Making SMART decisions in prophylaxis and treatment studies

Robert K. Mahar\textsuperscript{1,2,*}, Katherine J. Lee\textsuperscript{2,3}, Bibhas Chakraborty\textsuperscript{4,5,6}, Agus Salim\textsuperscript{1,7,8}, and Julie A. Simpson\textsuperscript{1}

\textsuperscript{1}Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, Faculty of Medicine, Dentistry, and Health Sciences, University of Melbourne, Parkville, Victoria, Australia

\textsuperscript{2}Clinical Epidemiology and Biostatistics Unit, Murdoch Children’s Research Institute, Parkville, Victoria, Australia

\textsuperscript{3}Department of Paediatrics, Melbourne Medical School, University of Melbourne, Parkville, Victoria, Australia

\textsuperscript{4}Centre for Quantitative Medicine, Duke-NUS Medical School, Singapore

\textsuperscript{5}Department of Statistics, Faculty of Science, National University of Singapore, Singapore

\textsuperscript{6}Department of Biostatistics and Bioinformatics, Duke University, Durham, North Carolina, United States of America

\textsuperscript{7}School of Mathematics and Statistics, Faculty of Science, University of Melbourne, Parkville, Victoria, Australia

\textsuperscript{8}Baker Department of Cardiometabolic Health, Melbourne Medical School, Faculty of Medicine, Dentistry and Health Sciences, University of Melbourne, Parkville, Victoria, Australia

\*Corresponding author: Robert Mahar, Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, 207 Bouverie Street, University of Melbourne, Victoria, 3053 Australia. robert.mahar@unimelb.edu.au

March 25, 2022
Abstract

The optimal prophylaxis, and treatment if the prophylaxis fails, for a disease may be best evaluated using a sequential multiple assignment randomised trial (SMART). A SMART is a multi-stage study that randomises a participant to an initial treatment, observes some response to that treatment and then, depending on their observed response, randomises the same participant to an alternative treatment. Response adaptive randomisation may, in some settings, improve the trial participants’ outcomes and expedite trial conclusions, compared to fixed randomisation. But ‘myopic’ response adaptive randomisation strategies, blind to multistage dynamics, may also result in suboptimal treatment assignments. We propose a ‘dynamic’ response adaptive randomisation strategy based on Q-learning, an approximate dynamic programming algorithm. Q-learning uses stage-wise statistical models and backward induction to incorporate late-stage ‘payoffs’ (i.e. clinical outcomes) into early-stage ‘actions’ (i.e. treatments).

Our real-world example consists of a COVID-19 prophylaxis and treatment SMART with qualitatively different binary endpoints at each stage. Standard Q-learning does not work with such data because it cannot be used for sequences of binary endpoints. Sequences of qualitatively distinct endpoints may also require different weightings to ensure that the design guides participants to regimens with the highest utility. We describe how a simple decision-theoretic extension to Q-learning can be used to handle sequential binary endpoints with distinct utilities. Using simulation we show that, under a set of binary utilities, the ‘dynamic’ approach increases expected participant utility compared to the fixed approach, sometimes markedly, for all model parameters, whereas the ‘myopic’ approach can actually decrease utility.

Keywords: adaptive trials, adaptive randomisation, Bayesian, COVID-19, decision theory, Q-learning, dynamic programming, sequential multiple assignment randomised trial, Stan, dynamic treatment regimen, adaptive treatment policy
1 Background

Compared to the general population, immunocompromised patients who are exposed to an infectious pathogen may be more likely to become infected, to develop severe disease, and to experience more severe treatment side-effects. Providing these patients first with prophylaxis, to reduce the chance of infection, followed by treatment if subsequently infected, is a common strategy to limit the damage caused by infectious diseases.

Although the standalone efficacy of prophylaxes and treatments may be well-established, the efficacy of their use in sequence may not be. For example, a relatively inefficacious prophylaxis that increases the efficacy of subsequent treatments may be eschewed in favour of a prophylaxis that may be only slightly more efficacious yet decrease subsequent treatment efficacy substantially. Likewise, a later-stage treatment may only be efficacious if a particular prophylaxis precedes it or is given conditional on some particular prognostic prophylaxis outcome being met (for example, disease severity).

