245

Pallmann, P., Bedding, A. W., Choodari-Oskooei, B., Dimairo, M., Flight, L., Hampson, L. V., Holmes, J., Mander, A. P., Odondi, L., Sydes, M. R., Villar, S. S., Wason, J. M. S., Weir, C. J., Wheeler, G. M., Yap, C., & Jaki, T. (2018). Adaptive designs in clinical trials: Why use them, and how to run and report them. *BMC Medicine, 16*, Article 29. https://doi.org/10.1186/s12916-018-1017-7

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika, 64*(2), 191–199. https://doi.org/10.1093/biomet/64.2.191

RECOVERY Collaborative Group. (2021). Dexamethasone in hospitalized patients with Covid-19. *New England Journal of Medicine, 384*(8), 693–704. https://doi.org/10.1056/NEJMoa2021436

Schouten, B., Peytchev, A., & Wagner, J. (2017). *Adaptive survey design*. CRC Press. https://doi.org/10.1201/9781315153964

Scott, S. L. (2010). A modern Bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry, 26*(6), 639–658. https://doi.org/10.1002/asmb.874

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika, 25*(3–4), 285–294. https://doi.org/10.1093/biomet/25.3-4.285

U.S. Food and Drug Administration. (2019). *Adaptive designs for clinical trials of drugs and biologics: Guidance for industry*. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adaptive-design-clinical-trials-drugs-and-biologics-guidance-industry

Wald, A. (1945). Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics, 16*(2), 117–186. https://doi.org/10.1214/aoms/1177731118

Wason, J. M. S., Brocklehurst, P., & Yap, C. (2019). When to keep it simple: Adaptive designs are not always useful. *BMC Medicine, 17*, Article 152. https://doi.org/10.1186/s12916-019-1391-9

Woodcock, J., & LaVange, L. M. (2017). Master protocols to study multiple therapies, multiple diseases, or both. *New England Journal of Medicine, 377*(1), 62–70. https://doi.org/10.1056/NEJMra1510062