*Biostatistics* (2020) **21**, 2, *pp.* 287–301
doi:10.1093/biostatistics/kxy048
Advance Access publication on September 10, 2018

# Sequential Monte Carlo for response adaptive randomized trials

SHIRIN GOLCHI*

*Department of Statistics and Actuarial Science, Simon Fraser University, 8888 University Dr, Burnaby, BC, V5A 1S6, Canada and MTEK Sciences, 802-777 W Broadway Street, Vancouver, BC V5Z 1J5, Canada*
sgolchi@sfu.ca

KRISTIAN THORLUND

*MTEK Sciences, 802-777 W Broadway, Vancouver, Canada and Department of Health Research Methods, Evidence, and Impact, McMaster University, 1280 Main St W, Hamilton, ON L8S 4L8, Canada*

## Summary

Response adaptive randomized clinical trials have gained popularity due to their flexibility for adjusting design components, including arm allocation probabilities, at any point in the trial according to the intermediate results. In the Bayesian framework, allocation probabilities to different treatment arms are commonly defined as functionals of the posterior distributions of parameters of the outcome distribution for each treatment. In a non-conjugate model, however, repeated updates of the posterior distribution can be computationally intensive. In this article, we propose an adaptation of sequential Monte Carlo for efficiently updating the posterior distribution of parameters as new outcomes are observed in a general adaptive trial design. An efficient computational tool facilitates implementation of more flexible designs with more frequent interim looks that can in turn reduce the required sample size and expected number of failures in clinical trials. Moreover, more complex statistical models that reflect realistic modeling assumptions can be used for analysis of trial results.

*Keywords*: Bayesian adaptive design; Bayesian interim analysis; Sequential updating.

## 1. Introduction

In a classic randomized clinical trial (RCT), key design components such as sample size and allocation ratio are pre-specified before the trial is launched and are held fixed through the course of the trial regardless of results obtained during any interim analysis. Adaptive clinical trial designs are alternatives to RCT that allow for continual adaptation of the design components throughout the trial. Particularly, response adaptive randomized (RAR) trials (Berry *and others*, 2010; Atkinson and Biswas, 2014; He *and others*, 2014) have gained popularity recently. In RAR trials, the allocation ratio among treatment arms is repeatedly adapted to increase enrollment to the best treatment according to the accumulating data. The trial may be stopped early if overwhelming statistical evidence suggest the superiority of one treatment

*To whom correspondence should be addressed.

over the rest. Further, treatments (or doses) may be excluded from the trial due to evidence of inferiority in interim analyses.

The main advantage of well-designed adaptive trials is that they allow for making more ethical and cost efficient decisions. RAR trials in particular ensure that fewer patients are randomized to an inferior treatment, thus reducing the overall number of negative outcomes. Adaptive designs (including RAR designs) may also aid in reaching firm evidence early on, thereby saving time and costs. These advantages, however, arrive with the potential pitfall that any trial adaptation may drastically affect the statistical properties of the trial such as power and Type I error rate. Unlike RCT, however, for adaptive designs the statistical criteria (power and error rates) are not given in closed form mathematical expressions as a function of various decision rules. Therefore, quantifying and controlling statistical errors typically require comprehensive simulation studies that due to the frequency of interim updates are often computationally intensive. Thus, improvements in the efficiency for adaptive trial planning simulations are in high demand.

Due to their flexible nature, nowadays Bayesian methods are used in a majority of adaptive clinical trials to establish informed decision rules for trial adaptations (Berry *and others*, 2010). RAR designs are not an exception; in particular, interim posterior or predictive probabilities of treatment superiority or inferiority have become the cornerstone of adaptive trial decision rules. For example, in the high-profile I-SPY2 trial, Bayesian predictive probabilities exceeding 0.85 exhibiting superior treatment effects across subgroup signatures were used to guide, which treatments and populations should be further subjected to randomized investigation in a seamless transition from Phase II to Phase III of the trial (Park *and others*, 2016). Bayesian models for clinical trial adaptations, however, are typically non-conjugate and therefore become most computationally intensive to simulate due to the need for employing Markov chain Monte Carlo (MCMC) algorithms to estimate posterior probabilities. In practice, the need for MCMC may limit the number of simulations that can be performed prior to trial launch, or delay trial launch due to the time required.

Alternatives to sampling are approximation methods such as variational inference. These methods aim at providing an approximate closed form posterior as opposed to drawing samples from the exact posterior. In general, if affordable, sampling from the exact posterior is preferred since Monte Carlo error may be reduced by increasing the Monte Carlo sample size, while approximation error is often difficult to quantify and control.

Sequential Monte Carlo (SMC) samplers or particle filters (Del Moral *and others*, 2006; Doucet *and others*, 2013) are a class of algorithms that use importance weights to update a sample with respect to a target distribution through a sequence of densities. The intuition behind SMC is that the sequence of densities is used as a bridge between a distribution that can be easily sampled and the target distribution. The initial sample is “filtered through the bridge” using weighting and sampling steps that are performed independently. Therefore, the algorithm is embarrassingly parallel. In addition to being easily parallelizabale, SMC samplers are known to be better at capturing and exploring challenging distributional features such as multimodality, flat (zero probability) regions and isolated peaks. Moreover, particle filters are specifically designed to make Bayesian inference on data that are collected/observed sequentially without the requirement of complete exploration of the data history. In this context, the sequence of densities is naturally defined by updating the likelihood with the accumulating data. Unlike the underlying intuition, implementation of SMC samplers is often complicated and extensive tuning is required to determine the optimal settings of the algorithm. The flexibility of SMC in capturing the problem specific sampling challenges is also the source of implementation difficulties since the tuning step and specification of algorithm settings will also become context specific and cannot be automated.

In this article, we introduce an SMC algorithm tailored for RAR designs and provide general guidelines for choices that need to be made in the sampling algorithm. While the SMC update can be used for any interim analysis, it is specially applicable for simulating RAR trials with complex models. Note that the focus of this article is to present SMC as a powerful tool to enhance the efficiency of RAR trial simulations rather than focusing on a simulation study for a specific RAR design. The rest of the article is organized as follows. In Section 2, we explain the challenges of Bayesian RAR trials and introduce the SMC algorithm. The implementation detail are described in Section 3 using two simulated examples. The performance of the proposed algorithm is compared with MCMC and variational Bayes (VB) approximation through simulations in Section 4 and Section 5 follows with a discussion.

## 2. Methods

### 2.1. *Response adaptive randomized trials and Bayesian inference*

RAR trials are a family of adaptive clinical trial designs that aim to limit the number of patients randomized to an inferior treatment arm as well as limiting the total sample size required for identifying the best performing treatment. While there exists a variety of adaptive designs in the literature (see for a review, Atkinson and Biswas, 2014), all RAR trials share the following features: allocation probabilities are adapted in the course of the trial according to the performance of each treatment arm based on available observations up to the adaptation points; a set of rules are defined to drop an arm due to futility or to stop the trial either due to decisiveness of evidence for efficacy or achieving a maximum affordable sample size. The efficiency of RAR designs are commonly assessed with respect to criteria such as expected number of failures (for binary and survival responses) or expected sample size at termination (Atkinson and Biswas, 2014). Due to the adaptive and sequential nature of these designs and the complexity that is added by defining different decision rules, simulations are required to obtain the statistical properties (e.g. power and Type I error rate) resulted from a set of chosen decision rules. In this vein, several simulations covering a multitude of "best" and "worst" case scenarios across a spectrum of decision rules are typically necessary to determine the optimal design.