Formalising such prophylaxis and treatment strategies with dynamic treatment regimens is useful. A dynamic treatment regimen is simply an algorithm that specifies sequences of treatment decision rules that depend on time-varying patient characteristics [1, 2]. Many dynamic treatment regimens can be candidates for a given clinical condition. The statistical problem is to identify the ‘optimal’ regimen from the candidates under consideration.

Obtaining the data to evaluate the efficacy of a dynamic treatment regimen is best done using a sequential multiple assignment randomised trial (SMART). A SMART is designed to randomise a patient to an initial treatment, observe some response to that treatment and then, depending on the observed response, possibly randomise the same patient to an alternative treatment. Although observational data can be used to optimise dynamic treatment regimens in principle, in practice it is not often done because it is exceedingly difficult to control for time-dependent confounding, among other things [3].

We are motivated by the study ‘COVID-19 prevention and treatment in cancer: a sequential multiple assignment randomised trial’ (C-SMART) [4], an ongoing Australian SMART that
was designed to evaluate COVID-19 prophylaxis and subsequent treatment in high-risk cancer
patients. Although designed as a SMART, C-SMART primarily aimed to estimate standalone
prophylaxis and treatment effects, rather than to evaluate dynamic treatment regimens. We
consider a stylised version of this trial that has a single initial prophylaxis stage, followed by
a single treatment stage, with a primary aim to optimise dynamic treatment decisions.

In this paper, we extend our study design to include Bayesian response adaptive randomi-
sation at each stage of the SMART. Although non-Bayesian response adaptive designs are
possible, Bayesian methods are useful because the interim inferences that are required in
an adaptive trial can be made by simply updating the posterior distribution as new data
become available. In contrast, frequentist designs often require complex adjustments and are
arguably more difficult to interpret.

Because early- and late-stage interventions may interact with one another, a ‘myopic’ response
adaptive randomisation strategy, one that ignores between stage information (e.g. outcomes
of the early-stage intervention), may lead to an adaptive design that favours suboptimal
regimens where only early-stage efficacy is apparent, leading to worse outcomes for trial
participants than a fixed design. In principle, this challenge can be overcome by using
statistical models for evaluating dynamic treatment regimens as the basis for the response
adaptive randomisation, herein referred to as ‘dynamic’ response adaptive randomisation,
but few guiding examples exist, and the extent that a ‘myopic’ approach is suboptimal is
not known [5, 6, 7, 8, 9]. We therefore aim to evaluate a simple two-stage design across a
comprehensive range of scenarios to gauge the relative performances of both the ‘myopic’
and ‘dynamic’ response adaptive randomisation approaches, relative to a fixed design.

We propose to use an approximate dynamic programming method, commonly known as Q-
learning [1, 10], to perform the ‘dynamic’ response adaptive randomisation. Our approach
is conceptually similar to that of Lee et al. who used dynamic programming to implement
decision-theoretic ‘dynamic’ response adaptive randomisation in dose finding studies with
joint binary [8] and ordinal [9] intermediate efficacy and toxicity endpoints, and Cheung et
al. [7], who used Q-learning to guide response adaptive randomisation in an implementation
design with only a single, continuous end-of-study endpoint.
Two key challenges that arise in evaluating dynamic prophylaxis and treatment regimens are that the endpoints at each stage are both binary and also distinct from one another, by which we mean that they are qualitatively different at each stage. Intermediate binary endpoints pose a challenge because Q-learning with intermediate outcomes relies on maximising the expectation of a ‘pseudo-outcome’, that is, the sum of the intermediate outcome and the maximised expectation of the later stage outcomes, conditional on patient history. Although the classical implementation of Q-learning is flexible enough to be used when there is either a single end-of-study binary endpoint or sequences of continuous endpoints, sequences of intermediate binary endpoints present a challenge because the pseudo-outcome at the first stage is an awkward sum of a binary endpoint and a probability \[\pi\]. We propose that the approach of Lee et al. \[8, 9\] can be adapted to overcome this challenge, thus identifying a methodological connection that has not been made thus far. Further, we propose that the decision-theoretic approach can be used with distinct endpoints, which also pose a challenge because, even if we can deal with intermediate binary endpoints, treating each of these endpoints as equal in utility may be hard to justify in some circumstances. For example, in C-SMART, the endpoints were COVID-19 infection in stage one and all-cause mortality in stage two. This decision-theoretic approach works seamlessly with the proposed Bayesian adaptive design, and thus provides further motivation for the use of Bayesian methods. In contrast to the existing studies, we evaluate the decision-theoretic approach across the entire range of parameter values to better guide practical application and further development of the method.

