A Functional Urn Model for CARA Designs

Giacomo Aletti, Andrea Ghiglietti, and William F. Rosenberger

Abstract We present a general class of covariate-adjusted response-adaptive (CARA) designs introduced in [1], which is based on a new functional urn model. We show strong consistency concerning the allocation probability and the proportion of subjects assigned to the treatment groups, in the whole study and for each covariate profile, allowing the distribution of the responses conditioned on covariates to be estimated nonparametrically. We also establish joint central limit theorems for these quantities and the joint sufficient statistics, which allow construction of inference procedures.

Abstract *In questo lavoro presentiamo una classe generale di disegni covariate-adjusted adattivi alla risposta (CARA) introdotti in [1], che è basato su un nuovo modello d'urna funzionale. Inoltre, dimostriamo la forte consistenza della probabilità di allocazione e della proporzione di soggetti assegnati ai gruppi dei trattamenti, nell'intero studio e per ciascun valore delle covariate, permettendo alla distributione delle risposte condizionate alle covariate di essere stimata in maniera non parametrica. Infine, abbiamo stabilito alcuni teoremi centrale del limite congiunti di queste quantità e delle statistiche sufficienti, che permoettono la costruzione di procedure inferenziali.*

Key words: asymptotics, clinical trials, covariate-adjusted response-adaptive designs, randomization

1 Introduction

In CARA designs the patients in the trial are randomly assigned to $d \geq 2$ treatment groups with an allocation probability that depends on the current patient covariate profile and on the previous patients' covariates, allocations and responses (e.g. see [3]). In this framework, it is desirable that the proportion of subjects of each covariate profile assigned to the treatments converges to a desired target, defined as a function of the response distribution conditionally on the covariates.

Giacomo Aletti
ADAMSS Center and Università degli Studi di Milano e-mail: giacomo.aletti@unimi.it,

Andrea Ghiglietti
Università degli Studi di Milano e-mail: andrea.ghiglietti@unimi.it

William F. Rosenberger
George Mason University, Fairfax, VA, USA e-mail: wrosenbe@gmu.edu

Ideally, the analysis of the ethical and inferential properties of the experimental designs should be based on theoretical results concerning the asymptotic behavior of the allocation proportion and adaptive estimators, and none of the previous work on CARA designs is able to provide such results. In fact, since the allocation and the estimation process depend on both the responses and the covariates, CARA designs are very complex to be formulated in a rigorous mathematical setting. Two papers, in particular, formalize CARA in a rigorous mathematical framework. The first of these is the groundbreaking paper of [4], in which consistency and second-order asymptotic results concerning both adaptive estimators and allocation proportions have been proved for a very wide class of CARA designs. In the second [2], compound optimal design theory was used to find target allocations of interest, and these target allocations are attained using an accelerated biased coin design.

Here we present a class of CARA designs introduced in [1], in which the allocation probability may depend on nonparametric estimates of the response distribution, and the patients' covariate profiles are not identically distributed.

2 The model

For any $n \geq 0$, let $\mathbf{Y}_n = (Y_n^1, .., Y_n^d)^\top$ be a vector of functions, with $Y_n^j : \tau \mapsto (0,1)$, where τ is the covariate space. For any $t \in \tau$, $\mathbf{Y}_n(t)$ represents an urn containing $Y_n^j(t)$ balls of color $j \in \{1, .., d\}$ and $\mathbf{Z}_n(t) = \mathbf{Y}_n / \sum_{j=1}^d Y_n^j$ indicates the proportion of the colors.