The requirement of sequential updates in adaptive designs makes the Bayesian analysis framework specially appealing—the results of the analysis from a previous interim look can be used as prior information for analyzing the most recent data without having to browse the historical data. Therefore, researchers have shifted attention toward Bayesian methods in the recent years. Examples of work in Bayesian RAR clinical trials are Cheng and Shen (2005), Berry *and others* (2010), and Saville and Berry (2016).

To be more specific, consider a RAR trial with $L$ arms where we denote the vector of parameters as $\boldsymbol{\theta}^T = (\theta_1, \ldots, \theta_L)$, where $\boldsymbol{\theta}^T$ denotes the transpose of vector $\boldsymbol{\theta}$. Let $x_n$, be a vector of size $L$ whose $l^{\text{th}}$ element is 1 if patient $n$ ($n = 1, \ldots, N$) receives treatment $l$ and zero otherwise, and $y_n$ be the response of patient $n$. Also we denote the cumulative data of $n$ patients as a $n \times L$ matrix $\mathbf{x}_n$ and a $n \times 1$ vector $\mathbf{y}_n$. When the outcome for a patient is observed the allocation probability of each arm $l$ is updated according to performance of the treatments inferred from the data, which is in turn extracted from the parameter estimates.

Bayesian inference is based on the posterior distribution of the unknowns $\boldsymbol{\theta}$ given the data, which is a product of a prior distribution on $\boldsymbol{\theta}$, the likelihood that describes the stochastic relationship between the data and $\boldsymbol{\theta}$, and a normalizing constant. In Bayesian RAR, therefore, the arm allocation probabilities after $n$ patients are obtained as a functional of the posterior, $\pi(\boldsymbol{\theta} \mid \mathbf{x}_n, \mathbf{y}_n)$, that can be written as,

$$\pi(\boldsymbol{\theta} \mid \mathbf{x}_n, \mathbf{y}_n) = \frac{\pi(\boldsymbol{\theta} \mid \mathbf{x}_{n-1}, \mathbf{y}_{n-1})\pi(y_n \mid \boldsymbol{\theta}, x_n)}{\int \pi(\boldsymbol{\theta} \mid \mathbf{x}_{n-1}, \mathbf{y}_{n-1})\pi(y_n \mid \boldsymbol{\theta}, x_n) d\boldsymbol{\theta}}, \tag{2.1}$$

where $\pi(y_n \mid \boldsymbol{\theta}, x_n)$ is the likelihood based on the most recent observation and the posterior from the last update, $\pi(\boldsymbol{\theta} \mid \mathbf{x}_{n-1}, \mathbf{y}_{n-1})$, is used as a prior distribution for $\boldsymbol{\theta}$. Therefore, the adaptations are performed by simply updating the posterior distribution with the most recently observed data without having to deal with the historical data since they are summarized in the prior (i.e. the posterior up until the very last patient's data was observed). This posterior update is straightforward as long as the integral in the denominator on the right hand side of (2.1) can be obtained analytically, and therefore, the posterior can be written in closed form. However, most adaptive designs comprise a set of decision rules and require complex statistical models for which a closed form posterior does not exist. As such, Bayesian inference utilizing MCMC has become popular in RAR trials as it bypasses these limitations since using MCMC samples can be drawn from a distribution as long as it is known up to a normalizing constant

The application of MCMC to RAR, however, is also accompanied by some difficulties and inefficiencies. As explained above, in Bayesian RAR trials, to avoid reanalyzing the cumulative data, for each patient $n$ the posterior distribution $\pi(\boldsymbol{\theta} \mid \mathbf{x}_n, \mathbf{y}_n)$ is updated using $\pi(\boldsymbol{\theta} \mid \mathbf{x}_{n-1}, \mathbf{y}_{n-1})$ as the prior distribution. However, $\pi(\boldsymbol{\theta} \mid \mathbf{x}_{n-1}, \mathbf{y}_{n-1})$ is not available in closed form and needs to be estimated from samples drawn from the last update performed also using MCMC. This is done either by choosing a parametric family and using moment estimation or using non-parametric estimates such as kernel density estimation methods. A major drawback of repeated density estimation is accumulation of error as well as the requirement for hands on choices that need to be made for an appropriate parametric family or a kernel smoother bandwidth. Therefore, automation of the updating process in not straightforward. In addition, running and monitoring a Markov chain for convergence as frequently as patients are enrolled in the trial is tedious and inefficient.

In addition to the application specific challenges mentioned above the common random walk based MCMC algorithms are known to suffer from poor mixing and low acceptance rates in challenging sampling problems. More specifically, entrapment of the Markov chain in local modes or in low probability regions in the posterior surface are common weaknesses of these algorithms. SMC algorithms, on the other hand, are know to perform well in capturing multimodality and other challenging distributional features. In the following, we provide a brief introduction to SMC and introduce an SMC algorithm tailored for RAR trial designs.

### 2.2. *Sequential Monte Carlo*

SMC samplers or particle filters (Del Moral *and others*, 2006; Doucet *and others*, 2013) are a family of algorithms that are designed to update a sample through a sequence of distributions, bypassing intermediate density estimations or convergence of a Markov chain at every step. Therefore, they avoid the major drawbacks of MCMC outlined above.

SMC samplers are based on importance sampling techniques. The idea is to create a link between the target distribution $\pi_N$ and a distribution $\pi_0$, that is straightforward to sample, using a sequence of intermediate distributions,

$$\{\pi_n\}_{n=1}^{N}, \tag{2.2}$$

At each step $n$, a sample of “particles” that is initially generated from $\pi_0$ is weighted according to $\pi_n$ and resampled with respect to the weights. Particle filters are most commonly used in making Bayesian inference for dynamic state-space models where a sequence of distributions is defined naturally as the evolving posterior distribution of parameters and states given the observations through time. The use of SMC, however, is not restricted to dynamic models as Chopin (2002) demonstrate that they can improve estimates in a static modeling framework where the posterior given incomplete data can be used to construct the intermediate sequence of distributions.