This article proceeds by summarising the motivating trial design in Section 2, describing the Bayesian decision-theoretic design in Section 3 that will allow sequential decisions to be optimised using intermediate and distinct binary endpoints, and describing a simple response adaptive randomisation strategy in Section 4. In Section 5, we present the results of a simulation study and make concluding remarks in Section 6.
2 Motivating trial design

C-SMART is a prospective, multi-centre, double-blind, placebo-controlled, Bayesian adaptive SMART. Our stylised version, outlined in Figure 1 includes two randomised ‘substudies’ that occur in sequence: a pre-exposure prophylaxis substudy, and a COVID-19 treatment substudy. Participants entering the trial do so within the pre-exposure prophylaxis substudy, where they are randomised to receive either an anti-viral prophylaxis or a placebo with a binary endpoint of a positive COVID-19 diagnosis. If a patient develops COVID-19, then they are randomised to receive either an active COVID-19 treatment or a placebo, with a primary binary endpoint of all-cause mortality.

3 Decision-theoretic model

3.1 Sequential actions and their utilities

Here we describe the general framework for our statistical decision problem [12,13]. Let the prophylaxis and disease treatment stages be indexed respectively by \( k = 1, 2 \), from which we define stage-specific actions, \( a_k \in A_k = \{0, 1\} \) and binary outcomes, \( y_k \in Y_k = \{0, 1\} \).

At each stage we define decision rules \( d_k(h_k) \in D_k \) as functions that map patient history \( h_k \) to certain course of action \( a_k \) such that \( d_k(h_k) \rightarrow a_k \), where \( h_k \) includes the patient history leading up to, but not including, \( a_k \). In our case, we assume that \( h_1 = \emptyset \), that is we do not consider history in the first stage, and \( h_2 = \{a_1, y_1\} \times (1 - m) \). Note that \( d_k(h_k) \) will be simply denoted as \( d_k \) hereafter. Also note a design constant \( m \in \{0, 1\} \) is included to induce either ‘dynamic’ or ‘myopic’ approaches by setting \( m = 0 \) or \( m = 1 \) respectively.

Because the utility of an outcome may depend on what treatments were allocated and patient history, we assign each possible realisation of history, treatment, and outcome at the terminal stage a specific utility, \( u(h_k, a_k, y_k) \). The utility of making a decision \( d_k \) is defined as the sum
of the utilities of each outcome under that decision, weighted by the probability of that outcome. We use a simple utility function that is defined generally as follows:

\[ U_k(\pi_k, h_k, d_k, y_k) = u(h_k, a_k, y_k = 0) \times (1 - \pi_k) + u(h_k, a_k, y_k = 1) \times \pi_k, \]  

(1)

where \( u(h_k, a_k, y_k = 0) \) and \( u(h_k, a_k, y_k = 1) \) are the utilities prescribed to the events where \( y_k = 0 \) and \( y_k = 1 \) for a given pair of \( h_k \) and \( a_k \), and \( \pi_k \) is the probability of an event in stage \( k \), conditional on \( h_k \) and \( a_k \).