When subject n enters the trial, his covariate profile T_n is observed. Then, a ball is sampled at random from the urn identified by T_n (i.e. with proportions $\mathbf{Z}_{n-1}(T_n)$), its color is observed and represented by $\bar{\mathbf{X}}_n$: $\bar{X}_n^j = 1$ when the color is $j \in \{1, .., d\}$, $\bar{X}_n^j = 0$ otherwise. Then, subject n receives the treatment associated to the sampled color and a response $\bar{\xi}_n$ is collected. The functional urn is then updated as: $\mathbf{Y}_n = \mathbf{Y}_{n-1} + D_n \mathbf{X}_n$, where $\mathbf{X}_n$ and D_n are appropriately defined. Specifically, the weighting function $\mathbf{X}_n : \tau \mapsto [0,1]^d$ should be such that, for any $t \in \tau$ and $j \in \{1, .., d\}$, $\sum_{j=1}^d X_n^j(t) = 1$ and $\mathbf{E}[X_n^j(t) | \mathscr{F}_{n-1}, T_n] = Z_n^j(t)$, where $\mathscr{F}_{n-1}$ is the σ-algebra of the information related with the first $(n-1)$ patients. This is straightforward for $t = T_n$ by setting $X_n^j(T_n) = \bar{X}_n^j$, since $\bar{X}_n^j$ is conditionally on $\mathscr{F}_{n-1}$ and T_n Bernoulli distributed with parameter $Z_n^j(T_n)$. Then, we define a family of Bernoulli random variables $\{\check{X}_n^j(t); t \in \tau\}$ with parameters $\{Z_n^j(t); t \in \tau\}$, representing the color that would be sampled in the trial if the covariate profile of subject n was equal to any $t \in \tau$. Finally, we use the quantile function that links this family to compute $\mathbf{X}_n(t)$ for all $t \in \tau$ as $\mathbf{X}_n := \mathbf{E}[\check{\mathbf{X}}_n | \mathscr{F}_{n-1}, T_n, \bar{\mathbf{X}}_n]$. Analogously, we can define the replacement functional matrix $D_n : \tau \mapsto [0,1]^{d \times d}$ as $D_n := \mathbf{E}[\check{D}_n | T_n, \bar{\mathbf{X}}_n, \bar{\xi}_n]$, where $\check{D}_n(t)$ is a function of a random variable having the same distribution of the response observed from a subject with covariate profile t, i.e. the response that would be observed in the trial if the covariate profile of subject n was equal to any $t \in \tau$. Naturally, $D_n(T_n) = \check{D}_n(T_n)$. Since the quantile functions of the response distributions are typically unknown, D_n is computed by using the corresponding (parametric or nonparametric) estimators obtained with the information in $\mathscr{F}_{n-1}$.

The key feature of the design is that quantile functions are used to update *all* urns, not just the urn for which $T_n = t$. In theory there could be an uncountably infinite number of urns, with only a finite subset of them used for patient allocation. However, in clinical practice, mathematically "continuous" covariates are really not continuous; for instance, cholesterol is represented by integer values, likely in some range, that would, for all intents and purposes, make it a finite discrete covariate. However, the procedure is well-defined for uncountably infinite urns, and first order asymptotic properties can be obtained, although some of

the covariate-specific metrics do not make sense in that context. When we move to second-order asymptotics, we partition τ into K strata, which could be intervals of a continuous set.

3 Consistency Results

We now present some consistency results for (i) the probability of allocation of the subjects for each covariate profile ($\mathbf{Z}_n(t)$), (ii) the proportion of subjects associated to each covariate profile assigned to the treatments ($\mathbf{N}_{t,n}/\sum_{j=1}^d N_{t,n}^j$, where $\mathbf{N}_{t,n} := \sum_{i=1}^n \bar{\mathbf{X}}_i \mathbf{1}_{\{T_i=t\}}$), (iii) the proportion of subjects assigned to the treatments ($\mathbf{N}_n/n$, where $\mathbf{N}_n := \sum_{i=1}^n \bar{\mathbf{X}}_i$). Consider the following assumptions:

(A1) for any $t \in \tau$ and $n \geq 1$, $D_n^\top \mathbf{1} = \mathbf{1}$ (*constant balance*);
(A2) denoting by $H(t) := \mathbf{E}[\breve{D}_1(t)]$ the average replacement when the covariate profile is t, we assume that $H(t)$ is irreducible, diagonalizable and there exists $\alpha > 0$ such that $\mathbf{E}[|\mathbf{E}[D_n(t)|\mathscr{F}_{n-1}, T_n, \bar{\mathbf{X}}_n] - H(t)|\mathscr{F}_{n-1}] = O(n^{-\alpha})$.

