Statistical Science
2023, Vol. 38, No. 2, 233–239
https://doi.org/10.1214/23-STS865REJ
Main article: https://doi.org/10.1214/22-STS865
This research was funded, in whole or in part, by [UK Medical Research Council, MC_UU_00002/14].

Rejoinder: Response-Adaptive Randomization in Clinical Trials

David S. Robertson, Kim May Lee, Boryana C. López-Kolkovska and Sofía S. Villar

We are very grateful to the editors of *Statistical Science* and the discussants for the opportunity to mark the finish line of our paper (Robertson et al., 2023c) with such a high note of clarity and optimism toward the future. Each discussant takes aspects briefly mentioned in our paper and expands on them with deeply insightful and thought-provoking comments. The combination of paper, discussions and rejoinder helps give the broad, balanced and fresh resource we were longing for in the response-adaptive randomization (RAR) literature.

RAR as One of Many Elements of an Adaptive Design

An important assumption we made in our paper was that RAR was the only adaptation used in an experiment, reflecting our main goal of identifying the effects of RAR on different aspects of clinical trial design and analysis. Hence, we did not address the use of RAR in combination with other forms of adaptation like early stopping or population enrichment. However, when the objective is to obtain an overall "best" trial design (rather than quantifying effects of individual components) then a combination of adaptive elements is likely to be the most effective way to achieve this, as illustrated by three of the discussants.

Jennison describes a procedure to define RAR, first introduced in Jennison and Turnbull (2000), in which the optimal sampling ratio (n_1/n_2) is $a^{\hat{\theta}/2\delta}$ where $\hat{\theta}$ is the current treatment effect estimate, δ is the anticipated treatment effect and a is a chosen constant. Jennison quantifies the potential advantages of such a procedure to reduce the expected number of patients receiving the inferior treatment, with the results in Tables 1 and 2 corresponding to RAR without early stopping. In the same setting, Jennison elegantly addresses other major issues we discussed in our paper, such as valid inference and the effects of temporal trends and delayed responses. Jennison also shows how early stopping (using a group sequential approach to define stopping boundaries) in combination with RAR can be more effective in minimizing the expected number of patients on inferior arms (and also reducing the total expected sample size). Similarly, the examples in practice presented by Berry and Viele and the future perspective of Trippa and Xu illustrate this idea that the design as a whole is greater than the sum of its parts.

RAR procedures defined as in Jennison require the prespecification of a, which controls the level of imbalance. As well, by using a $1:1$ ratio for the first block, such procedures include a *burn-in* period before adaptive sampling is used. These are caveats to consider when comparing designs like this to RAR with no burn-in or no control over the maximum level of imbalance (as these can impact operating characteristics). More generally, when comparing different types of trial adaptations, the distinction between comparing candidate designs against an objective versus identifying the effects of individual adaptive elements may be important. In Jennison's example, with a normal endpoint and common variance, the use of RAR as the single adaptation to reduce the number of patients treated with an inferior arm requires a larger total sample size to achieve the same power as equal randomization. However, trade-offs also exist even when considering the inclusion of early stopping for a fixed $1:1$ randomization ratio, which achieves reductions in expected sample size but requires a larger maximum sample size.

Duan, Müller and Jin ask us: *"Do the authors recommend practitioners to use RAR? In which cases?"* Given that RAR considered in isolation may not give the "best" trial design, the answer depends on whether the superior trial design (according to the goals of the study) includes RAR as one component. This suggests the question of practical importance is determining when the "price" of a design element is worthwhile overall. Jennison alludes to this when pointing out that using a simpler fixed 3:2 ratio in a group sequential design may be comparable to a design with RAR and early stopping. We expand on our answers to these questions in the section after the next one.