The optimal decisions at each stage, \( d_k^{opt} \), are defined using dynamic programming, which identifies the optimal stage-specific decision on the basis of the expected utility of that decision and the assumption that all subsequent decisions are optimal. We proceed with a similar approach to Lee et al. [9] by choosing the decision that maximises the expected posterior utility taken with respect to the posterior probabilities of the events at each stage, \( p(\pi_k \mid h_k, a_k) \). We use the familiar Q-function notation of Q-learning to simplify what follows. For any stage, we define a general Q-function that represents the posterior expected utility, as follows:

\[ Q_k(\pi_k, h_k, d_k, y_k) = E_{\pi_k}\left[ U_k(\pi_k, h_k, d_k, y_k) \mid h_k, a_k, y_k \right] \]

\[ = \int U_k(\pi_k, h_k, d_k, y_k)p(\pi_k \mid h_k, a_k) \ d\pi_k. \]

(2)

In general, \( d_k^{opt} = \arg \max_{d_k} Q_k(\pi_k, d_k, h_k, y_k) \) is the optimal stage \( k \) decision, conditional on \( h_k \).

The stage two utility function is of the same form as (1), but with \( k = 2 \),

\[ U_2(\pi_2, h_2, d_2, y_2) = u(h_2, a_2, y_2 = 0) \times (1 - \pi_2) + u(h_2, a_2, y_2 = 1) \times \pi_2. \]

(3)

The stage one utility function is defined similarly, however the utility of a stage one infection outcome \( y_1 = 1 \), is replaced by the expected utility under the optimal decision in the second
stage,

\[ U_1(\pi_1, h_1, d_1, y_1) = u(h_1, a_1, y_1 = 0) \times (1 - \pi_1) + \max_{d_2} Q_2(\pi_2, h_2, d_2, y_1) \times (1 - m) \times \pi_1. \tag{4} \]

In other words, the utility function is the sum of the utility of no infection, weighted by the probability of no infection, and the posterior expected utility of making the likely best treatment decision upon infection, weighted by the probability of developing infection.

Note also in (4) the constant \( m \in \{0, 1\} \). If \( m = 0 \) then the second term in the summand in (4) is unchanged, in other words, all of the later-stage expected utilities are included. If \( m = 1 \) then the second summand equals zero, in other words, the later-stage expected utilities are not included and the stage-specific utility for the event is zero (i.e. the myopic approach).

### 3.2 Probability model

In general, the posterior distributions defined in (2), \( p(\pi_k \mid h_k, a_k) \) are, by Bayes Theorem, proportional to the product of the likelihood \( L_k(\pi_k, h_k) \) and the prior probability density of \( \pi_k \), \( p(\pi_k) \). The outcomes \( y_k \) are modelled by Bernoulli probabilities with a logistic link and parametric model indexed by stage \( k \):

\[
y_k \sim \text{Bernoulli}(\pi_k) \\
\pi_k = \logit^{-1}\left[f_k(\beta_k, h_k, a_k)\right],
\]

where \( \pi_k \) denotes the probability of an event \( y_k = 1 \) and \( f_k(\beta_k, h_k) \) is a linear function of history up to stage \( k \), and the \( \beta_k \) are vectors containing the set of model parameters used for
each stage $k$. In general, we can define our posterior distribution as

$$p(\pi_k \mid h_k, a_k) \propto \prod_{i \in I_k} \mathcal{L}_k(\pi_k, y_{ik}) p(\pi_k) \propto \prod_{i \in I_k} f_k(\beta_k, h_k)^{y_{ik}} (1 - f_k(\beta_k, h_k))^{1 - y_{ik}} p(\pi_k)$$

(6)

where $y_{ik}$ is the specific outcome observed for patient $i \in \{1, \ldots, N\}$ in stage $k$.

4 Response adaptive randomisation

The overwhelming majority of trial designs randomise between treatments with fixed (and usually equal) probability and only analyse data once the study has been completed. In a SMART design the treatment for patient $i$ at stage $k$ (i.e. $a_{ik}$) is typically allocated randomly with equal probability conditional on $h_{ik}$. This means that, by the end of the trial, many patients will have received sub-optimal treatments.

Response adaptive randomisation is an approach that aims to improve the probability that trial participants are given the most promising treatment at randomisation by continually reweighting the randomisation ratio in favour of the more promising treatments. In the extreme case of a very poorly performing treatment, this approach can simply stop randomly allocating new patients to that treatment.