In the following, we use a RCT with $L$ treatment arms and a pre-fixed sample size $N$ to describe the particle filter in a static modeling framework. The goal is to estimate the parameter $\theta_l$ for treatment $l = 1, \ldots, L$ based on the observed responses for the $N$ enrolled patients, $\mathbf{y}_N^T = (y_1, \ldots, y_N)$ as well as at any interim analysis. Based on $N$ patients data, the target distribution is given by,

$$\begin{aligned}\pi_N(\boldsymbol{\theta}) &= \pi(\boldsymbol{\theta} \mid \mathbf{y}_N, \mathbf{x}_N)\\ &\propto \pi_0(\boldsymbol{\theta})\pi(\mathbf{y}_N \mid \boldsymbol{\theta}, \mathbf{x}_N),\end{aligned} \tag{2.3}$$

where, as in the previous section, $\boldsymbol{\theta}^T = (\theta_1, \ldots, \theta_L)$ is the vector of parameters and $\mathbf{x}_N$ is a $N \times L$ design matrix whose $[n, l]$ element is 1 if patient $n$ received treatment $l$ and 0, otherwise. Suppose that an interim analysis is to be done based on the $n < N$ patients' data available. The posterior distribution of parameters given these intermediate data can be written as

$$\pi_n(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta} \mid \mathbf{y}_n, \mathbf{x}_n). \tag{2.4}$$

Analogously, a sequence of posterior distributions is given as,

$$\{\pi_n(\boldsymbol{\theta})\}_{n=0}^{N}, \tag{2.5}$$

where $\pi_0(\boldsymbol{\theta})$ is the prior distribution of parameters before any patient data is available and

$$\begin{aligned} \pi_n(\boldsymbol{\theta}) &\propto \pi(\boldsymbol{\theta} \mid \mathbf{y}_{n-1}, \mathbf{x}_{n-1})\pi(y_n \mid \boldsymbol{\theta}, x_n) \\ &\propto \pi_{n-1}(\boldsymbol{\theta})\pi(y_n \mid \boldsymbol{\theta}, x_n). \end{aligned} \tag{2.6}$$

An SMC sampler can then be used to draw a sample from the prior, $\pi_0(\boldsymbol{\theta})$ and filter this sample after the data for new patients are observed. The filtering process comprises two steps; first, samples are weighted and resampled according to the intermediate posterior, $\pi_n(\boldsymbol{\theta})$. Given that $\pi_n(\boldsymbol{\theta})$ is given by (2.6) the weights are calculated as the likelihood evaluated for the most recently observed data. That is, for a sample of size $M$ drawn from $\pi_n(\boldsymbol{\theta})$, each particle's weight is calculated as,

$$W_i^n = W(\boldsymbol{\theta}_i^{n-1}) = \pi(y_n \mid \boldsymbol{\theta}_i^{n-1}, x_n), \quad i = 1, \ldots, M. \tag{2.7}$$

The particles are then resampled with respect to these weights. Next, a MCMC transition for each particle $\boldsymbol{\theta}_i^n$ is performed by generating $\boldsymbol{\theta}^*$ from a proposal distribution, $q_n(\boldsymbol{\theta} \mid \boldsymbol{\theta}_i^n)$, following by an accept/reject step according to the intermediate posterior $\pi_n$. The MCMC transition step should not be confused with using MCMC to sample form the posterior. This step consists of mutating the particles to assure that the final sample does not consist of multiple copies of a few unique values due to sequential importance sampling. The use of MCMC transition kernels for mutation assures convergence of the overall SMC with the specified weight expressions in (2.7) (Del Moral *and others*, 2006).

Pseudo code for an SMC sampler to sample from the posterior given in (2.3) through intermediate posteriors of the form (2.4) is provided in Algorithm 1. In step 4:f, the compact notation $K^n(\boldsymbol{\theta})$ is used to represent the MCMC transition kernel explained above.

Note that Algorithm 1 aims at updating the posterior sample after the outcomes are available for every single new patient. However, it can easily be generalized to the case that updates are made for batches of patient data. Denoting data corresponding to a group of patients as $(\mathbf{x}_b, \mathbf{y}_b)$, the weights at step 4:b of Algorithm 1 are calculated as $\pi(\mathbf{y}_b \mid \boldsymbol{\theta}_i^{b-1}, \mathbf{x}_b)$. The SMC sampler in Algorithm 1 can be stopped at any intermediate step if the results based on the intermediate data include enough evidence to conclude the trial.

### 2.3. *SMC for Bayesian RAR trial simulation*

While interim analysis may be performed in any clinical trial design, frequent interim looks are a necessity in RAR designs. On the other hand, statistical properties of RAR trials as a function of different decision criteria and stopping rules need to be explored via simulation studies in the planning phase. Simulations are generally performed to study various statistical properties, such as statistical power and error rates, as well as cost related properties such as the expected sample size at termination or expected number

**Algorithm 1** SMC sampling for sequential Bayesian analysis

```
Inputs: Patient data $\{(x_1,y_1),\ldots,(x_N,y_N)\}$; prior distribution $\pi_0(\boldsymbol{\theta})$; MCMC transition kernels $K^n(\boldsymbol{\theta})$;
    number of particles $M$.
$n \leftarrow 0$;
Generate a sample, $\boldsymbol{\theta}^0_{1:M}$ from the prior, $\pi_0(\boldsymbol{\theta})$;
Initiate the weights $W^0_i \leftarrow \frac{1}{M}$ for $i = 1,\ldots,M$;
while $n < N$ do
    a. $n \leftarrow n+1$;
    b. $W^n_i \leftarrow \pi(y_n \mid \boldsymbol{\theta}^{n-1}_i, x_n)$;
    c. Normalize $W^n_{1:M}$, i.e., $W^n_i \leftarrow \frac{W^n_i}{\sum_{i=1}^M W^n_i}$;
    d. Resample $\boldsymbol{\theta}^{n-1}_{1:M}$ with weights $W^n_{1:M}$;
    e. Sample $\boldsymbol{\theta}^n_{1:M} \sim K^n(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{n-1}_{1:M})$ (refer to the text for $K^n(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{n-1}_{1:M})$);
Return: $N$ samples $\boldsymbol{\theta}^{1:N}_{1:M}$ from $\pi_{1:N}(\boldsymbol{\theta})$.
```

of failures in binary and survival outcomes. However, the goal of the present work is not to perform a simulation study to investigate the properties of a specific RAR design but to provide an efficient recipe for simulating from any RAR trial design while taking advantage of the flexibility of Bayesian framework. The SMC algorithm introduced above enhances the feasibility of such simulation studies to a great extent. In the following, we explain simulation of a single trial given a RAR design and statistical model with implementation detail for the embedded SMC updates.

Consider a RAR trial design with $L$ treatment arms. Patients are registered in the trial one at a time and the responses are supposed to be observed instantaneously. An initial fixed sample of patients of size $n_0$ are enrolled in the trial with equal probability of assignment to different arms. After this initial sample, the allocation probabilities are adapted for any newly enrolled patient according to the performance of each treatment inferred from the responses observed thus far. The allocation rate of each arm is proportional to the square root of posterior probability that the corresponding parameter is greater than that of all the other treatments. The use of square root of the probabilities of superiority as randomization probabilities is a more conservative adaptation approach, which reduces the risk of extreme imbalance between the arm sample sizes early on in the trial (see for alternative adaptation measures, Rosenberger *and others*, 2001; Zhang and Rosenberger, 2006; Thall and Wathen, 2007; Brown *and others*, 2016; Saville and Berry, 2016). The probability of superiority for arm $l$ is estimated as the proportion of particles for which $\theta_l$ is the greatest parameter in the vector of parameters. A treatment arm is dropped if the corresponding probability of it being the best performing arm falls below a lower threshold, $\ell$. The trial is terminated if the probability of a treatment being the best performing treatment exceeds an upper threshold, $u$. In addition, a maximum sample size is specified such that if none of the treatments outperform the rest with probability greater than or equal to $u$ the trial is stopped when this sample size is achieved.