Denote by $\mathbf{v}(t)$ the right eigenvector of $H(t)$ associated to $\lambda = 1$, with $\sum_{j=1}^d v^j(t) = 1$, and let μ_{n-1} be the probability distribution of T_n conditioned on $\mathscr{F}_{n-1}$. Then,

(a) for any probability measure ν on τ, we have $\int_\tau \|\mathbf{Z}_n(t) - \mathbf{v}(t)\| \nu(dt) \overset{a.s.}{\to} 0$;
(b) if $\sum_{i=1}^n \mu_{i-1}(\{t\}) \overset{a.s.}{\to} \infty$, we have $\|\mathbf{N}_{t,n}/\sum_{j=1}^d N_{t,n}^j - \mathbf{v}(t)\| \overset{a.s.}{\to} 0$;
(c) if $\int_\tau |\mu_n(dt) - \mu(dt)| \overset{a.s.}{\to} 0$, we have $\|\mathbf{N}_n/n - \int_\tau \mathbf{v}(t)\mu(dt)\| \overset{a.s.}{\to} 0$.

The convergence results consider a general covariate space τ. In order to show second-order properties, we now partition τ into K finite elements, which could, for instance, be K intervals of a continuous covariate space. This partitioning induces K urns used to allocate subjects with covariate profiles in the set $\{1, ..., K\}$. In clinical trials practice, K must be considerably smaller than the total sample size.

4 Central Limit Theorems

We now present further assumptions that are required for establishing the second-order asymptotic properties.

- **Finite partition of the covariate space.** We assume that the covariate space τ is composed by a finite number $K \in \mathbb{N}$ of distinct elements. When τ contains infinite elements, we can take a partition of τ, i.e. $\{\tau_1, .., \tau_K\}$ such that $\cup_k \tau_k = \tau$ and $\tau_{k_1} \cap \tau_{k_2} = \emptyset$ for $k_1 \neq k_2$, and consider these sets to be the elements of τ, i.e. $\tau := \{\tau_1, .., \tau_K\}$. To facilitate the notation, without loss of generality in the sequel we redefine $\tau = \{1, .., K\}$ and $\mu_{n-1}(t) = \mu_{n-1}(\{t\}) = \mathbb{P}(T_n = t|\mathscr{F}_{n-1})$ for any $t \in \tau$.
- **Conditional response distributions.** The analog of the null hypothesis in classical inferential statistics is given here by assuming that the conditional response distributions $\pi_t^1, .., \pi_t^d$ are known for any $t \in \tau$. As a direct consequence, we have that $D_n = D_n^*$ and $H_n = H$ with probability one for any $n \geq 1$.
- **Eigenvalues of the limiting generating matrix.** Denoting $\lambda_H^*(t)$ the eigenvalue of $Sp(H(t)) \setminus \{1\}$ with largest real part, assume that $\max_{t\in\tau} \mathscr{R}e(\lambda_H^*(t)) < 1/2$.
- **Dynamics of adaptive estimators.**

– *(Covariate-stratification approach)* For some $t \in \tau$ and $j \in \{1,..,d\}$, consider that there are features of interest θ_t^j related with the distribution π_t^j of the responses to treatment j conditionally on the covariate profile t. Then, we assume that the corresponding adaptive estimator $\hat{\theta}_{t,n}^j$ is strongly consistent and its dynamics can be expressed as follows: there exists $n_0 \geq 1$ such that for any $n \geq n_0$

$$\hat{\theta}_{t,n}^j - \hat{\theta}_{t,n-1}^j = -\frac{\bar{X}_n^j \mathbb{1}_{\{T_n=t\}}}{N_{t,n}^j}(f_{t,j}(\hat{\theta}_{t,n-1}^j) - \Delta\mathbf{M}_{t,j,n} - \mathbf{R}_{t,j,n}), \tag{1}$$

where

(i) $f_{t,j}$ is a Lipschitz continuous function such that $f_{t,j}(\theta_t^j) = 0$;

(ii) $\Delta\mathbf{M}_{t,j,n} \in \mathscr{F}_n$ is a martingale increment such that $\mathbb{E}[\Delta\mathbf{M}_{t,j,n}|\mathscr{F}_{n-1}, T_n, \bar{X}_n^j] = 0$, and it converges stably to $\Delta\mathbf{M}_{t,j}$ with kernel K independent of $\mathscr{F}_{n-1}$: $\mathscr{L}(\Delta\mathbf{M}_{t,j,n}|\mathscr{F}_{n-1}, T_n = t, \bar{X}_n^j = 1) \xrightarrow{a.s.} K(t,j)$;

(iii) $\mathbf{R}_{t,j,n} \in \mathscr{F}_n$ is such that $n\mathbb{E}[\|\mathbf{R}_{t,j,n}\|^2] \to 0$.