Our approach uses the posterior expected utility to weight the random allocation of treatment in proportion to the total expected utility for the decision being considered:

$$p(a_{ik} \mid h_{ik}) = \frac{Q_k(\pi_k, h_{ik}, d_k = a_{ik}, y_{ik})^c}{\sum_{d_k} Q_k(\pi_k, h_{ik}, d_k, y_{ik})^c}$$

(7)

Note the exponent $c$. Where $c = 0$ the probability of random allocation is equal and fixed for all treatments and where $c = 1$ the probability of random allocation is proportional to the relative expected utility of that treatment.
5 Simulation study

Having defined the decision-theoretic model to guide the outcome adaptive randomisation, we proceed by examining the finite-sample properties of our approach using simulated data. The details of the simulation study are provided in the following subsections.

5.1 Design types

The four main design types we considered in this simulation study are summarised in Table 1. If the design was fixed, then the trial continued until the final outcome was observed at which point the data were analysed. If the design was adaptive, then the trial conducted four scheduled analyses at equal timepoints corresponding to observed outcomes, with randomisation probabilities updated, according to (7), for each of the first three of the scheduled analyses, before continuing recruitment. If the design was ‘myopic’ then the response adaptive randomisation algorithm ignored between-stage information. If the design was ‘dynamic’ then the response adaptive randomisation algorithm used between-stage information.

5.2 Data generation and utilities

Data were randomly generated using the Bernoulli distribution with values of $\pi_k$ as described in Table 2. The infection probabilities at the first stage $r_0$, for placebo, and $r_1$, for prophylaxis, were specified from the following set of 21 elements:

$$R = \{0, 5, 10, \ldots, 100\}/100,$$

for example, $(r_0, r_1) = (0.1, 0.3)$. Likewise, the death probabilities in the second stage either following stage one placebo, $s_0$, or following stage one prophylaxis, $s_1$, were specified from
the following set of 8 elements:

\[ S = \{5, 10, 20, 40, 60, 80, 90, 95\}/100, \]  

for example, \((s_0, s_1) = (0.45, 0.5)\). Note that there are \(21^2 = 441\) and \(8^2 = 64\) unique pairs of \((r_0, r_1)\) and \((s_0, s_1)\), respectively, resulting in 28224 data generating scenarios.

Also described in Table 2, we assumed only a single set of utilities, whereby participants who died received zero utility and participants who lived received unit utility, irrespective of their histories and treatments. Although this is undoubtedly a strong assumption, in life versus death examples it is arguably reasonable to assume that death has no utility and survival is preferred regardless of whether there was infection and subsequent treatment, at least as a demonstrative approximation. This utility set also has the useful property of allowing us to examine the performance of the various design approaches whereby the optimisation problem is reduced to simply maximising the probability of survival.

We set a maximum sample size of 2,000 patients, which was approximately equal to that of the original C-SMART study, however we note that for our evaluation any sample size could be used. To simplify the simulations, patients were assumed to be recruited at a single time point and their outcomes observed immediately (or within a very short time frame). For each trial design and simulation scenario, we simulated 10 trials. Based on pilot simulations, we judged 10 trials as adequate to allow both the results to be interpreted meaningfully and to complete the required computations within a reasonable amount of time.

### 5.3 Analysis

For our motivating example, we computed the posterior expected utilities as described in (1)–(5) using the following linear models for (5):

\[
\begin{align*}
f_1(\beta_1, a_1) &= \beta_{10} + \beta_{11} a_1 \\
f_2(\beta_2, h_2, a_2) &= \beta_{20} + \beta_{21} a_2 + \beta_{22} a_1 + \beta_{23} a_1 a_2.
\end{align*}
\]  

(10)
where $a_1, a_2 \in \{0, 1\}$, and $a_1 = 1$ and $a_1 = 0$, indicates prophylaxis and placebo in the first stage, respectively, and $a_2 = 1$ and $a_2 = 0$, indicates treatment and placebo in the second stage, respectively.