A statistical model needs to be specified as the data generating process, which is also used for analyzing the data. As explained above, a Bayesian model comprises a likelihood, $\pi(y_n \mid \boldsymbol{\theta}, x_n)$ that links the data to the parameters, and a prior, $\pi(\boldsymbol{\theta})$ that incorporates any prior information about the parameters.

Algorithm 2 outlines the steps for simulating a single RAR trial. Note that in the RAR design the rows of the design matrix are random variables whose distribution evolves according to the observed data through the course of the trial:

$$x_n \sim \text{Multinomial}(1, \boldsymbol{\rho} = (\rho_1, \ldots, \rho_L)), \tag{2.8}$$

**Algorithm 2** RAR trial simulation

```
Inputs: "True" value of the parameters $\boldsymbol{\theta}^*$; data generating model; MCMC transition kernels $K^n(\boldsymbol{\theta})$;
    number of particles $M$.
$n \leftarrow 0$;
Generate a sample, $\boldsymbol{\theta}^0_{1:M}$ from the prior, $\pi_0(\boldsymbol{\theta})$;
Initiate the weights $W^0_i \leftarrow \frac{1}{M}$ for $i = 1, \ldots, M$;
while $\max(\boldsymbol{\rho}^n) < u$ and $n < N$ do
    a. $n \leftarrow n + 1$;
    b. Randomize patient $n$ according to the current allocation probabilities, i.e., generate $x_n$ from
       (2.8) where $\boldsymbol{\rho}$ is given by (2.9);
    c. Generate observation $y_n$ from the data model, $\pi(y_n \mid \boldsymbol{\theta}^*, x_n)$;
    d. $W^n_i \leftarrow \pi(y_n \mid \boldsymbol{\theta}^{n-1}_i, x_n)$;
    e. Normalize $W^n_{1:M}$, i.e., $W^n_i \leftarrow \frac{W^n_i}{\sum_{i=1}^M W^n_i}$;
    f. Resample $\boldsymbol{\theta}^{n-1}_{1:M}$ with weights $W^n_{1:M}$;
    g. Sample $\boldsymbol{\theta}^n_{1:M} \sim K^n(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{n-1}_{1:M})$ (refer to the text for $K^n(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{n-1}_{1:M})$);
    h. Update $\boldsymbol{\rho}^n$: $\rho^n_l \leftarrow \left[\frac{1}{M}\sum_{i=1}^M \mathbf{1}(\theta^n_{li} = \max(\theta^n_{1i}, \ldots, \theta^n_{Li}))\right]^{\frac{1}{2}}$;
    i.     if $\rho^n_l < \ell$ then $\rho^n_l = 0$, for $l = 1, \ldots, L$.
Return: A sample $\boldsymbol{\theta}^{n_T}_{1:M}$ from $\pi_{n_T}(\boldsymbol{\theta})$, where $n_T$ is the sample size at trial termination.
```

where

$$\begin{aligned} \rho_1 = \ldots = \rho_L &= \frac{1}{L} && \text{for } n \leq n_0; \\ \rho_l &= [\mathrm{P}(\theta_l = \max(\theta_1, \ldots, \theta_L) \mid \mathbf{y}_n)]^{\frac{1}{2}}, \quad l = 1, \ldots, L && \text{for } n > n_0. \end{aligned} \tag{2.9}$$

The MCMC kernel $K^n(\boldsymbol{\theta} \mid \boldsymbol{\theta}^n_{1:M})$ is a Metropolis–Hastings step for the components of $\boldsymbol{\theta}$ with a normal proposal distribution centered at the current value of $\theta^n_{li}$ and a variance that is specified as the sample variance of $\theta_l$ from the previous step,

$$\hat{\mathrm{var}}(\theta_l \mid \mathbf{y}_{n-1}, \mathbf{x}_{n-1}) = \frac{1}{n-1}\sum_{i=1}^{M}(\theta^{n-1}_{li} - \bar{\theta}^{n-1}_l)^2, \tag{2.10}$$

where $\bar{\theta}^{n-1}_l$ is the previous step posterior sample mean. This proposal is accepted/rejected according to the current target distribution $\pi_n(\boldsymbol{\theta})$.

Convergence results for SMC (Chopin, 2004; Del Moral *and others*, 2006) show that in addition to the convergence to the target distribution, $\pi_N$, the intermediate SMC posteriors converge to the true intermediate posteriors almost surely as $M \to \infty$. Moreover, for any measurable function $\phi$, $E_{\pi^M_n}(\phi) \to E_{\pi_n}(\phi)$ almost surely as $M \to \infty$, where $\pi^M_n$ is the SMC posterior. The probability of superiority can be written as,

$$\mathrm{P}(\theta_l = \max(\theta_1, \ldots, \theta_L) \mid \mathbf{y}_n) = \mathrm{E}_{\pi_n}\left(\mathbf{1}(\theta_l = \max(\theta_1, \ldots, \theta_L))\right),$$

where $\mathbf{1}(\cdot)$ is an indicator function. Therefore, we have

$$\mathrm{E}_{\pi^M_n}(\mathbf{1}(\theta_l = \max(\theta_1, \ldots, \theta_L))) \to \mathrm{E}_{\pi_n}(\mathbf{1}(\theta_l = \max(\theta_1, \ldots, \theta_L))),$$

almost surely as $M \to \infty$, where

$$\mathrm{E}_{\pi_n^M}(\mathbf{1}(\theta_l = \max(\theta_1, \ldots, \theta_L))) = \frac{1}{M}\sum_{i=1}^{M} \mathbf{1}(\theta_{li}^n = \max(\theta_{1i}^n, \ldots, \theta_{Li}^n)).$$

Note that Algorithm 2 can be used in the presence of covariates in the statistical model. The vector of parameters $\boldsymbol{\theta}$ should be extended to included the covariate coefficients in that case.

## 3. Examples

In this section, we consider two modeling examples where the posterior distribution cannot be obtained in closed form. The first example is a simple case with binary responses and a non-conjugate prior specification. The second example represents a less trivial situation where model parameters are assumed correlated, resulting in a likelihood for which a conjugate prior is not available.