Moreover, let $f_{t,j}$ be differentiable at θ_t^j, denote by $\lambda^*_{\theta_t^j}$ the eigenvalue of $Sp(\mathscr{D}f_{t,j}(\theta_t^j))$ with largest real part and assume that $\min_{t\in\tau}\mathscr{R}e(\lambda^*_{\theta_t^j}) > 1/2$. We also assume that for some $\delta > 0$,

$$\sup_{n\geq 1}\mathbb{E}\left[\left\|\Delta\mathbf{M}_{t,j,n}\right\|^{2+\delta} | \mathscr{F}_{n-1}\right] < +\infty \, a.s., \tag{2}$$

and

$$\mathbb{E}\left[\Delta\mathbf{M}_{t,j,n}(\Delta\mathbf{M}_{t,j,n})^\top | \mathscr{F}_{n-1}\right] \underset{n\to+\infty}{\xrightarrow{a.s.}} \Gamma_{t,j}, \tag{3}$$

where $\Gamma_{t,j}$ is a symmetric positive matrix.

– *(Covariate-adjusted approach)* For some $j \in \{1,..,d\}$, consider that there are features of interest β^j related with the entire family of distributions $\{\pi_t^j; t \in \tau\}$ of the responses to treatment j conditionally on the covariates. Then, we assume that the corresponding adaptive estimator $\hat{\beta}_n^j$ is strongly consistent and its dynamics can be expressed as follows:

$$\hat{\beta}_n^j - \hat{\beta}_{n-1}^j = -\frac{\bar{X}_n^j}{N_n^j}(f_j(\hat{\beta}_{n-1}^j) - \Delta\mathbf{M}_{j,n} - \mathbf{R}_{j,n}), \tag{4}$$

where the quantities in (4) fulfill the same conditions presented above for the dynamics (1).

We first provide the convergence rate and the joint asymptotic distribution concerning the quantities of interest in the design in the framework of covariate-stratification response-adaptive designs. This result is established in the following central limit theorem. We introduce the variables independent of $\sigma(\mathscr{F}_n; n \geq 1)$: $T \in \tau$ with distribution $\mu(t)$, $\bar{\mathbf{X}} \in \{0,1\}^d \in \mathscr{S}$ such that $\mathbb{P}(\bar{X}^j = 1|T) = v^j(T)$, $D := \mathbb{E}[\breve{D}|T, \bar{\mathbf{X}}, \bar{\xi}]$, where the distribution of $\bar{\xi}$ conditioned on $\{T = t\}$ and $\{\bar{X}^j = 1\}$ is π_t^j.

Theorem 4.1. *Define* $\mathbf{W}_n := (\mathbf{Z}_n(t), \mathbf{N}_{t,n}/w(\mathbf{N}_{t,n}), \hat{\theta}_{t,n}, t \in \tau)^\top$, $\mathbf{W} := (\mathbf{v}(t), \mathbf{v}(t), \theta_t, t \in \tau)^\top$. *Then,*

$$\mu_n(t) \xrightarrow{a.s.} \mu(t) = f_{\mu,t}(\mathbf{v}(t), \theta_t), \qquad \mathbf{W}_n \xrightarrow{a.s.} \mathbf{W}, \tag{5}$$

$$\sqrt{n}(\mathbf{W}_n - \mathbf{W}) \xrightarrow{\mathscr{L}} \mathscr{N}(\mathbf{0}, \Sigma), \qquad \Sigma := \int_0^\infty e^{u(\frac{\mathbf{I}}{2}-A)}\Gamma e^{u(\frac{\mathbf{I}}{2}-A^\top)}du, \tag{6}$$

where

$$A := \begin{pmatrix} A_{ZZ} & 0 & 0 \\ -I & I & 0 \\ 0 & 0 & A_{\theta\theta} \end{pmatrix}, \qquad \Gamma := \begin{pmatrix} \Gamma_{ZZ} & \Gamma_{ZN} & \Gamma_{Z\theta} \\ \Gamma_{ZN}^\top & \Gamma_{NN} & 0 \\ \Gamma_{Z\theta}^\top & 0 & \Gamma_{\theta\theta} \end{pmatrix},$$

and A_{ZZ}, $A_{\theta\theta}$, Γ_{NN}, $\Gamma_{\theta\theta}$ are block-diagonal matrices whose t^{th} block is

