Bayesian adaptive design: the future for regional anesthesia trials?

Robert Gotmaker,[1,2] Michael J Barrington,[2,3] John Reynolds,[1] Lorenzo Trippa,[4] Stephane Heritier[1]

INTRODUCTION

Over the past decade there has been a substantial growth in ultrasound-guided interfascial plane blocks,[1 2] and these procedures are now embedded in routine clinical practice. One example, transabdominis plane (TAP) blockade, was made popular by the availability of ultrasound technology and promoted by enthusiastic clinicians well in advance of supporting evidence. A recent meta-analysis examining the analgesic efficacy of TAP block identified 31 randomized controlled trials (RCTs)[3]; however, this body of literature has taken a decade to accumulate. The evidence and expert consensus is that TAP blocks have minimal postoperative analgesic efficacy in adults, and therefore during the last decade a large number of patients were needlessly exposed to an inefficacious therapy. Several other ultrasound-guided interfascial plane blocks including quadratus lumborum, pectoralis plane and erector spinae appear to be following a clinical trajectory similar to TAP blocks, with anatomical studies and numerous case reports but very few controlled clinical trials.[2] Another problem has been the tendency to perform meta-analysis of a modest number of small RCTs of regional anesthesia procedures with heterogeneous comparator groups and outcomes measures. This may be misleading since a large RCT using the same intervention and comparator would be unlikely to yield the same perioperative treatment effect size.[4 5]

In this Daring discourse, we explore the potential of multiarm Bayesian adaptive trial designs (BATD) to accelerate discovery in regional anesthesia. Although infrequently used in perioperative care, these methods are gaining traction, especially in oncology,[6–8] and we believe they are also well suited to studying the numerous emerging interfascial plane blocks.

Adaptive trials alter key-study characteristics according to accruing data and can be more efficient compared with classical parallel-arm trials.[9] Efficiency is gained by addressing multiple treatments simultaneously with a shared contemporaneous control, increasing allocation to promising treatments (adaptive randomization) or stopping early for futility or efficacy according to predetermined rules. More complex designs allow for the addition of a new treatment (eg, a new version of an existing block) when that treatment becomes available[10] or the enrichment of the study population with subjects in whom the intervention appears promising. Adaptive designs are particularly advantageous when there are multiple competing therapies for a given clinical condition and early outcome measures are available to drive adaptations. This is precisely the situation in regional anesthesia where we have a large number of potential therapies available for a given surgical procedure and outcomes are measured early. Although frequentist adaptive designs are possible, they are generally less flexible than Bayesian approaches.[11] The frequentist operating characteristics (eg, false positive rate, true positive rate) of a Bayesian design can be determined through computer simulation, and design parameters can be chosen to achieve given targets, for example prespecified type I error.[12]

Our objective is to illustrate the potential advantages of a BATD compared with a standard frequentist approach using a four-arm superiority trial in regional anesthesia as motivation.

METHODS

Motivating example

We consider a hypothetical superiority trial comparing wound infiltration, TAP, and quadratus lumborum blocks with standard of care (SOC) in lower abdominal surgery. Experimental treatments are not compared with each other. For simplicity, we assume all patients come from a homogeneous population (lower abdominal surgery) and that all clinicians are equally proficient at all techniques. We also assume that an experimental arm could be inferior to standard care.

For this example trial, the primary outcome measure is the patient-reported Quality of Recovery-15 score (QoR15) at 24 hours postoperatively.[13] QoR15 comprises five domains of testing: pain (two questions), physical comfort (five questions), physical independence (two questions), psychological support (two questions), and emotional state (four questions). Each question uses a 10-point scale ranging from 0=‘none of the time’ to 10=‘all of the time’ (scoring is reversed for negative questions). The maximum total score of 150 signifies the best possible recovery. A minimum 8-point difference in QoR15 is regarded as clinically significant.[14] Based on a study of hip arthroplasty patients,[15] the median (IQR) for controls is assumed to be approximately 107 (85–123). Since the follow-up period is short, it is assumed that there will be minimal missing data over the course of the trial.