5.4 Evaluation

We evaluated the performance of each four trial designs under each simulation scenario using the empirical mean utility of the trial participants, $\bar{u}$ averaged over the 10 simulated trials, for which we use the shorthand $\bar{u}$. We then evaluated the $\tilde{u}$ of the adaptive-myopic design relative to the fixed-myopic design, and the $\bar{u}$ of the adaptive-dynamic design relative to the fixed-dynamic design.

5.5 Computational methods

Markov-chain Monte Carlo (MCMC) methods implemented in Stan, a probabilistic programming language [14], were used to numerically approximate the joint posterior distributions of the parameters for each model. The posterior distribution of each parameter of the analysis models was sampled using the Hamiltonian Monte Carlo algorithm implemented within Stan, called from the R software environment [15]. For each parameter, we ran four MCMC chains in parallel for a ‘warm-up’ phase followed by a ‘sampling’ phase of 1000 iterations each, resulting in 4000 samples from the posterior distribution.

5.6 Results

We summarised the results using empirical mean utilities, $\bar{u}$, for trial participants (i.e. the average utility of participants, averaged over ten simulations). In each of the figures 2 and 3 we show results for all unique 441 combinations of infection probabilities in the placebo arm (inner x-axes) and the prophylaxis arm (inner y-axes), for each of the 64 unique combinations
of death probabilities following stage 1 placebo (upper margin) and following stage 1 pro-
phylaxis (right margin). Figure 2 shows the mean empirical utilities of the design that used
myopic response adaptive randomisation, relative to the fixed design. Figure 3 shows the
mean empirical utilities of the design that used dynamic response adaptive randomisation,
relative to the fixed design. Note that where the relative $\bar{u}$ was greater than one, the response
adaptive design had better utility than the fixed design, and where the relative $\bar{u}$ was less
than one, the response adaptive design had worse utility than the fixed design. Clearly, where
the relative $\bar{u}$ equalled one, the adaptive and fixed designs had the same utility.

Myopic response adaptive randomisation favoured the most efficacious prophylaxis, regardless
of the subsequent treatment outcomes. In many cases this did not result in relative $\bar{u}$ values
materially lower than one. For example, if for both prophylaxis histories the infection rates
were both very low, or the death rates were similar, then $\bar{u}$ was typically close to or greater
than one. But where the death rates differed between stage one histories, a very small
difference between infection rates could ultimately weight first stage randomisation in favour
of a ultimately more fatal treatment regimen and therefore lower utilities, and we see this
manifested in areas coloured definitively red in Figure 2.

Dynamic response adaptive randomisation, by incorporating patient histories, treatments,
and expectations about future outcomes in the randomisation algorithm, ensured that the
empirical expected utility was either greater than or approximately equal to the fixed design
for all possible data generating scenarios, as demonstrated by the absence of red coloured
areas in Figure 3. Note the relative utility was approximately equal to one in some scenarios.
This is obvious wherever the infection or death probabilities were approximately similar (i.e.
$r_0 \sim r_1$ and $s_0 \sim s_1$), or where the probabilities and death probabilities were very different
(e.g. $r_0 \ll r_1$ and $s_0 \ll s_1$). Otherwise, the dynamic response adaptive randomisation
algorithm tended to guide participants towards the most efficacious regimens which resulted
in, often substantially higher, expected utilities, and this is reflected by the areas coloured
blue in Figure 3.
6 Discussion

We have shown how a Bayesian decision-theoretic dynamic programming algorithm can be used to model intermediate binary endpoints in a simple two-stage prophylaxis and treatment SMART. With a simple binary set of utilities, assigning zero utility to death and unit utility to survival, our approach can be seen as one that simply maximises survival probabilities. Such a utility set may be reasonable in other clinical contexts. Our approach provides a solution to the longstanding challenge for Q-learning of how to model sequential binary endpoints [1, 16]. More complicated and perhaps more relevant SMART trials may now be designed and analysed. Uniquely, our approach allows any utility set to be assigned to different realisations of histories, treatments, and outcomes, allowing decisions based on distinct intermediate binary endpoints to be optimised. When we use the ‘dynamic’ SMART design with response adaptive randomisation the expected utility of trial participants is either equal or higher, sometimes substantially, to that of a SMART with fixed randomisation. However, we have shown that a ‘myopic’ SMART design with response adaptive randomisation can substantially decrease the expected utility of trial participants relative to a SMART design with fixed randomisation, and for this reason we recommend against using ‘myopic’ response adaptive randomisation for this type of SMART.