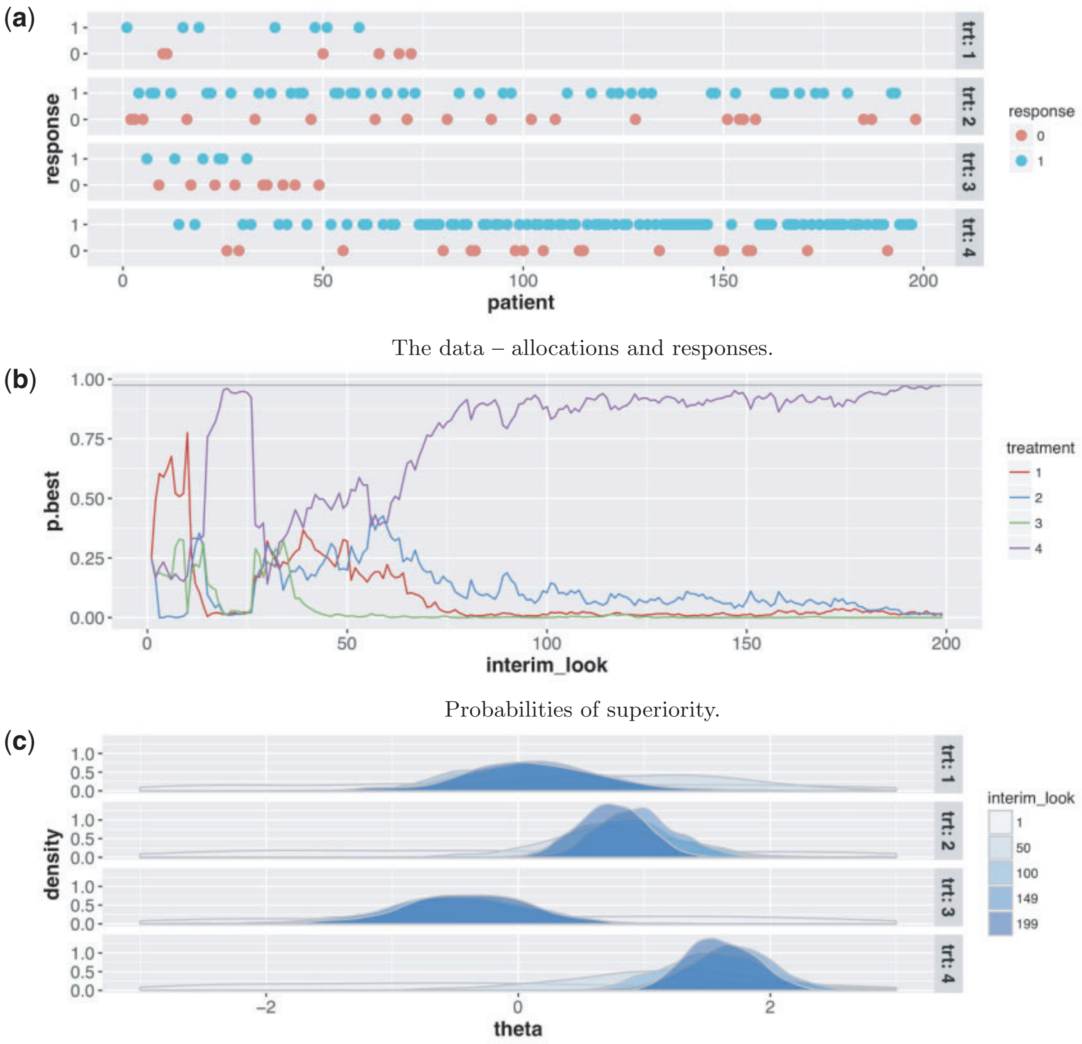
The initial sample size, $n_0$, is determined as 10 times the number of arms in each example trial. While this can be used as a rule of thumb in simulations, in practice, this initial pre-adaptation sample size needs to be specified by considering available resources, the size of the population of interest, and any prior information available on the parameters. The number of particles used to estimate the posteriors is $M = 1000$.

Example 1 Consider a four arm RAR design where each patient enrolled in the trial is given one of four competing treatments. The allocation is decided by generating $x_n$ from a multinomial distribution as given in (2.8). No adaptation is made for the allocation probabilities $\boldsymbol{\rho}$ for the first $n_0 = 40$ patients. The instantaneous response is either a success, $y = 1$, or a failure, $y = 0$ and is generated from the following Bernoulli likelihood,

$$y_n \sim \text{Bernoulli}\left(\frac{1}{1 + \exp(-x_n^T\boldsymbol{\theta})}\right). \quad (3.11)$$

An observation pair $(x_n^T = (0, 0, 0, 1), y_n = 1)$ means that patient $n$ is assigned to treatment arm 4 and the response is a success. We fix the parameters for the data generating process at $\boldsymbol{\theta}^T = (0, 0.5, -0.4, 1.2)$. A normal prior with mean 0 and variance 100 is used for $\theta_l$, $l = 1, 2, 3, 4$. The allocation probability of an arm is set to zero if the probability that the corresponding treatment is the best performing treatment, according to the historical data, falls below $\ell = 0.01$ and the trial is stopped if a treatment outperforms the rest with probability $u = 0.975$ or a total sample size of $N = 500$ patients is reached. Algorithm 2 is used to simulate a trial with the above specifications.

An example simulated RAR trial is presented in Figure 1. Figure 1a summarizes the data: each panel corresponds to a treatment arm. If patient $n$ receives treatment $l$ a dot appears in panel $l$ for patient $n$. If the dot is green the response is positive and if it is red the response is negative. The allocations are performed with equal probabilities for the four arms up to 40 patients. The probabilities of superiority are quite volatile early on due to small sample sizes, and therefore, large estimation variance (Figure 1b). Arms 1 and 3 are dropped shortly after adaptation is allowed. The reason is evident in Figure 1b where the probability of superiority drops close to zero for treatments 1 and 3 along the way. The second arm remains with a decreasing probability of superiority until the trial is terminated with a total 198 patients enrolled where enough evidence is provided that arm 4 is the superior arm. The evolving posterior distributions of the parameters at five equally spaced interim looks are presented in Figure 1c. The posterior distributions are diffuse for arms 1 and 3 since fewer data are used to estimate the corresponding parameters. For the



The data – allocations and responses.

Probabilities of superiority.

Posterior distribution of parameters.

Fig. 1. Results for Example 1. (a) The data: each panel corresponds to a treatment arm and each dot represents the response for a patient who received the treatment. A response with value 1 is positive while zero indicates a negative response. (b) The performance of treatment arms: each curve shows the evolution of the probability of superiority of the corresponding treatment as more data is collected. The horizontal line shows the upper threshold that if reached by any of the treatment curves the trial is stopped for efficacy. (c) Posterior distributions: each panel corresponds to a treatment arm. The posterior kernel density estimates of the corresponding parameters are plotted for five selected interim looks.

remaining two arms and specifically the winner (arm 4) the gradual concentration of the posterior is visible as more data are collected.

Example 2 Multi-arm trials with combination treatments: Consider a multi-arm trial where in one or more of the arms patients are given a combination of treatments that are given separately to patients in other arms. Conventionally, in such trials, the combination treatments are treated as independent treatments.

Thorlund *and others* (2017) propose to express the parameter corresponding to the combination arms as a function of the constituents, thereby incorporating the dependence through a proposed “fractional additivity” model. They show through simulation studies that their proposed model can result in efficiency and precision gains in a Bayesian adaptive framework. Incorporating the fractional additivity assumption, however, results in a non-conjugate model where Bayesian inference requires sampling/approximating the posterior. Thorlund *and others* (2017) use MCMC sampling in scenarios where a maximum of four interim looks is allowed. In the following, we briefly explain the model proposed by Thorlund *and others* (2017) and simulate an example trial from the model with interim updates that are done using the SMC algorithm.