Basic concepts for Bayesian adaptive designs

Monitoring of a standard two-arm trial is often based on group sequential testing occurring at different interim analysis times when decisions to stop the trial for efficacy or lack of benefit can be made.[16] Performing multiple interim analyses could inflate the chance of a false positive result, so adjustments are used to control the type I error at a desired level (alpha level).[17] Bayesian monitoring of a trial is also based on multiple looks at the data, but decisions are based on computation of posterior probabilities. Examples include the probability of a positive treatment effect given the available data, or the probability that a specific arm is superior to SOC, given the data. This approach can be evaluated for any event/comparison of interest and has the natural flexibility that suits interim analysis of accruing data. Essentially, Bayesian methods use available knowledge and early estimates of the most relevant parameters, here the median QoR15 in each arm. Posterior probability computations are often based on a prior probability distribution chosen as uninformative or weakly informative in the absence of existing data. The prior is then combined with the observed

[1]Department of Epidemiology and Preventive Medicine, School of Public Health and Preventive Medicine, Monash University, Melbourne, Victoria, Australia
[2]Department of Anaesthesia and Acute Pain Medicine, St Vincent’s Hospital, Melbourne, Victoria, Australia
[3]Melbourne Medical School, Faculty of Medicine, Dentistry and Health Sciences, University of Melbourne, Melbourne, Victoria, Australia
[4]Department of Biostatistics, Dana-Farber Cancer Institute and Harvard TH Chan School of Public Health, Boston, Massachusetts, USA

Correspondence to Dr Stephane Heritier, Department of Epidemiology and Preventive Medicine, School of Public Health and Preventive Medicine, Monash University, Melbourne, Vic, Australia; stephane.heritier@monash.edu

data to generate the posterior distribution of the parameter of interest. The interested reader can refer to these sources—refs 8 18 19—for further information regarding Bayesian methods for clinical trials.

In the context of the four-arm trial described above, we can compute the (posterior) probability of success (PPS), defined as the probability that the median QoR15 in an experimental arm is greater than that of the SOC arm. PPS can then be used to make a decision to stop an arm for efficacy or futility or to change the probability of allocation to an arm.

Table 1 Summary of simulation scenarios

Scenario	Effective arms	Median QoR15 score			
		Control	Group 2	Group 3	Group 4
1	0	107	107	107	107
2	0	107	101	107	107
3	1	107	107	107	115
4	1	107	101	107	115
5	2	107	107	115	115
6	2	107	101	115	115
7	3	107	115	115	115
8	1	127	127	127	135
9	1	127	119	127	135

Note: The SD of the QoR15 score on the transformed scale is 0.81 for all scenarios based on historical data.
QoR15, Quality of Recovery-15 score.

Stopping for futility or efficacy

Typically, if PPS is higher than the success cut-off at a given interim analysis (s_i), then substantial evidence exists to declare that the experimental arm is better than SOC. Conversely, if PPS is lower than the futility cut-off (f_i), we can argue that there is little evidence that this particular experimental treatment is superior and stop that arm for futility. The cut-offs s_i and f_i need not be constant over the course of the trial but can be made to vary. For example, s_i can start high (eg, s_i=0.995) and then decrease as data accrue to account for greater variability early in the trial. Similarly, we can select increasing values for f_i to stop for lack of benefit. Overall, these cut-offs are chosen to achieve particular frequentist targets: (1) a small comparison-wise false positive rate, for example, 2.5%, to control the risk of claiming that an experimental arm is effective when it is not, equivalent to a two-sided type I error of 5% in the frequentist setting; (2) a high true positive rate (probability of detecting an effective treatment), for example, 80%. Details about the stopping rules made for our example are given in online supplementary appendix 1.