RAR as Nonmyopic Solutions to Sequential Designs

We welcome the presentation by Duan, Müller and Jin of RAR as a computationally feasible solution to an optimal sequential problem, though we note that not all

David S. Robertson is a Senior Research Associate, MRC Biostatistics Unit, University of Cambridge, Cambridge, UK (e-mail: david.robertson@mrc-bsu.cam.ac.uk). Kim May Lee is a Research Fellow, King's College London, London, UK (e-mail: kim.lee@kcl.ac.uk). Boryana C. López-Kolkovska is a Statistical Science Associate Director at AstraZeneca, Cambridge, UK (e-mail: boryana.kolkovska@astrazeneca.com). Sofía S. Villar is an MRC Investigator, MRC Biostatistics Unit, University of Cambridge, Cambridge, UK (e-mail: sofia.villar@mrc-bsu.cam.ac.uk).

RAR may be suitably described as such. We appreciate their recognition of the challenges associated with deriving good approximations to solve such problems, as well as their proposal to reduce computational cost by replacing these problems with a simpler nonsequential one of finding a "near-optimal" boundary. As they point out, the RAR literature heavily focuses on myopic approaches because of the "curse of dimensionality" of optimal solutions. For example, setting $c_2 = c_3 = 0$, $K_2 = 1$, $K_3 = 0$ in equation (2) in Duan, Müller and Jin and allowing $\mathcal{A}_t = \{0, 1\}$ to be a (binary) treatment decision variable for just two doses and cohorts of 1 patient recovers the classical two-armed bandit problem as in Berry (1978). Press (2009, Figure 1) shows the topology of this problem to be well approximated by a boundary defined by two metrics corresponding to statistical significance and sample size imbalance.

A key point about the computational complexity of optimal solutions is that two-armed sequential problems have become increasingly tractable over the years. For example, Berry (1978, Table 2) reports optimal value functions for a two-armed bandit problem with a maximum size of $N = 100$ (i.e., 100 treatment decisions) while four decades later Jacko (2019, Table 2) reports $N = 4440$ obtained on a computer with 32 GB RAM.[1] More practically, Jacko (2019, Figure 1) shows finding the next optimal treatment decision for $N = 1200$ only requires 1 GB RAM and 10 min to run. For the multiarm setting, where the potential of RAR is the largest (as agreed by Trippa and Xu, Ivanova and Rosenberger and illustrated by the examples in practice of Berry and Viele), the computational bottleneck remains pressing. A recent approach to the design of nonmyopic RAR that remains computationally tractable was introduced by Villar, Wason and Bowden (2015) and noted by Ivanova and Rosenberger. The computational advantage of such approaches relies on the use of Gittins (or Whittle) indices which deal with computational barriers via a 'divide and conquer' approach (see Villar, Bowden and Wason, 2015, Figure 1).

We see many advantages of optimal designs such as the one presented by Duan, Müller and Jin. Adding restrictions directly into the formulation, as they do in their example, is one of them. The approach in Cheng and Berry (2007) and Williamson et al. (2017) illustrates the addition of a restriction on the minimal sample size per arm (which may be a preferable way to indirectly reduce undesirable imbalances while increasing power). While it is possible to enforce a restriction to prevent sample size imbalance (as Duan, Müller and Jin say, forcing it to be less than 10%) this also implies a strict limit to the sample size in the best arm (e.g., to never be larger than 90%). The reason for the *zeroes* (for imbalance in the wrong direction beyond 10%) in Table 1 of our paper is not caused by the inclusion of such a constraint but rather linked to two points we could have emphasized more, as noted by Giovagnoli: (i) the distinction between deriving an optimal target and determining how to effectively target it and (ii) the degree of randomness of RAR procedures used to target an optimal ratio (linking to metrics of the "amount of randomness").

ERADE is the fastest converging implementation algorithm for a given target ratio (e.g., the one given by equation (6) in our paper). In Table 1 of our paper, this corresponds to a final ratio of $N_1/n \approx 0.54$. If by the end of the trial ERADE attains an overall allocation very close to that ratio, then $N_1 - N_0 \approx 0.08n$ and imbalances larger than 10% in the wrong direction (i.e., $N_0 - N_1 > 0.1n$) will be almost impossible. This also illustrates the arbitrariness of such metrics for imbalance: if the threshold was set to be smaller than 10%, there would be fewer zeroes in Table 1.[2] An additional point is that these zeroes suggest high predictability of the randomization sequence as well as fast convergence (Berger et al., 2021). Such trade-offs are common between different objectives and metrics, as we discuss next.