(i) $A_{ZZ}^{tt} = (I - H(t) + \mathbf{v}(t)\mathbf{1}^\top)$;
(ii) $A_{\theta\theta}^{tt}$ *is a block-diagonal matrices whose* j^{th} *block is* $[A_{\theta\theta}^{tt}]^{jj} := \mathscr{D} f_{t,j}(\theta_t^j)$;
(iii) $\Gamma_{NN}^{tt} := \mu^{-1}(t)(diag(\mathbf{v}(t)) - \mathbf{v}(t)\mathbf{v}^\top(t))$;
(iv) $\Gamma_{\theta\theta}^{tt}$ *is a block-diagonal matrices whose* j^{th} *block is*
$[\Gamma_{\theta\theta}^{tt}]^{jj} := (v^j(t)\mu(t))^{-1}\mathbb{E}[\Delta\mathbf{M}_{t,j}(\Delta\mathbf{M}_{t,j})^\top | T = t, \bar{X}^j = 1]$;

and Γ_{ZZ}, Γ_{ZN}, $\Gamma_{Z\theta}$ are matrices defined as follows: for any $t_1, t_2 \in \tau$

(v) $\Gamma_{ZZ}^{t_1 t_2} := \mathbb{E}[D(t_1)\mathbf{g}(t_1, T, \bar{\mathbf{X}})\mathbf{g}^\top(t_2, T, \bar{\mathbf{X}})D^\top(t_2)] - \mathbf{v}(t_1)\mathbf{v}^\top(t_2)$;
(vi) $\Gamma_{ZN}^{t_1 t_2} := H(t_1)G(t_1, t_2)diag(\mathbf{v}(t_2)) - \mathbf{v}(t_1)\mathbf{v}^\top(t_2)$;
(vii) $[\Gamma_{Z\theta}^{t_1 t_2}]^j := \mathbb{E}[D(t_1)\mathbf{g}(t_1, t_2, \mathbf{e}_j)\Delta\mathbf{M}_{t_2,j}^\top | T = t_2, \bar{X}^j = 1]$;

where $\mathbf{g}$ *is a d-multivariate function with values in $\mathscr{S}$ and $G(t_1, t_2)$ is a matrix with columns $\{\mathbf{g}(t_1, t_2, \mathbf{e}_j); j \in \{1, .., d\}\}$.*

We now provide the convergence rate and the joint asymptotic distribution of the quantities interest in the design in the framework of covariate-adjusted response-adaptive designs. This result is established in the following central limit theorem.

Theorem 4.2. *Define* $\mathbf{W}_n := (\mathbf{Z}_n(t), t \in \tau, \mathbf{N}_n/n, \hat{\beta}_n)^\top$, $\mathbf{W} := (\mathbf{v}(t), t \in \tau, \mathbf{x}_0, \beta)^\top$. *Then,*

$$\mu_n(t) \xrightarrow{a.s.} \mu(t) = f_{\mu,t}(\mathbf{x}_0, \beta), \qquad \mathbf{W}_n \xrightarrow{a.s.} \mathbf{W}, \tag{7}$$

$$\sqrt{n}(\mathbf{W}_n - \mathbf{W}) \xrightarrow{\mathscr{L}} \mathscr{N}(\mathbf{0}, \Sigma), \qquad \Sigma := \int_0^\infty e^{u(\frac{\mathbf{I}}{2} - A)} \Gamma e^{u(\frac{\mathbf{I}}{2} - A^\top)} du, \tag{8}$$