Response adaptive randomization

Another feature of the Bayesian approach is the ability to adapt the randomization probabilities at each interim look as a function of the PPS. A more traditional approach consists of fixing the allocation ratio to 1, that is, the same number of patients will be randomized to each arm at each interim analysis, a method called equal randomization. Using response adaptive randomization (RAR), after a period of equal randomization (burn-in phase) more patients are allocated to arms that are performing well and less patients to arms that are performing poorly. In our four-arm BATD, we decided to conduct the first interim analysis after the data are recorded for 80 patients (approximately 20 patients per arm). Subsequently, the probability of assignment to each group is calculated after every 20 patients according to a function of the PPS,[20] with the idea of increasing the probability of randomization to the most promising treatment(s). In our example we assume that updating can occur in real time without the need to 'pause' randomization. In order to maintain statistical efficiency, the randomized allocation to the SOC group is protected[20] to ensure the number of controls matches the number in the best experimental arm (details are given in online supplementary appendix 1).

Maximum sample size and practicalities about the adaptations

A decision is made after each 20 patients to stop for futility or efficacy according to the rules given above. If these rules are not met, the trial continues until the maximum sample size is reached. In the context of this four-arm BATD, the maximum sample size was set at N=500. This is the estimated number required to detect a median QoR15 difference of 8 points (see table 1, scenario 7), in a frequentist trial with one control and three active arms, with power of 0.8 and two-sided comparison-type I error of 0.05. As discussed earlier, functions are chosen to control early stopping and RAR. These functions are tuned through simulations to achieve the desired frequentist characteristics, that is, 2.5% comparison-wise false positive rate on average when all treatments are equally effective and a true positive rate of 80% (details in online supplementary appendix 1).

Analysis

QoR15 has a left-skewed distribution,[15 21] but preliminary data showed that the empirical logit transformation to the QoR15 removes skewness and achieves approximate normality in this population. Calculations can be sped up significantly by choosing a conjugate prior probability (this means prior and posterior distributions are mathematically related and closed form mathematical formulas are available). In principle, any non-conjugate prior could be used to compute the PPS but at considerable computational cost. Since the mean and the median are the same in the normal model, we can deal with means after transformation. We set a weakly informative prior distribution for both the mean and variance of transformed QoR15 in each arm. The choice is guided by both practical considerations, that is, a direct calculation of the posterior distribution of the parameters of interest and the expected effect of treatment (details in online supplementary appendix 1).

Scenarios

We use computer simulation to assess the performance of our BATD under nine scenarios (table 1) where the 'true' distribution of QoR15 is known for each group. Each simulation represents a full randomized trial where subjects' outcomes are randomly drawn from their group's distribution. The 'trials' are repeated 10 000 times to estimate the frequentist operating characteristics and expected sample sizes under each scenario (table 2).

Each scenario has a different number of null, superior and inferior arms in the trials. For example, scenario 1 has all experimental arms with 'true' effect equal to standard care, while scenario 4 has one inferior, one effective and one null arm. Scenarios 1–7 all assume a median QoR15 of 107 (0.92 on the empirical logit scale) for the control arm, and median QoR15 of 115 and 101 for the superior arms and inferior arms,

Table 2 Summary of simulation results

Scenario	Number of patients, mean (SD)					Reached maximum accrual (proportion)§	Proportion of trials found one or more arms efficacious at completion*	Proportion of trials found one or more arms efficacious and stopped early†	Proportion of trials in which arm was found efficacious‡		
	Total sample size	Control	Group 2	Group 3	Group 4				Group 2	Group 3	Group 4
1	456 (76)	140 (28)	105 (40)	106 (41)	105 (40)	0.63	0.05	0.01	0.021	0.021	0.021
2	437 (93)	150 (37)	42 (20)	122 (47)	122 (47)	0.53	0.03	0.01	0.000	0.025	0.024
3	466 (58)	140 (21)	96 (34)	96 (34)	135 (22)	0.63	0.52	0.37	0.024	0.026	0.887
4	431 (80)	143 (29)	42 (20)	107 (39)	140 (29)	0.42	0.35	0.58	0.000	0.039	0.929
5	457 (667)	129 (20)	89 (32)	120 (21)	120 (21)	0.58	0.52	0.42	0.039	0.846	0.849
6	374 (95)	117 (33)	41 (13)	108 (33)	108 (33)	0.21	0.17	0.79	0.000	0.883	0.887
7	400 (109)	108 (29)	97 (28)	97 (28)	97 (28)	0.39	0.34	0.61	0.806	0.808	0.808
8	441 (84)	128 (24)	98 (37)	98 (37)	117 (22)	0.53	0.53	0.47	0.025	0.024	1.000
9	395 (107)	129 (35)	33 (14)	113 (43)	120 (31)	0.35	0.35	0.65	0.000	0.042	1.000