Consider a three arm trial where treatment $A$ is given to patients in the first arm, treatment $B$ is given to patients in the second arm and patients assigned to the third arm receive both treatments $A$ and $B$. Denoting the corresponding parameters of the three arms by $\theta_A$, $\theta_B$ and $\theta_{AB}$, Thorlund *and others* (2017) proposed the following parametrization for $\theta_{AB}$ as a function of its constituents,

$$\theta_{AB} = \max(\theta_A, \theta_B) + f \min(\theta_A, \theta_B), \tag{3.12}$$

where $f$ is referred to as the “fractionality parameter” which facilitates incorporating “fractional additivity” assumption as opposed to full additivity which corresponds to $f = 1$, i.e. $\theta_{AB} = \theta_A + \theta_B$. While, Thorlund *and others* (2017) focus on cases where $f \in (0, 1)$ they do not restrict $f$ under the model to allow for disagreements between prior and data. We consider the following model including the prior on the parameters, $f$, $\theta_A$ and $\theta_B$ and the likelihood,

$$\begin{aligned} & f \sim \mathcal{N}(0.5, 1) \\ & \theta_A \sim \mathcal{N}(0, 100) \quad \theta_B \sim \mathcal{N}(0, 100) \\ & y_n \sim \text{Bernoulli}\left(\frac{1}{1 + \exp(-x_n^T \boldsymbol{\theta})}\right) \end{aligned} \tag{3.13}$$

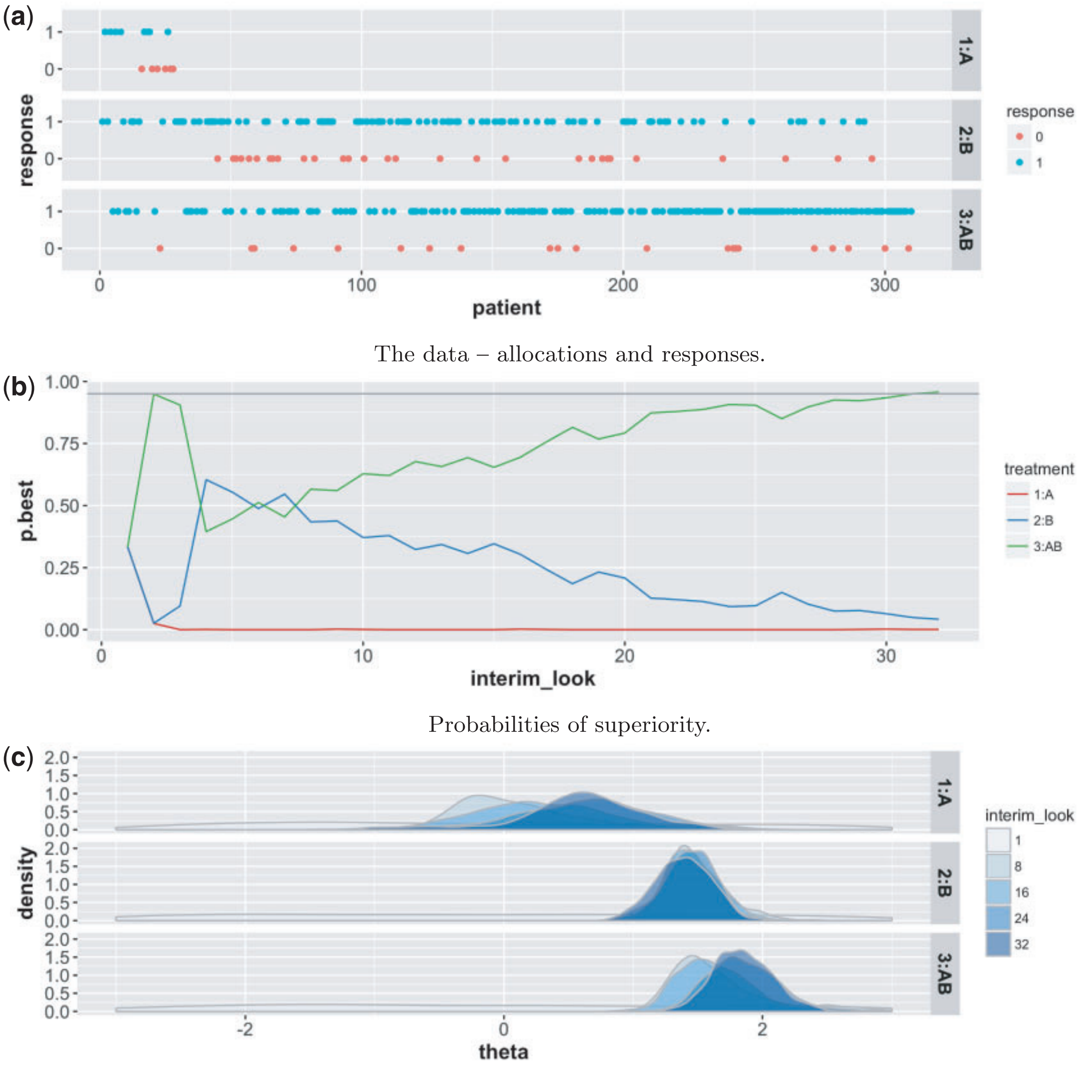
where $\boldsymbol{\theta} = (\theta_A, \theta_B, \theta_{AB})^T$.

The trial is simulated for $f = 0.75$, $\theta_A = 1$ and $\theta_B = 1.2$. The maximum sample size is set to $N = 500$ and the initial sample size is set to $n_0 = 30$. The superiority and inferiority thresholds are set to $u = 0.975$ and $l = 0.05$, respectively. The interim sample size is $n_b = 10$, i.e., posterior updates are made for every 10 patients.

An example simulated RAR trial with fractional additivity assumption is presented in Figure 2. Figure 2a summarizes the data: each panel corresponds to a treatment arm. The allocations are performed with equal probabilities for the three arms up to 30 patients. Treatment A is dropped at the third interim look, i.e. as soon as adaptation is allowed (Figure 2b). Treatments B and AB remain in the trial until the trial is terminated with a total of 310 patients concluding that AB is the superior treatment. The evolving posterior distributions of the parameters at five equally spaced interim looks are presented in Figure 2c. The posterior distribution is diffuse for treatment A since fewer data are used to estimate this parameter. For the two other arms the gradual concentration of the posterior is evident as more data are collected.

## 4. Comparison with Markov chain Monte Carlo and variational Bayes

In this section, the proposed SMC algorithm is compared with MCMC and VB approximation through a small simulation study. The comparisons are made for a three-arm trial design with parameter values $\boldsymbol{\theta} = (0.1, 1.2, 0)$. The methods are to be compared in terms of the precision of estimates of the probabilities of superiority. To reduce data variability, no adaptive decision rules are included in the design: the design



The data – allocations and responses.

Probabilities of superiority.

Posterior distributions of parameters.