RAR: Objectives, Metrics and Benchmarking Designs

Perhaps the largest advantage of optimal designs is the formal incorporation of a study's objective(s) into the problem formulation. As Duan, Müller and Jin illustrate, optimal designs naturally allow explicitly linking phase II with phase III trial considerations (in their example, through the p_D variable). The challenge of suitably linking phases to increase overall success in the drug development process means the phase II trial design choice may need to go beyond the aims of this phase in isolation, as echoed by Trippa and Xu. When a study has multiple key objectives, an optimal design (which may or may not include RAR as part of it) can find solutions to best balance these. However, this relies on the specification of an appropriate objective function. For example, Duan, Müller and Jin's equation (2) represents a utility function that when optimized returns a design minimizing recruitment costs and maximizing the expected number of patients in the best arm during the phase II trial as well as the probability of a successful confirmatory phase III trial. In our experience, we have witnessed the difficulty of both articulating objectives and agreeing on to how to weigh them appropriately. How would we determine at the design stage how large K_2 (in Duan, Müller and Jin) or a (in

[1] When the optimal decisions and not just the final value function are needed (i.e., an offline implementation), a similar computer is able to solve the problem for $N = 1440$.

[2] This reasoning is consistent with DBCD being slower to converge to the target allocation than ERADE, which is noticeable for $n = 200$, but for $n = 654$, DBCD has essentially converged.

Jennison) should be? Such difficulties may explain why it is common to enforce restrictions to existing designs (to limit undesirable properties) instead of formulating designs based on utility/loss functions. Pitt (2021) discusses this in the context of Phase I dose-finding studies.

Similar considerations apply to the choice of metrics to accurately match the design objectives. For example, what should the threshold level be for the imbalance metrics in Section 3.1 of our paper? More importantly, exactly what objective is this metric trying to capture and is it the best reflection of that aim? And how would this objective interact with others? Perhaps agreeing on metrics is easier than making objectives explicit; nonetheless, those metrics that will drive design choice deserve careful consideration. The latter point is nicely illustrated in dose-finding studies where a typical metric is the proportion of correct selections of doses. In that context, benchmarking is presented as a way to compare designs using an optimal benchmark corresponding to a (theoretical) upper bound on the performance of a design under a given scenario. This promotes a more accurate evaluation of dose-finding designs (Mozgunov, Jaki and Paoletti, 2020, 2022).

Duan, Müller and Jin refer to benchmarking as evaluating the performance of a simpler RAR procedure against a relevant (upper or lower) performance bound from an optimal RAR design under a given scenario. Jennison's comparison of a simpler group sequential design to the full RAR + group sequential approach (Tables 3–5) represents a similar form of benchmarking to find simpler designs with a similar performance on the key objectives. Meanwhile, Trippa and Xu mention the need for *"new and improved methods for accurate, comprehensive and context-specific assessments of candidate designs."* Our view of benchmarking[3] is perhaps more aligned with that of Trippa and Xu, in the sense that given the key objectives of the study (which are context dependent, e.g., the phase of the trial, prevalence of the disease) and suitable metrics to measure these across designs (including simplicity of the design), then all candidate designs can be fairly and accurately assessed against each other. Those assessments would involve well-performed simulations to guide design choice (as one of many aspects that may require simulations, as we discuss below). Our view is that such benchmarking holds the answer to the question by Duan, Müller and Jin on when RAR (as part of a design) may fit well with a trial's context and goals.

Aside from the choice of objectives and metrics, the choice of the set of candidate designs to consider is also not necessarily straightforward. The set would naturally include optimal (or near-optimal) designs if these are known or can be derived. The set would also include simpler (e.g., nonadaptive) designs as feasible comparators, such as a fixed randomization design. The comparators will crucially depend on the trial context. For example, in the phase II setting, *nonrandomized* single-arm trials are commonplace and could be reasonably included as a comparator. We note in passing that while RAR may be criticized in the phase III context for moving away from the (arguable) equipoise of fixed equal randomization, in the phase II setting, the fact that RAR allows for randomization at all is already a major statistical advantage when compared with single-arm trials.