The table shows the summary of 10 000 trial simulations. The total sample size for a frequentist trial with one control and three active arms, with power of 0.8 and two-sided comparison-type I error of 0.05, would be 500. The total sample size for three two-arm frequentist trials run in series, with power of 0.8 and two-sided comparison-type I error of 0.05, would be 774.

*Proportion of trials that concluded one or more arms were efficacious at the last look.

†Proportion of trials that finished early and concluded one or more arms were efficacious.

‡Success group-proportion of trials in which the arm was found efficacious.

§Proportion of trials that reached the maximum sample size

respectively. Scenarios 8 and 9 explore a situation in which the median QoR15 is 127 for the standard care group, and 135 and 119 for the superior and inferior arms, respectively. For all scenarios, the scale parameter after transformation was set at 0.81, as observed in the original data. For simplicity, we omit the parameter values on the transformed scale as they have no direct clinical interpretation.

All simulations were performed in R (R Development Core Team, 2018).[22] The 'metRology' package[23] was used to draw random deviates from the scaled t-distribution.

RESULTS

Figure 1 and table 2 summarize the results of 10 000 trial simulations for each scenario. Under all scenarios 37%–79% of trials stopped early without sacrificing the ability to detect effective arms.

Sample size and operating characteristics

Under scenario 1, where all arms are equal to control, the chance of erroneously concluding an arm was superior (false positive rate) was less than 2.1% (target <2.5%) for each arm or 5.4% for any one of the three non-control arms (overall false positive rate under the global null). This is similar to a balanced randomization four-arm trial without interim analysis, but with approximately 9% less patients on average. For perspective, the standard practice of performing three separate two-arm trials with the same characteristics would require 774 patients, so approximately 270 subjects are saved simply by using a shared control arm. Where all arms are effective (scenario 7), the ability to detect an effective arm is 80%. Again this is similar to a balanced four-arm trial but with 20% less patients.

In scenarios 4 and 9, where one arm is inferior, one neutral and one superior, the power to detect an effective arm ranges from 93% to approximately 100%. Additionally, the effective arm is allocated between 87 and 99 more patients than the ineffective arm on average. The false positive rate for the neutral arms is approximately 4%. In scenario 5, where two arms are superior and one neutral, the power to detect the effective arms is 85% with a false positive rate of 3.7% for the neutral arm. The superior arms are allocated 30 more patients than the neutral arm on average.

Under all scenarios, the false positive rate for a strictly inferior arm was close to zero. Scenarios 8 and 9 have greater power than scenarios 3 and 4 since the difference between median QoR15 values of 127 and 135 leads to a larger difference on the transformed scale than the difference between 107 and 115. Scenarios 8 and 3 both have one effective arm, but due to the larger effect size on the transformed scale, scenario 8 stops an average of 15 patients earlier, is almost certain to detect the effective arm and maintains the comparison-wise false positive rate at 2.5%.

Chance of randomization to an inferior arm

Under all other scenarios, on average, more patients were randomized to effective arms and less to inferior arms. Table 3 summarizes the chance of more patients being randomized to an inferior arm. In our simulated scenarios, the probability of an inferior arm containing more subjects than a neutral arm was always less than 10%. The chance of an inferior arm containing more patients than a superior arm was less than 0.02%. Since more patients are allocated to effective arms, the sample size and thus the power are increased for those arms. For example, this is marked in scenario 4 with a 93% chance of correctly selecting the effective arm.