Either the myopic or dynamic approach can easily be extended to multiple stages and/or treatments, and indeed can be applied for SMARTs that re-randomise after all intermediate outcomes, however such designs can become quickly become intractable. Lee et al. [9] note that one approach around this intractability is to make a Markovian assumption, in other words, to ignore history. Our ‘myopic’ approach is not strictly Markovian because it also ignores information in later stages, however our results suggest that investigators should proceed with caution if intending to make the Markovian assumption.
Data and code availability

An R package containing all code used for this simulation study is available online at: https://github.com/robert.mahar/dynamicar. Alternatively, the R package is available from the authors upon request.

Acknowledgements

This work was supported by the Australian National Health and Medical Research Council (NHMRC) Centre for Research Excellence grants to the Victorian Centre for Biostatistics (ID 1035261) and to the Australian Trials Methodology Research Network (ID 1171422). Robert Mahar acknowledges the support of the University of Melbourne Early Career Researcher Grant Scheme and Australian Trials Methodology Research Network Seed Funding Program. Julie Simpson is funded by a NHMRC Investigator Grant (ID 1196068) and Katherine Lee by a NHMRC Career Development Fellowship (ID 1127984). Bibhas Chakraborty acknowledges start-up funds from Duke-NUS Medical School, Singapore. Research at the Murdoch Children’s Research Institute is supported by the Victorian Government’s Operational Infrastructure Support Program. We thank the C-SMART investigator team for developing the study design that motivated this work.
References

[1] Chakraborty B, Murphy SA. Dynamic treatment regimes. Annual Review of Statistics and Its Application. 2014;1(1):447–464. Available from: http://www.annualreviews.org/doi/10.1146/annurev-statistics-022513-115553

[2] Murphy SA. An experimental design for the development of adaptive treatment strategies. Statistics in Medicine. 2005;24(10):1455–1481. Available from: https://onlinelibrary.wiley.com/doi/10.1002/sim.2022

[3] Mahar RK, McGuinness MB, Chakraborty B, Carlin JB, IJzerman MJ, Simpson JA. A scoping review of studies using observational data to optimise dynamic treatment regimens. BMC Medical Research Methodology. 2021;21(1):39. Available from: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-021-01211-2

[4] ClinicalTrials.gov [Internet]. Bethesda (MD): National Library of Medicine (US). 2000 Feb 29–. Identifier NCT04534725, COVID-19 Prevention and Treatment in Cancer; a Sequential Multiple Assignment Randomised Trial. 2020 Sep 1 [cited 2022 Feb 4];. Available from: //clinicaltrials.gov/ct2/show/NCT04534725

[5] Thall PF, Sung HG, Estey EH. Selecting Therapeutic Strategies Based on Efficacy and Death in Multicourse Clinical Trials. Journal of the American Statistical Association. 2002;97(457):29–39. Available from: http://www.tandfonline.com/doi/abs/10.1198/016214502753479202

[6] Thall PF, Wathen JK. Covariate-adjusted adaptive randomization in a sarcoma trial with multi-stage treatments. Statistics in Medicine. 2005;24(13):1947–1964. Available from: http://onlinelibrary.wiley.com/doi/abs/10.1002/sim.2077

[7] Cheung YK, Chakraborty B, Davidson KW. Sequential multiple assignment randomized trial (SMART) with adaptive randomization for quality improvement in depression treatment program: SMART with Adaptive Randomization. Biometrics. 2015;71(2):450–459. Available from: http://doi.wiley.com/10.1111/biom.12258
[8] Lee J, Thall PF, Ji Y, Müller P. Bayesian Dose-Finding in Two Treatment Cycles Based on the Joint Utility of Efficacy and Toxicity. Journal of the American Statistical Association. 2015;110(510):711–722. Available from: http://www.tandfonline.com/doi/full/10.1080/01621459.2014.926815