Fig. 2. Results for Example 2. (a) The data: each panel corresponds to a treatment arm and each dot represents the count response for a patient who received the treatment. A response with value 1 is positive while zero indicates a negative response. (b) The performance of treatment arms: each curve shows the evolution of the probability of superiority of the corresponding treatment as more data is collected. The horizontal line shows the upper threshold that if reached by any of the treatment curves the trial is stopped for efficacy. (c) Posterior distributions: each panel corresponds to a treatment arm. The posterior kernel density estimates of the corresponding parameters are plotted for five selected interim looks.

consists of a fixed sample size of 150 and 30 equally spaced interim updates (i.e. every five patients). The allocation probabilities remain equal through the course of the trial. Data is generated for 100 trials from the Bernoulli likelihood in (3.11) and analyzed in 30 steps (interim looks) by the SMC in Algorithm 1, MCMC, and VB. We use Hamiltonian Monte Carlo (HMC) implemented in Rstan (Stan Development Team, 2018) considering that HMC is one the most efficient MCMC methods available. Also, for VB the

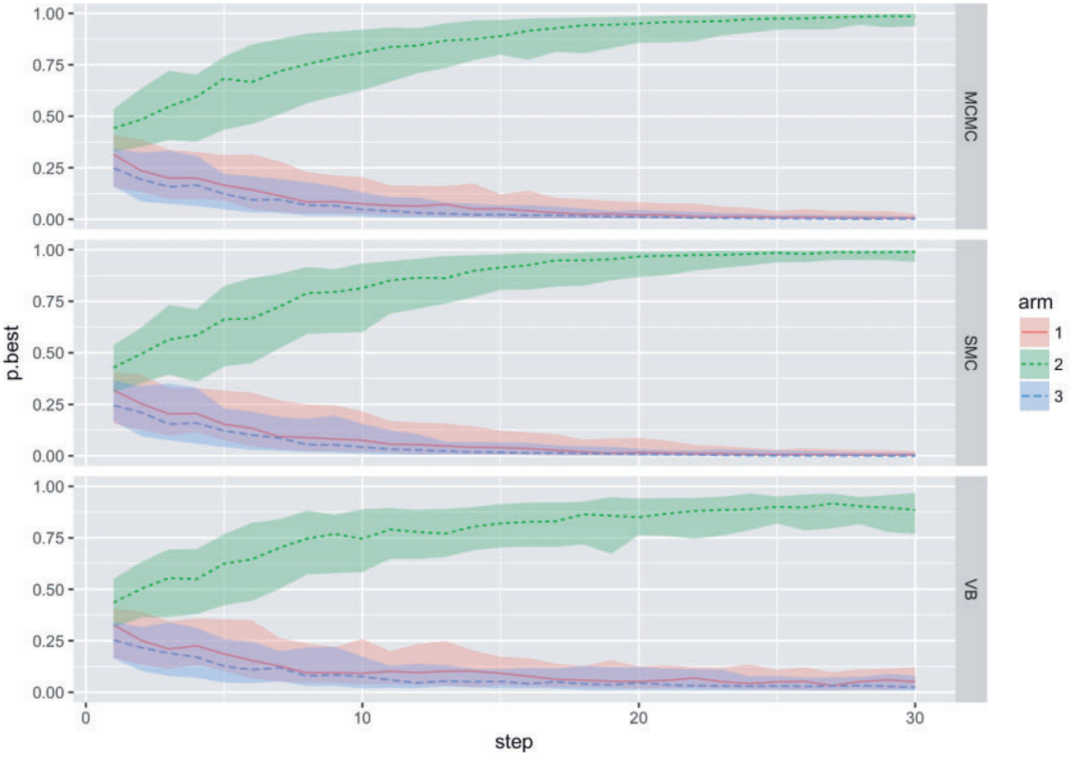


Fig. 3. Comparison of MCMC, SMC, and VB in estimates of probabilities of superiority; the lines connect the medians of the estimates for the 30 interim updates and the bands show the middle 50% of the estimates.

black-box VB implementation in Rstan is used. The updates in HMC and VB are performed by estimating the moments of the Gaussian prior from the posterior sample obtained from the previous step.

Figure 3 summarizes the distribution of the probability of superiority estimates for the three arms by the two methods. The solid lines connect the medians of the estimates over the 30 updates and the bands are obtained by 25% and 75% quantiles of the estimates. The estimates under the three methods are calculated from the SMC and HMC posterior samples, and a sample drawn for the approximated VB posterior. The size of all posterior samples is 1000. As evident in the figure, the estimated probabilities of superiority converge to 1 for the superior arm and to zero for the inferior arms faster by sampling than by approximation and the estimates obtained by sampling the exact posterior have smaller variances than those obtained from the approximate posterior. The estimation variance is slightly lower for SMC in compare with MCMC.

Figure 4 shows the distribution of final estimates of the probability of superiority for the second arm (superior treatment arm) obtained by the three methods. Variability of the estimates is lowest for SMC with MCMC following in the second place, while VB estimates are distributed more diffusely. Considering the true value of the probability of superiority for the superior arm as 1, the root mean squared error at the final step is calculated for the three methods as,

$$RMSE = \frac{1}{100}\sum_{i=1}^{100}\left(\mathrm{P}\left(\theta_2 = \max(\theta_1,\theta_2,\theta_3) \mid \mathbf{y}_{150}\right) - 1\right)^2, \tag{4.14}$$

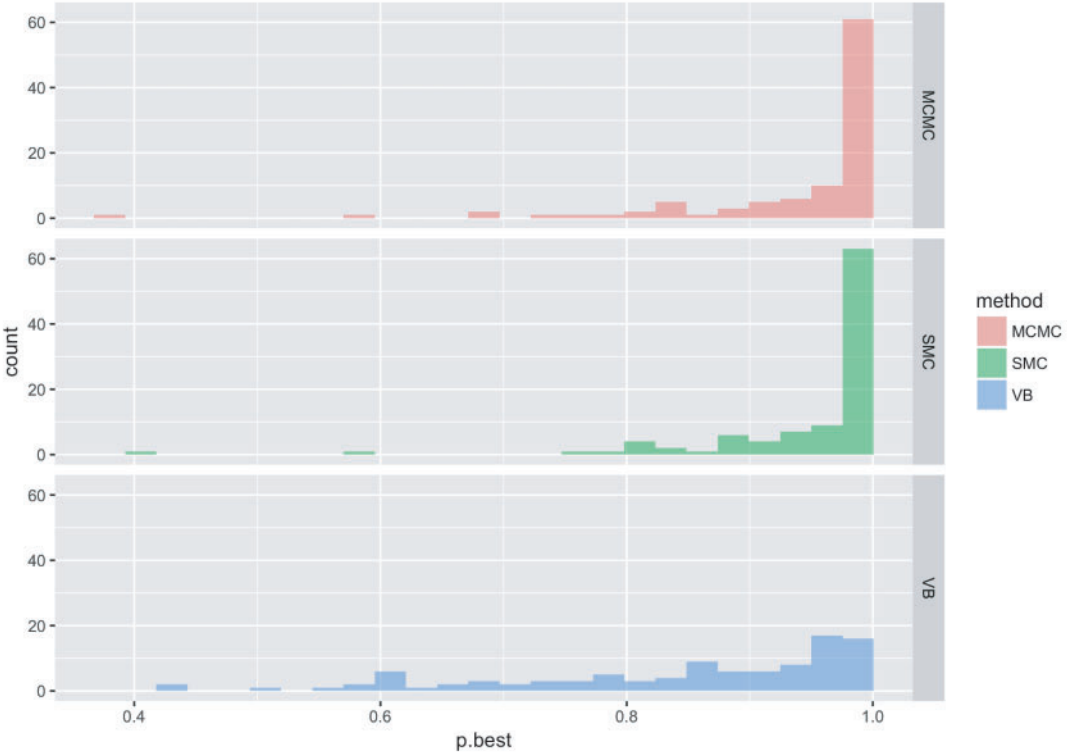


Fig. 4. The distribution of the estimated probability of superiority for the superior arm at the final step for the three methods.

where $\mathbf{y}_{150}$ is the data of 150 patients at the 30th interim step. The smallest RMSE is that of the SMC (0.10) with MCMC (0.11) following with a small difference while the RMSE obtained for VB (0.21) is twice as large.