DISCUSSION

We have shown it is possible to design a multiarm Bayesian trial with good control of the false positive rate and power, but with smaller expected sample size than a standard frequentist approach, under a range of plausible scenarios. This is an especially marked improvement over current practice in regional anesthesia trials, where interventions are studied in a series of three two-arm trials without interim analysis. This approach would require 774 patients in total to detect a difference of 8 QoR15 points with 80% power, compared with an expected mean sample size ranging from 394 to 466 for our simulated scenarios.

Our example design increases the chance of discovering a difference

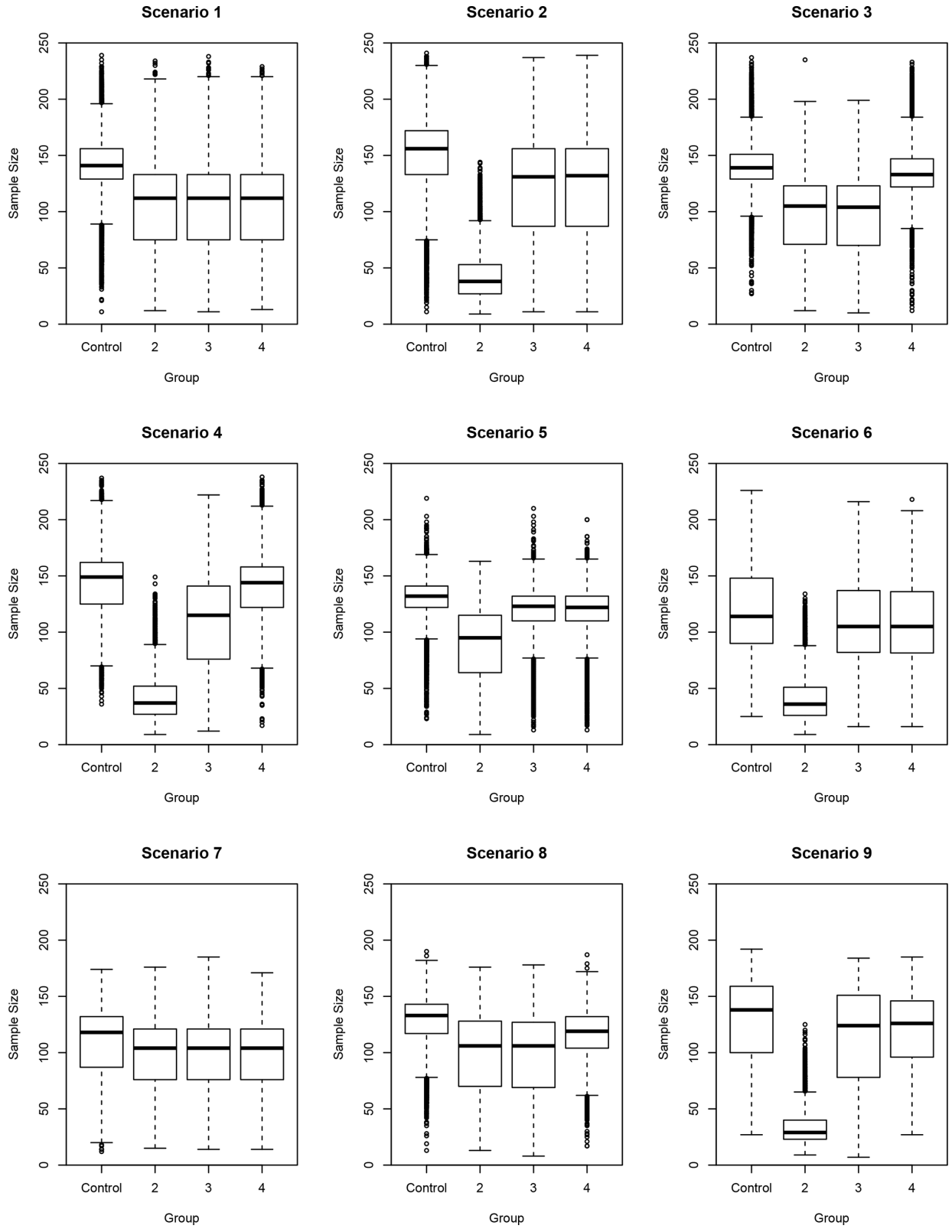

Figure 1 Boxplots of empirical sample size distribution for each arm. Control is the standard of care group, and groups 2, 3 and 4 are experimental arms. Refer to table 1 for scenario parameters.