The Role of Simulations and Data Sharing

One key aspect featured by all the discussants is the use of simulations for designing and evaluating complex trial designs, including (but not limited to) those that use RAR. Giovagnoli discusses the lack of formalization and guidelines to conduct simulation studies, which leads to concerns around reproducibility and selection bias, as well as inconclusive and potentially contradictory simulation studies. Similarly, Ivanova and Rosenberger point out it is possible to find specific simulation scenarios that show a RAR procedure does *not* work well (or indeed, *does* work well), which echoes concerns raised by Pawel, Kook and Reeve (2023). Duan, Müller and Jin discuss using "benchmarks" for simulation studies evaluating RAR, while Jennison uses simulations with group sequential designs as comparators. Berry and Viele discuss how calibrating and understanding the properties of complex Bayesian designs require "*proper exploration and simulation*," which were used to "*extensively hone*" the designs of their example RAR trials in practice. Trippa and Xu agree on the importance of comprehensive simulations to compare candidate designs and note how methodological and computational advances have made this task much easier. Crucially, they add the dimension of data sharing as a key component of rigorous and context-specific simulation comparisons. They envision prospectively planned simulation studies of candidate designs using data generated from multiple randomized trials, including evaluating the use of adaptive designs such as RAR in early phases of the drug development process on later phases.

We share the general concerns raised by the discussants around the need for specific guidelines and structured approaches around conducting simulation studies for evaluating trial designs, and agree that the current state of the debate around RAR is partially explained by the sometimes inadequate reporting of simulation studies of RAR designs. However, this issue is not unique to RAR as noted by recent relevant work. One example is Morris, White and Crowther (2019), also mentioned in the Discussion of our paper and by Giovagnoli. This may appear to be focused on simulations of statistical methods

[3]The term "benchmarking" has a different interpretation in machine learning applications. There, benchmarking refers to comparing competing methods/algorithms on a common, "gold-standard" *data set*.

for data analysis, but the principles can be applied more broadly to simulations for trial designs (personal communication with the author), as evidenced by Section 3.4: "*The term "method" is generic. Most often it refers to a model for analysis, but might refer to a design or some procedure (such as a decision rule).*" Another useful reference is Mayer et al. (2019), which provides insights from industry on simulations for adaptive designs. More broadly, simulations form a key part of a recently proposed framework for the phases of methodological research in biostatistics (Heinze et al., 2022).

The challenges of conducting simulation studies become evident in Section 3.1 of the paper. This was a response to the lack of reporting of imbalance metrics in simulation studies for RAR other than for Thompson sampling (and its variants), and we aimed to investigate this for other types of RAR. We focused on a few scenarios, but (contrary to the comment by Giovagnoli) we did not only report results for $p_0 = 0.25$ and $p_1 = 0.35$ (see Figure 2 and Table A2). For the Randomized Play-the-Winner rule in particular, as pointed out by Giovagnoli, the performance looks different for other points in the parameter space (namely, where $p_0 + p_1 > 3/2$). For example, when $n = 200$, $p_0 = 0.75$ and $p_1 = 0.85$, the imbalance metric $P(N_0 > N_1 + 0.1n) = 0.142$. This highlights the importance of exploring a wide range of the parameter space, as well as Ivanova and Rosenberger's point on how a design may look good or bad for a given metric by underreporting of scenarios.

Looking to the future, we agree with Trippa and Xu that continued computational and methodological advances will greatly aid the systematic evaluation of complex trial designs such as RAR. One such example is the recent innovative methodological work by Sklar (2022), who proposes a rigorous framework for simulation-based verification of adaptive design properties, including type I error rate control. We also echo the point by Trippa and Xu about the importance of data sharing of clinical trials, which has many benefits including allowing the evaluation of candidate designs on multiple real-world data sources; see, for example, the commentary by Law et al. (2022).