[9] Lee J, Thall PF, Ji Y, Müller P. A decision-theoretic phase I–II design for ordinal outcomes in two cycles. Biostatistics. 2016;17(2):304–319. Available from: https://academic.oup.com/biostatistics/article/17/2/304/1744296

[10] Watkins CJCH, Dayan P. Q-learning. Machine Learning. 1992;8(3):279–292. Available from: https://doi.org/10.1007/BF00992698

[11] Moodie EEM, Chakraborty B, Kramer MS. Q-learning for estimating optimal dynamic treatment rules from observational data. Canadian Journal of Statistics. 2012;40(4):629–645. Available from: http://doi.wiley.com/10.1002/cjs.11162

[12] Berger JO. Statistical Decision Theory. Springer Series in Statistics. New York, NY: Springer New York; 1980. Available from: http://link.springer.com/10.1007/978-1-4757-1727-3

[13] Robert CP. The Bayesian choice: from decision-theoretic foundations to computational implementation. 2nd ed. Springer texts in statistics. New York: Springer; 2007.

[14] Team SD. The Stan Core Library; 2018. Available from: http://mc-stan.org

[15] Team RC. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2019. Available from: https://www.R-project.org/

[16] Moodie EEM, Dean N, Sun YR. Q-Learning: Flexible Learning About Useful Utilities. Statistics in Biosciences. 2014;6(2):223–243. Available from: http://link.springer.com/10.1007/s12561-013-9103-z
List of Figures

1  C-SMART design ............................................................... 19
2  Empirical mean utilities of *myopic* SMART with response adaptive randomisation relative to SMART with fixed randomisation ..................... 20
3  Empirical mean utilities of *dynamic* SMART with response adaptive randomisation relative to SMART with fixed randomisation ................. 21
Figure 1: C-SMART design

Entry → 1:1

Placebo

No COVID

COVID

Prophylaxis

1:1

Treatment

Placebo

No COVID

COVID

1:1

Treatment

Placebo

Alive

Dead

Alive

Dead

Alive

Dead

Alive

Dead
Figure 2: Empirical mean utilities of *myopic* SMART with response adaptive randomisation relative to SMART with fixed randomisation.
Figure 3: Empirical mean utilities of *dynamic* SMART with response adaptive randomisation relative to SMART with fixed randomisation.
# List of Tables

| Table | Title | Page |
|-------|-------|------|
| 1     | Adaptive and dynamic design settings | 23   |
| 2     | Data generation and utilities       | 24   |
Table 1: Adaptive and dynamic design settings

|                      | Dynamic  | Myopic   |
|----------------------|----------|----------|
| Fixed randomisation  | $m = 0, c = 0$ | $m = 1, c = 0$ |
| Response adaptive randomisation | $m = 0, c = 1$ | $m = 1, c = 1$ |
## Table 2: Data generation and utilities

| Stage | Prophylaxis | Infected | Treatment | Dead | P(Infected) | P(Dead) | Utility |
|-------|-------------|----------|-----------|------|-------------|---------|---------|
| $k$   | $a_1$       | $y_1$    | $a_2$     | $y_2$| $\pi_1$ | $\pi_2$ | $u_k(h_k, a_k, y_k)$ |
| 1     | 0           | 0        |          |      |            |         | 1       |
| 2     | 0           | 1        | 0        | 0    | $r_0$     | $s_0$   | 0       |
| 2     | 0           | 1        | 1        | 0    |           | $s_0$   | 1       |
| 1     | 1           | 0        |          |      |            |         | 1       |
| 2     | 1           | 1        | 0        | 0    | $r_1$     | $s_1$   | 0       |
| 2     | 1           | 1        | 1        | 0    |           | $s_1$   | 1       |
| 2     | 1           | 1        | 1        | 1    | $s_1$     |         | 0       |

Note: $r_0, r_1 \in R : R = \{0, 5, 10, ..., 100\}/100$ and $s_0, s_1 \in S : S = \{5, 10, 20, 40, 60, 80, 90, 95\}/100$. 

24