between groups when there is an ineffective arm by randomizing more patients to the effective arm, or arms, and by reducing the chance of randomization to ineffective arms. Moreover, the advantages of our example are greater if a treatment is more (or less) effective than expected with minimal effects on the false positive rate. Finally, unlike cancer drug development,[9] it is likely that more than one treatment will be effective; therefore, our design allows the study to continue if more than one arm appears to be better than control. This prevents randomly stopping when

Table 3 Empirical proportion of trials that randomize in the wrong direction

	Efficacy of group			Probability (group$_i$ is larger than group$_j$)		
Scenario	Group 2	Group 3	Group 4	Group 2 > group 3	Group 2 > group 4	Group 3 > group 4
1	Neutral	Neutral	Neutral	0.50	0.50	0.50
2	Inferior	Neutral	Neutral	0.06	0.06	0.50
3	Neutral	Neutral	Superior	0.50	0.16	0.16
4	Inferior	Neutral	Superior	0.07	0.00	0.16
5	Neutral	Superior	Superior	0.18	0.18	0.49
6	Inferior	Superior	Superior	0.02	0.02	0.48
7	Superior	Superior	Superior	0.48	0.48	0.48
8	Neutral	Neutral	Superior	0.50	0.33	0.33
9	Inferior	Neutral	Superior	0.04	0.01	0.39

Note: Neutral—no treatment effect; superior—superior to control; inferior—inferior to control. This table summarizes the proportion of trials in which a worse arm was allocated more patients than a better arm. For example, in scenario 4, group 2 (inferior to control) is allocated more patients than group 3 (no treatment effect) in 7% of trials but never allocated more patients than group 4 (superior to control). Also group 3 is allocated more patients than group 4 in 16% of trials.

one arm reaches the success boundary, thereby missing another effective treatment.

Although we have shown increased efficiency in terms of sample size in this Bayesian-monitored, multiarm trial, this type of trial has some disadvantages. It is more complex to implement; data collection must be centralized and timely with close cooperation between sites. This problem is not insurmountable and a recent paper describes the successful implementation of a platform for Bayesian trials using open-source software.[24] Efficiency gains can also be achieved with frequentist multi-arms multistage designs (MAMS) that use prespecified stopping boundaries to determine if a study arm should be stopped. However, due to difficulties calculating stopping boundaries, this methodology is only available for a limited number of outcomes.[11] Moreover, for a binary outcome, BATD with adaptive randomization has been shown to be more efficient than MAMS when there is an effective treatment.[25]

Planning of a Bayesian trial is also typically more complex than for similar frequentist trials since the operating characteristics of the trial cannot be determined analytically. Extensive computer simulations are required and custom-code or proprietary software must be used. Since many concepts will be new to clinicians, ethics committees, funders and readers, more effort may be required to communicate the study methods and design benefits.

In the two-arm setting, Thall *et al*[26] argue that RAR may increase bias and create statistical inefficiencies that outweigh any potential ethical advantages. They also point out that in any study using RAR, there is a probability that more patients will be randomized to an inferior arm than a superior arm. However, we have shown in our example that the chance of unfavorable imbalance can be minimized by tuning the design, and as has been discussed elsewhere the advantages of RAR increase with increasing numbers of study arms.[27]