RAR in Current Practice

We thank Berry and Viele for providing a detailed summary of six successful examples of RAR trials in practice (some of which are still ongoing). This valuable snapshot of RAR provides a practical view that uniquely complements our paper's methodological focus, illustrating how with careful choice of the design elements and for the right trial context, RAR can be operationally feasible and greatly beneficial in meeting trial objectives. Nonetheless, Berry and Viele also note that RAR is no panacea and that they have been involved with more equally randomized trials than those with RAR. Some common features of the RAR examples include:

- Phase II trials evaluating multiple treatments/doses, with the selected dose(s) then evaluated in phase III
- A goal of identifying the single best active arm
- A "burn-in" stage using fixed (but not necessarily equal) randomization to the treatments and control
- Combination of RAR with early stopping rules
- RAR probabilities being updated in groups
- Maintaining or even increasing the allocation to the control compared with equal randomization
- Bayesian RAR that goes far beyond the vanilla Thompson sampling implementation

As Berry and Viele point out, these trials have additional innovative features besides RAR, such as the ability to add treatment arms seamlessly during the trial. The common features above give concrete examples of when RAR may be used by practitioners, as asked by Duan, Müller and Jin.

The Bayesian RAR used in practice is more complex than a simple application of Thompson sampling, which (as noted by Giovagnoli) may not have sensible properties. As discussed by both Ivanova and Rosenberger and Giovagnoli, other types of Bayesian RAR (distinct from Thompson sampling and its generalizations) have been developed, for example, the Bayesian biased coin design of Xiao, Liu and Hu (2017). Berry and Viele note that the use of (Bayesian) RAR typically reduces the type I error rate. In our experience, this feature depends on the primary outcome type (e.g., binary or continuous), the hypothesis testing procedure and the stopping rules used. For example, Smith and Villar (2018) find that RAR procedures including Bayesian RAR can inflate the type I error rate. More generally, the interplay between RAR, arm selection and early stopping can be complex; see, for example, Shin, Ramdas and Rinaldo (2019).

Finally, a notable feature of the AWARD-5 example reported by Berry and Viele is the rigor in terms of blinding and the logistical ease of the automatic updates to the randomization probabilities. Such an algorithmic blinded implementation gives an insight into how "*experienced groups may seamlessly update RAR probabilities quickly . . . without interrupting other trial processes*." Clearly, experience and capacity are key to realizing this, which naturally brings us to the future outlook for RAR designs.

The Future to Come for RAR

We share the optimism of Trippa and Xu around the bright outlook for RAR designs (and adaptive trials more generally) to be further developed and increasingly used in practice in the future. First, as Trippa and Xu point out, and as seen in the trial examples already used in practice by Berry and Viele, the growing use of multiarm and platform trials provide new opportunities for RAR to be applied successfully. An interesting feature of such designs is that in some trial settings, there may not be a

natural "control" arm to include (e.g., where no standard of care currently exists). Second, Trippa and Xu point to the advancement in methods for rigorously evaluating candidate trial designs in the appropriate context, which we feel is a key part of the "benchmarking" idea discussed earlier. Third, Trippa and Xu point to the rapid advancement of the use of biomarkers (e.g., genomic information) to inform precision medicine trials, including adaptive enrichment designs. These trials aim to match the best-suited treatments to patient subgroups defined by biomarkers. Ivanova and Rosenberger ask us if the addition of RAR can enhance the individualized ethical considerations of such trials. Our answer is that enrichment or biomarker-stratified designs using RAR could enable more efficient and timely decision-making as well as a larger expected number of patients treated with their best responding arm. Implementing RAR in such designs is likely to be paired with features such as early stopping rules for enrichment and Bayesian hierarchical models to account for between-subgroup heterogeneity. For example, Ventz et al. (2017) present a class of Bayesian RAR procedures and report the allocation to a superior arm per cancer type increases from that of balanced randomization by 10–32%. We envisage that the potential efficiency gains brought by RAR, when combined with principled information borrowing (Zheng and Wason, 2022, Ouma et al., 2022) or enrichment strategies, can help deliver the full potential of precision medicine trials evaluating multiple treatment arms in multiple patient subgroups. Such trial settings, as well as multiarm and platform trials, provide contexts where RAR may be increasingly used in the future, again helping answer the question asked by Duan, Müller and Jin.