and

$$A := \begin{pmatrix} A_{ZZ} & 0 & 0 \\ A_{NZ} & A_{NN} & A_{N\beta} \\ 0 & 0 & A_{\beta\beta} \end{pmatrix}, \qquad \Gamma := \begin{pmatrix} \Gamma_{ZZ} & \Gamma_{ZN} & \Gamma_{Z\beta} \\ \Gamma_{ZN}^\top & \Gamma_{NN} & 0 \\ \Gamma_{Z\beta}^\top & 0 & \Gamma_{\beta\beta} \end{pmatrix},$$

where again and A_{ZZ}, $A_{\beta\beta}$, $\Gamma_{\beta\beta}$ are block-diagonal matrices whose t^{th} or j^{th} block is

(i) $A_{ZZ}^{tt} = (I - H(t) + \mathbf{v}(t)\mathbf{1}^\top)$;
(ii) $A_{\beta\beta}^{jj} = \mathscr{D} f_j(\beta^j)$;
(iii) $\Gamma_{\beta\beta}^{jj} := (\mathbb{E}[v^j(T)])^{-1}\mathbb{E}[\Delta\mathbf{M}_j(\Delta\mathbf{M}_j)^\top | \bar{X}^j = 1]$;

and

(iv) $A_{NN} := I - \sum_{s=1}^K \mathbf{v}(s)\mathscr{D}_N f_{\mu,s}(\mathbf{x}_0, \beta)^\top$;
(v) $A_{N\beta} := -\sum_{s=1}^K \mathbf{v}(s)\mathscr{D}_\beta f_{\mu,s}(\mathbf{x}_0, \beta)^\top$;
(vi) $\Gamma_{NN} := diag(\mathbb{E}[\mathbf{v}(T)]) - \mathbb{E}[\mathbf{v}(T)]\mathbb{E}[\mathbf{v}^\top(T)]$;

and A_{NZ}, Γ_{ZZ}, Γ_{ZN}, $\Gamma_{Z\beta}$ are matrices defined as follows: for any $t_1, t_2 \in \tau$

(vii) $A_{NZ}^{t_2} := -\mu(t_2)I$;

(viii) $\Gamma_{ZZ}^{t_1 t_2} := \mathbb{E}[D(t_1)\mathbf{g}(t_1,T,\bar{\mathbf{X}})\mathbf{g}^\top(t_2,T,\bar{\mathbf{X}})D^\top(t_2)] - \mathbf{v}(t_1)\mathbf{v}^\top(t_2)$;
(ix) $\Gamma_{ZN}^{t_1} := H(t_1)\mathbb{E}[G(t_1,T)diag(\mathbf{v}(T))] - \mathbf{v}(t_1)\mathbb{E}[\mathbf{v}^\top(T)]$;
(x) $\Gamma_{Z\beta}^{t_1 j} := \mathbb{E}[D(t)\mathbf{g}(t_1,T,j)\Delta\mathbf{M}_j^\top | \bar{X}^j = 1]$.

where we recall that $\mathbf{g}$ *is a* d*-multivariate function with values in* $\mathscr{S}$ *and* $G(t_1,t_2)$ *is a matrix with columns* $\{\mathbf{g}(t_1,t_2,\mathbf{e}_j); j \in \{1,..,d\}\}$.

Remark 4.1. *We recall that Theorem 4.1 allows inferential procedures based on stratified estimators, while Theorem 4.2 allows inference on covariate-adjusted regression parameters representing the covariate-adjusted treatment effect.*

References

1. Aletti G., Ghiglietti A., Rosenberger W.F. (2018). Nonparametric covariate-adjusted response-adaptive design based on a functional urn model, *Ann. Statist.*, in press.
2. Baldi Antognini, A. and Zagoraiou, M. (2012). Multi-objective optimal designs in comparative clinical trials with covariates: the reinforced doubly adaptive biased coin design, *Ann. Statist.* **40** 1315–1345.
3. Hu F., Rosenberger W.F. (2006). *The Theory of Response-Adaptive Randomization in Clinical Trials*, John Wiley & Sons, New York.
4. Zhang L.-X., Hu F., Cheung S. H., Chan W. S. (2007). Asymptotic properties of covariate-adjusted response-adaptive designs, *Ann. Statist.* **35** 1166–1182.