One significant concern is that time trends, often referred to as 'parameter drift', lead to biased estimates in adaptive designs. An example might be a gradual improvement in QoR15 related to improvement in surgical technique unrelated to study intervention. In this situation, it is easy to see how a treatment that had low randomization probability due to random poor performance early in the study would be further penalized as overall outcomes improved. It may be possible to account for drift with modeling or by only making comparisons to contemporaneous controls; however, it is likely that with excessive secular trends, adaptive designs may be inappropriate.[27]

We have shown the benefits of one relatively simple Bayesian adaptive design. Further advantages can be gained through platform trial design. A platform trial allows multiple interventions to be studied under one master protocol, with the emphasis being on the overall clinical problem rather than any individual intervention.[28] Possible extensions include the adding of new intervention arms[10] as they become available or the 'enrichment' of the study population with subjects in whom an intervention looks most promising.[29] A platform trial in regional anesthesia would allow multiple blocks to be studied in multiple patient populations and new blocks to be added to the study as they are developed. By having multiple patient groups, we could learn which blocks are effective in which population(s) and increase recruitment of those subgroups. In our example, we compare with an SOC group, but a modified design could make it possible to design a trial where the aim is to find the best performing block. Perhaps the most ambitious of adaptive designs is the 'randomized, embedded, multifactorial, adaptive platform'. This design embeds a perpetual randomized trial into clinical practice with the idea of integrating causal inference with continuous quality improvement. The study ensures that future patients are more likely to be randomized to the best available treatment, effectively creating a health system that learns and translates research automatically.[28]

CONCLUSION

Regional anesthesia is a field in which evidence from traditional small two-arm RCTs cannot keep pace with innovation. We have shown how a multiarm Bayesian adaptive trial could accelerate research by sharing controls, allocating more patients to effective arms and stopping earlier for either success or futility. This would provide clinicians with information on the potential efficacy of ultrasound-guided interfascial plane blocks and could reduce the risk of ineffective techniques being adopted early.

Correction notice This article has been corrected since it published Online First. Table 2 and reference 33 have been updated.

Acknowledgements We thank Dr Niki Tan for providing pilot data for the outcome model.

Contributors All authors made substantial contributions to the conception and design of the work, and the acquisition, analysis and interpretation of data. All authors were involved in writing the drafts and approving the final manuscript.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/rapm-2018-100248).

To cite Gotmaker R, Barrington MJ, Reynolds J, *et al*. *Reg Anesth Pain Med* 2019;**44**:617–622.

Received 9 November 2018
Revised 13 January 2019
Accepted 26 January 2019
Published Online First 2 March 2019