The trials reported by Berry and Viele also give a glimpse into what the future of complex trial designs (including RAR as a possible component) could look like as groups gain experience and confidence in applying RAR and other trial adaptations in practice. As complex trial designs become more routinely used, a preference for simplicity of a design may evolve, with design features such as (well-understood) RAR procedures being seen as an increasingly logistically feasible option to consider. Hence, the difference in complexity between a group sequential design with a fixed 3:2 sampling ratio and a RAR procedure updated in groups, as considered by Jennison, may be viewed as less influential as some of the costs of implementing complex designs decrease. This helps answer the question Jennison posed around when a more symmetric attitude to the assumed treatment effect around zero may be justified. Depending on the trial context, an assumption that the new treatment is at least as effective as the control may be difficult to justify (e.g., when there is no standard of care), and (all else being equal) using a RAR procedure that does not depend on it may be preferable.

Further Questions Increased Uptake Might Bring

As noted by many of the discussants, and with the prospect of the increased uptake of RAR designs in the future, there remain potential barriers and further questions to answer. Ivanova and Rosenberger ask if there is scope to adjust recruitment for ethical purposes. As an example of this, Ventz et al. (2017) comment that the increased expected number of patients on the best arm is higher for the group with the slowest accrual rate, and note that "*this gives the algorithm more time to accumulate information.*" Our view is that recruitment is challenging enough to predict, so we cannot easily imagine manipulating this. However, slowing down recruitment is more feasible than speeding it up (e.g., delaying the opening of new centers in a multicenter study). Perhaps an ideal approach would match the adaptation to the expected pace of recruitment, with this being a key part of the simulations performed to choose a design. The idea of recruiting patients in groups (as mentioned by Ivanova and Rosenberger, utilized in practice by Berry and Viele, and methodologically described by Jennison) is one way to help match adaptations such as RAR to recruitment rates.

Ivanova and Rosenberger also point to the possibility of *accrual bias* (patients choosing to delay consenting into a trial) being an issue. This relates to the above point in that while slowing recruitment may be possible, this may not be desirable both from a sponsor and patient perspective. If the disease is such that it is imperative to seek a potential superior treatment earlier on, patients may have to weigh these two incentives to make a decision. Also, it is an open question as adaptive trials remain in the minority and work is needed to understand what level of information is given to patients at recruitment. We need evidence to determine if this is an issue just as much as we need research on how to best approach patients to enter an adaptive trial to minimize this and other forms of biases.

We agree with Giovagnoli that the "*ground remains fertile and there is ample scope*" for further methodological developments around RAR. We gave our view of some of the key remaining open areas for methods research at the very end of our paper. As we discussed there, one open area is that of efficient and valid inference methods for RAR. This links to the point raised by Giovagnoli about metrics related to "*inference and estimation.*" We regard estimation as a branch of statistical inference, but made the distinction between hypothesis testing and estimation metrics explicit in our paper as the two issues can have very different considerations within the context of adaptive designs; see, for example, Robertson et al. (2023a, 2023b).

Meanwhile, clearly defining objectives and suitable metrics is in our opinion one of the biggest hurdles to overcome so that practitioners can use optimal and complex designs (including those with RAR). Linked with

this, the derivation of optimal designs for complex utility functions remains a difficult methodological and computational challenge. Another area mentioned by Trippa and Xu (and discussed above) is the development of RAR that incorporates biomarker information to be used in the context of precision medicine. Finally, as mentioned by Ivanova and Rosenberger there is a need for software for RAR, so that methodology can be implemented in practice. This links to the idea from Berry and Viele of the importance of blinded algorithmic implementations to allow RAR to be logistically feasible.