In addition to estimation precision, the computational complexity of the methods should be considered. Suppose that the goal is to simulate a trial with $N$ interim updates (including the final analysis) and the output is a sample of size $M$ from the posterior (or approximate posterior) of model parameters. The computational complexity for Algorithm 2 is $\mathcal{O}(NM)$ and the algorithm is embarrassingly parallel. MCMC algorithms vary in their time complexities. For the HMC algorithm implemented in Stan the time complexity depends highly on the geometry of posterior. For the simple example used for the comparisons the run time of Stan sampling function was very close to the SMC update parallelized over four processors (about 0.1 CPU seconds). Note that MCMC algorithms are generally more challenging to parallelize. As for variational inference, the time complexity only depends on the number of interims, $\mathcal{O}(N)$. It is worth mentioning, however, that the derivation/optimization cost increases with the complexity of the model.

## 5. Discussion

In this article, we propose an efficient algorithm for simulating a RAR trial design, while taking advantage of the flexibility of Bayesian framework. While Bayesian methods are becoming more popular in design and analysis of clinical trials, the statistical models used in practice are restricted to conjugate models where the posterior can be obtained analytically. Adding any layer of complexity to the model that breaks the conjugacy would require costly computations for performing the Bayesian update. Computational cost and implementation difficulties are therefore the main reason that the Bayesian framework, despite providing a natural setting for updating results, is not yet widely adopted in practice for clinical trials.

Trial simulation is specifically important for more complex adaptive designs since the design properties cannot be analytically investigated in most interesting design scenarios. Therefore, exploration of various criteria such as power and Type I error rate as well as cost and ethics related criteria such as total sample size, overall proportion of failures, and proportion of patients allocated to an inferior arm is performed through simulation studies. While this article does not aim at directly addressing a comprehensive simulation study to investigate the properties of a specific design, the proposed efficient algorithm for simulating a single trial for any adaptive design enhances the efficiency of such simulation studies.

Our proposed algorithm uses SMC for updating samples that are drawn from the posterior distribution of the model parameters (generally, including treatment and covariate effects plus any hyperparameters) given the cumulative data through the course of the trial. SMC is a more efficient and easily parallelizable alternative to MCMC and bypasses arbitrary and hands on tuning questions that arise in MCMC for updating a posterior sample. In addition SMC samplers are known to effectively capture features in the posterior surface such as multimodality, low probability regions and isolated peaks. Implementation of SMC is however a challenging task and requires problem specific tuning. The SMC algorithm laid out in the article is specifically tailored for simulation of RAR trial designs.

We showcase the implementation of the proposed algorithm by simulating trials with two example modeling scenarios. The first example represents a simple but non-conjugate model, while in the second example prior dependence assumptions result in a model for which a conjugate prior choice is not available.

The performance of the proposed algorithm, in terms of the accuracy and precision of the probabilities of superiority obtained from posterior samples, is compared with HMC and black-box VB implemented in RStan. The proposed SMC algorithm stands in the first place in terms of precision with HMC and VB in the second and third place, respectively. In terms of computation time SMC is comparable to HMC, which is one of the most efficient MCMC algorithms while VB, being an approximation method, is a fast alternative to any sampling technique. Considering a combination of performance, ease of implementation and run time, we consider the proposed algorithm as a promising approach for simulating RAR trials.

## 6. Software

Software in the form of R code including the examples in the article is available at the public GitHub repository, https://github.com/sgolchi/SMC.RAR.

## Acknowledgments

The authors would like to thank Dr Edward Mills for his support and consultations. *Conflict of Interest*: None declared.

## References

Atkinson, A. C. and Biswas, A. (2014). *Randomized Response-adaptive Designs in Clinical Trials*. New York: CRC Press.

Berry, S. M., Carlin, B. P., Lee, J. J. and Muller, P. (2010). *Bayesian Adaptive Methods for Clinical Trials*. New York: Chapman and Hall/CRC Biostatistics Series.

Brown, A. R., Gajewski, B. J., Aaronson, L. S., Mudaranthakam, D. P., Hunt, S. L., Berry, S. M., Quintana, M., Pasnoor, M., Dimachkie, M. M., Jawdat, O. *and others*. (2016). A bayesian comparative effectiveness trial in action: developing a platform for multisite study adaptive randomization. *Trials* **17**, 428–437.

Cheng, Y. and Shen, Y. (2005). Bayesian adaptive designs for clinical trials. *Biometrika* **92**, 633–646.

Chopin, N. (2002). A sequential particle filter for static models. *Biometrika* **89**, 539–551.

Chopin, N. (2004). Central limit theorem for sequential Monte Carlo methods and its application to bayesian inference. *Annals of Statistics* **32**, 2385–2411.

Del Moral, P., Doucet, A. and Jasra, A. (2006). Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **68**, 411–436.

Doucet, A., Freitas, N. and Gordon, N. (2013). *Sequential Monte Carlo methods in practice*. New York: Springer Science and Business Media.

He, W., Pinheiro, J. and Kuznetsova, O. M. (2014). *Practical Considerations for Adaptive Trial Design and Implementation*. New York, NY: Springer.

Park, J. W., Liu, M. C., Yee, D., Yau, C., van't Veer, L. J., Symmans, W. F., Paoloni, M., Perlmutter, J., Hylton, N. M., Hogarth, M. *and others*. (2016). Adaptive randomization of neratinib in early breast cancer. *New England Journal of Medicine* **375**, 11–22.

Rosenberger, W. F., Stallard, N., Ivanova, A., Harper, C. N. and Ricks, M. L. (2001). Optimal adaptive designs for binary response trials. *Biometrics* **57**, 909–913.

Saville, B. R. and Berry, S. M. (2016). Efficiencies of platform clinical trials: a vision of the future. *Clinical Trials* **13**, 358–366.

Stan Development Team. (2018). RStan: the R interface to Stan. R package version 2.17.3. http://mc-stan.org.

Thall, P. F. and Wathen, J. K. (2007). Practical bayesian adaptive randomization in clinical trials. *European Journal of Cancer* **43**, 589–866.

Thorlund, K., Golchi, S. and Mills, E. (2017). Bayesian adaptive clinical trials of combination treatments. *Contemporary Clinical Trials Communications* **8**, 227–233.

Zhang, L. and Rosenberger, W. F. (2006). Response-adaptive randomization for clinical trials with continuous outcomes. *Biometrics* **62**, 562–569.

[*Received September 19, 2017; revised August 3, 2018; accepted for publication August 10, 2018*]