Reg Anesth Pain Med 2019;**44**:617–622.
doi:10.1136/rapm-2018-100248

REFERENCES

1 Abrahams M, Derby R, Horn JL. Update on ultrasound for truncal blocks: a review of the evidence. *Reg Anesth Pain Med* 2016;41:275–88.
2 Elsharkawy H, Pawa A, Mariano ER. Interfascial plane blocks: back to basics. *Reg Anesth Pain Med* 2018;43:341–6.
3 Baeriswyl M, Kirkham KR, Kern C, *et al*. The analgesic efficacy of ultrasound-guided Transversus abdominis plane block in adult patients: a meta-analysis. *Anesth Analg* 2015;121:1640–54.
4 Sivakumar H, Peyton PJ. Poor agreement in significant findings between meta-analyses and subsequent large randomized trials in perioperative medicine. *Br J Anaesth* 2016;117:431–41.
5 Hussain N, Ferreri TG, Prusick PJ, *et al*. Adductor canal block versus femoral canal block for total knee arthroplasty: a meta-analysis: what does the evidence suggest? *Reg Anesth Pain Med* 2016;41:314–20.
6 Berry DA. Adaptive clinical trials in oncology. *Nat Rev Clin Oncol* 2011;9:199–207.
7 Kim C, Giaccone G. Lessons learned from BATTLE-2 in the war on cancer: the use of Bayesian method in clinical trial design. *Ann Transl Med* 2016;4.
8 Trippa L, Lee EQ, Wen PY, *et al*. Bayesian adaptive randomized trial design for patients with recurrent glioblastoma. *J Clin Oncol* 2012;30:3258–63.
9 Saville BR, Berry SM. Efficiencies of platform clinical trials: a vision of the future. *Clin Trials* 2016;13:358–66.
10 Ventz S, Alexander BM, Parmigiani G, *et al*. Designing clinical trials that accept new arms: an example in metastatic breast cancer. *J Clin Oncol* 2017;35:3160–8.
11 Curtin F, Heritier S. The role of adaptive trial designs in drug development. *Expert Rev Clin Pharmacol* 2017;10:727–36.
12 Ventz S, Parmigiani G, Trippa L. Combining Bayesian experimental designs and frequentist data analyses: motivations and examples. *Appl Stoch Models Bus Ind* 2017;33:302–13.
13 Stark PA, Myles PS, Burke JA. Development and psychometric evaluation of a postoperative quality of recovery score: the QoR-15. *Anesthesiology* 2013;118:1332–40.
14 Myles PS, Myles DB, Galagher W, *et al*. Minimal clinically important difference for three quality of recovery scales. *Anesthesiology* 2016;125:39–45.
15 Tan NLT, Hunt JL, Gwini SM. Does implementation of an enhanced recovery after surgery program for hip replacement improve quality of recovery in an Australian private hospital: a quality improvement study. *BMC Anesthesiol* 2018;18.
16 Jennison C, Turnbull BW. *Group sequential methods with applications to clinical trials*. Boca Raton: Chapman & Hall/CRC, 2000.
17 DeMets DL. Methodological issues in AIDS clinical trials. Data monitoring and sequential analysis--an academic perspective. *J Acquir Immune Defic Syndr* 1990;3 Suppl 2:S124–33.
18 Berry DA. Bayesian clinical trials. *Nat Rev Drug Discov* 2006;5:27–36.
19 Berry SM. *Bayesian adaptive methods for clinical trials*. Boca Raton: CRC Press, 2011.
20 Cellamare M, Ventz S, Baudin E, *et al*. A Bayesian response-adaptive trial in tuberculosis: the endTB trial. *Clin Trials* 2017;14:17–28.
21 Chazapis M, Walker EM, Rooms MA, *et al*. Measuring quality of recovery-15 after day case surgery. *Br J Anaesth* 2016;116:241–8.
22 R Core Team. *R: a language and environment for statistical computing*, 2018.
23 Ellison SLR. *metRology: support for Metrological applications*, 2018.
24 Brown AR, Gajewski BJ, Aaronson LS, *et al*. A Bayesian comparative effectiveness trial in action: developing a platform for multisite study adaptive randomization. *Trials* 2016;17.
25 Wason JM, Trippa L. A comparison of Bayesian adaptive randomization and multi-stage designs for multi-arm clinical trials. *Stat Med* 2014;33:2206–21.
26 Thall P, Fox P, Wathen J. Statistical controversies in clinical research: scientific and ethical problems with adaptive randomization in comparative clinical trials. *Ann Oncol* 2015;26:1621–8.
27 Berry DA. Adaptive clinical trials: the promise and the caution. *J Clin Oncol* 2011;29:606–9.
28 Angus DC. Fusing randomized trials with big data: the key to Self-learning health care systems? *JAMA* 2015;314:767–8.
29 Park JJ, Thorlund K, Mills EJ. Critical concepts in adaptive clinical trials. *Clin Epidemiol* 2018;10:343–51.
30 Stevens S, Valderas JM, Doran T, *et al*. Analysing indicators of performance, satisfaction, or safety using empirical logit transformation. *BMJ* 2016;352.
31 Arostegui I, Núñez-Antón V, Quintana JM. Statistical approaches to analyse patient-reported outcomes as response variables: an application to health-related quality of life. *Stat Methods Med Res* 2012;21:189–214.
32 Lesaffre E, Rizopoulos D, Tsonaka R. The logistic transform for bounded outcome scores. *Biostatistics* 2007;8:72–85.
33 Gelman A*Bayesian data analysis. Boca Raton*. CRC Press, 2014.