Concluding Remarks

When we started to write our paper, we found that (as Trippa and Xu put it) many "*are often influenced by their experience with a single algorithm.*" Our paper was an attempt to move the debate away from that stale position. In doing so, we did not fully answer some questions (such as Duan, Müller and Jin's question of when to use RAR, or which RAR procedures have been superseded as Giovanoli would like to see in a review), but we have given our views on many others that have been part of a long-lasting debate on the use of RAR in clinical trials. We are delighted to see that the discussion on the topic appears to have evolved in such a promising direction. Ivanova and Rosenberger looked back at the *Statistical Science* review by Rosenberger (1996) in relation to ours and reflected upon what has changed. We look forward with eager expectation to the changes to come in the next 25 years.

ACKNOWLEDGMENTS

The authors thank Peter Jacko, Tim Morris, Pavel Mozgunov, Lukas Pin and Haiyan Zheng for helpful discussion and comments when preparing the rejoinder.

REFERENCES

BERGER, V. W., BOUR, L. J., CARTER, K., CHIPMAN, J. J., EVERETT, C. C., HEUSSEN, N., HEWITT, C., HILGERS, R., LUO, Y. A. et al. (2021). A roadmap to using randomization in clinical trials. *BMC Med. Res. Methodol.* **21** 1–24.

BERRY, D. A. (1978). Modified two-armed bandit strategies for certain clinical trials. *J. Amer. Statist. Assoc.* **73** 339–345.

CHENG, Y. and BERRY, D. A. (2007). Optimal adaptive randomized designs for clinical trials. *Biometrika* **94** 673–689. MR2410016 https://doi.org/10.1093/biomet/asm049

HEINZE, G., BOULESTEIX, A., KAMMER, M., MORRIS, T. P. and WHITE, I. R. (2022). Phases of methodological research in biostatistics-building the evidence base for new methods. *Biom. J.* https://doi.org/10.1002/bimj.202200222

JACKO, P. (2019). The finite-horizon two-armed bandit problem with binary responses: A multidisciplinary survey of the history, state of the art, and myths. arXiv preprint. arXiv:1906.10173.

JENNISON, C. and TURNBULL, B. W. (2000). *Group Sequential Methods with Applications to Clinical Trials*. CRC Press/CRC, Boca Raton, FL. MR1710781

LAW, M., COUTURIER, D., CHOODARI-OSKOOEI, B., CROUT, P., GAMBLE, C., PALLMANN, P., PILLING, M., ROBERTSON, D. S., ROBLING, M., SYDES, M. R., VILLAR, S. S., WASON, J. M. S., WHEELER, G., WILLIAMSON, S. F., YAP, C. and JAKI, T. (2022). Medicines and healthcare products regulatory agency's "Consultation on proposals for legislative changes for clinical trials". Preprint. https://doi.org/10.21203/rs.3.rs-2210654/v1

MAYER, C., PEREVOZSKAYA, I., LEONOV, S., DRAGALIN, V., PRITCHETT, Y., BEDDING, A., HARTFORD, A., FARDIPOUR, P. and CICCONETTI, G. (2019). Pitfalls and potentials in simulation studies: Questionable research practices in comparative simulation studies allow for spurious claims of superiority of any method. *Stat. Biopharm. Res.* **11** 325–335.

MORRIS, T. P., WHITE, I. R. and CROWTHER, M. J. (2019). Using simulation studies to evaluate statistical methods. *Stat. Med.* **38** 2074–2102. MR3937487 https://doi.org/10.1002/sim.8086

MOZGUNOV, P., JAKI, T. and PAOLETTI, X. (2020). A benchmark for dose finding studies with continuous outcomes. *Biostatistics* **21** 189–201. MR4132543 https://doi.org/10.1093/biostatistics/kxy045

MOZGUNOV, P., PAOLETTI, X. and JAKI, T. (2022). A benchmark for dose-finding studies with unknown ordering. *Biostatistics* **23** 721–737. MR4454951 https://doi.org/10.1093/biostatistics/kxaa054

OUMA, L. O., GRAYLING, M. J., WASON, J. M. S. and ZHENG, H. (2022). Bayesian modelling strategies for borrowing of information in randomised basket trials. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **71** 2014–2037. MR4511139 https://doi.org/10.1111/rssc.12602

PAWEL, S., KOOK, L. and REEVE, K. (2023). Pitfalls and potentials in simulation studies: Questionable research practices in comparative simulation studies allow for spurious claims of superiority of any method. *Biom. J.* https://doi.org/10.1002/bimj.202200091

PITT, L. (2021). Optimising first in human trials. PhD Thesis, Univ. Bath. Available at https://researchportal.bath.ac.uk/en/studentTheses/optimising-first-in-human-trials.

PRESS, W. H. (2009). Bandit solutions provide unified ethical models for randomized clinical trials and comparative effectiveness research. *Proc. Natl. Acad. Sci. USA* **106** 22387–22392.

ROBERTSON, D. S., CHOODARI-OSKOOEI, B., DIMAIRO, M., FLIGHT, L., PALLMANN, P. and JAKI, T. (2023a). Point estimation for adaptive trial designs I: A methodological review. *Stat. Med.* **42** 122–145. MR4527724 https://doi.org/10.1002/sim.9605

ROBERTSON, D. S., CHOODARI-OSKOOEI, B., DIMAIRO, M., FLIGHT, L., PALLMANN, P. and JAKI, T. (2023b). Point estimation after adaptive trials II: Practical considerations and guidance. *Stat. Med.* https://doi.org/10.1002/sim.9734

ROBERTSON, D. S., LÓPEZ-KOLKOVSKA, B. C., LEE, K. M. and VILLAR, S. S. (2023c). Response-adaptive randomization in clinical trials: From myths to practical considerations. *Statist. Sci.* **38** 185–208.

ROSENBERGER, W. F. (1996). New directions in adaptive designs. *Statist. Sci.* **11** 137–149.

SHIN, J., RAMDAS, A. and RINALDO, A. (2019). Are sample means in multi-armed bandits positively or negatively biased? *Adv. Neural Inf. Process. Syst.* **32**.

SKLAR, M. J. (2022). Adaptive experiments and a rigorous framework for type I error verification and computational experiment design. PhD Thesis, Stanford Univ. arXiv:2205.09369.

SMITH, A. L. and VILLAR, S. S. (2018). Bayesian adaptive bandit-based designs using the Gittins index for multi-armed trials with normally distributed endpoints. *J. Appl. Stat.* **45** 1052–1076. MR3774531 https://doi.org/10.1080/02664763.2017.1342780

VENTZ, S., BARRY, W. T., PARMIGIANI, G. and TRIPPA, L. (2017). Bayesian response-adaptive designs for basket trials. *Biometrics* **73** 905–915. MR3713124 https://doi.org/10.1111/biom.12668

VILLAR, S. S., BOWDEN, J. and WASON, J. (2015). Multi-armed bandit models for the optimal design of clinical trials: Benefits and challenges. *Statist. Sci.* **30** 199–215. MR3353103 https://doi.org/10.1214/14-STS504

VILLAR, S. S., WASON, J. and BOWDEN, J. (2015). Response-adaptive randomization for multi-arm clinical trials using the forward looking Gittins index rule. *Biometrics* **71** 969–978. MR3436722 https://doi.org/10.1111/biom.12337

WILLIAMSON, S. F., JACKO, P., VILLAR, S. S. and JAKI, T. (2017). A Bayesian adaptive design for clinical trials in rare diseases. *Comput. Statist. Data Anal.* **113** 136–153. MR3662397 https://doi.org/10.1016/j.csda.2016.09.006

XIAO, Y., LIU, Z. and HU, F. (2017). Bayesian doubly adaptive randomization in clinical trials. *Sci. China Math.* **60** 2503–2514. MR3736374 https://doi.org/10.1007/s11425-016-0056-1

ZHENG, H. and WASON, J. M. S. (2022). Borrowing of information across patient subgroups in a basket trial based on distributional discrepancy. *Biostatistics* **23** 120–135. MR4366039 https://doi.org/10.1093/biostatistics/kxaa019