The Annals of Statistics
2017, Vol. 45, No. 6, 2537–2564
https://doi.org/10.1214/16-AOS1534

TARGETED SEQUENTIAL DESIGN FOR TARGETED LEARNING INFERENCE OF THE OPTIMAL TREATMENT RULE AND ITS MEAN REWARD

BY ANTOINE CHAMBAZ[*,†,1,2], WENJING ZHENG[†,1] AND MARK J. VAN DER LAAN[†,3]

UPL, Université Paris Nanterre[*] and University of California, Berkeley[†]

This article studies the targeted sequential inference of an optimal treatment rule (TR) and its mean reward in the nonexceptional case, *that is*, assuming that there is no stratum of the baseline covariates where treatment is neither beneficial nor harmful, and under a companion margin assumption.

Our pivotal estimator, whose definition hinges on the targeted minimum loss estimation (TMLE) principle, actually infers the mean reward under the current estimate of the optimal TR. This data-adaptive statistical parameter is worthy of interest on its own. Our main result is a central limit theorem which enables the construction of confidence intervals on both mean rewards under the current estimate of the optimal TR and under the optimal TR itself. The asymptotic variance of the estimator takes the form of the variance of an efficient influence curve at a limiting distribution, allowing to discuss the efficiency of inference.

As a by product, we also derive confidence intervals on two cumulated pseudo-regrets, a key notion in the study of bandits problems.

A simulation study illustrates the procedure. One of the cornerstones of the theoretical study is a new maximal inequality for martingales with respect to the uniform entropy integral.

1. Introduction. This article contributes theoretically to the burgeoning field of precision medicine, whose general focus is on identifying which treatments and preventions will be effective for which patients based on genetic, environmental and lifestyle factors. It studies the targeted data-adaptive inference of an optimal treatment rule (TR) from data sampled based on a targeted sequential design. A TR is an individualized treatment strategy in which treatment assignment for a patient is based on her measured baseline covariates. Eventually, a reward is measured on the patient. Optimality is meant in terms of maximization of the mean reward. We also infer the mean reward under the optimal TR.

Received April 2016; revised September 2016.

[1]Supported by the French Agence Nationale de la Recherche (ANR), under Grant ANR-13-BS01-0005 (project SPADRO).

[2]This research has been conducted as part of the project Labex MME-DII (ANR11-LBX-0023-01).

[3]Supported by NIH Grant Number 2R01 A1074345-07.

MSC2010 subject classifications. Primary 62G05, 62G20; secondary 62K99.

Key words and phrases. Bandits, optimal treatment rule, precision medicine, pseudo-regret, targeted minimum loss estimation (TMLE).

We choose not to frame our statistical model into a causal one. To give an idea of how the targeted sequential design unfolds, suppose nonetheless (in this paragraph only) that there exists an infinite sequence of independent and identically distributed (i.i.d.) full data structures consisting each of a set of baseline covariates and a couple of potential rewards measured on a randomly sampled patient. The baseline covariates describe the corresponding patient and the rewards are the two potential outcomes for the patient corresponding with the two possible treatments. Only one reward can be observed, that corresponding with the assigned treatment. A TR maps deterministically the baseline covariates to one treatment; a *stochastic* TR does so randomly. The optimal TR is that TR which maps the baseline covariates to that treatment with the larger mean reward. The mean reward of the optimal TR is the average of the larger mean with respect to (wrt) the baseline covariates. Until a number of observations deemed sufficient to begin to learn from them is reached, treatment is assigned equiprobably regardless of the baseline covariates. Then, sequentially, all observations collected so far (each consisting of baseline covariates, a single treatment assignment and the reward resulting from it) are exploited to learn the optimal TR, which is approximated by a stochastic TR from which the next treatment assignment is drawn conditionally on the next observed baseline covariates. Our statistical model is a submodel of the above causal model derived from it under sequential missingness.

The targeted sequential elaboration of our design and its companion inference procedure are driven by two objectives. First, increasing the chance that each patient enrolled in the study be assigned that treatment which is more favorable to her according to data accrued so far. Second, increasing the robustness and efficiency of statistical inference through the construction of targeted, narrower confidence intervals. The latter objective is appealing to the investigators of the study, and the former to the patients enrolled in it and their doctors. Indeed, a known disadvantage of traditional randomized clinical trials is that randomization interferes with the doctor-patient relationship: clinicians must admit to each potential patient that it is not known which of the treatments would be best for her, thereby potentially eroding their relationship. From an ethical perspective, a second disadvantage is that the clinicians should believe that the treatments are equivalent wrt potential patient benefit, a situation many of them find uncomfortable [28]. These two disadvantages would be respectively considerably diminished and irrelevant in a trial based on our design, at the cost of a more complex implementation. In addition, one may expect a gain in compliance.

The authors of [4] present an excellent unified overview on the estimation of optimal TRs, with a special interest in dynamic rules (where treatment assignment consists in successive assignments at successive time points). The estimation of the optimal TR from i.i.d. observations has been studied extensively, with a recent interest in the use of machine learning algorithms to reach this goal [20, 24, 27, 34–37]. In contrast, we estimate the optimal TR (and its mean reward) based on sequentially sampled dependent observations by empirical risk minimization over

sample-size-dependent classes of candidate estimates with a complexity controlled in terms of uniform entropy integral.

The estimation of the mean reward under the optimal TR is more challenging than that of the optimal TR. In [36, 37], the theoretical risk bound evaluating the statistical performance of the estimator of the optimal TR can also be interpreted in terms of a measure of statistical performance of the resulting estimator of the mean reward under the optimal TR. However, it does not yield confidence intervals.

Constructing confidence intervals for the mean reward under the optimal TR is known to be more difficult when there exists a stratum of the baseline covariates where treatment is neither beneficial nor harmful [25]. In this so-called "exceptional" case, the definition of the optimal TR has to be disambiguated. Assuming nonexceptionality, confidence intervals are obtained in [34] for the mean reward under the (sub-)optimal TR defined as the optimal TR over a parametric class of candidate TRs, and in [18] for the actual mean reward under the optimal TR. In the more general case where exceptionality can occur, different approaches have been considered [3, 12, 17, 19]. Here, we focus on the nonexceptional case under a companion margin assumption [21].

Because we are committed to providing robust and (more) efficient inference, we rely on the targeted minimum loss estimation (TMLE) principle [31]. Succinctly, and focusing on targeted maximum likelihood estimation, the first instance of TMLE [32], the procedure consists in the following steps: (a) viewing the parameter of interest as a smooth functional Ψ evaluated at a law P_0, (b) computing a possibly highly data-adaptive initial estimator P_n^0 of P_0 (*e.g.*, a super learner), (c) defining a least favorable model through P_n^0 (*i.e.*, a parametric model through P_n^0 whose score spans the so called canonical gradient at P_n^0 of the derivative of Ψ), (d) maximizing the log-likelihood over this model to define an updated estimator P_n^* (possibly iteratively), (e) defining the TMLE as the plug-in estimator $\Psi(P_n^*)$ obtained by evaluating Ψ at the last update of the estimator of P_0. Targeted maximum likelihood estimation was naturally extended to targeted minimum loss based estimation by replacing the log-likelihood loss and least favorable model by any pair of loss and submodel whose generalized score spans the canonical gradient. The method has been applied and advanced across a large variety of data structures, models and target parameters. Since it involves substitution estimators, TMLE should typically yield better performances in small sample than one-step [13, 23] or estimating equations-based [30] estimators, which are also not as general. We can build upon previous studies on the construction and statistical analysis of targeted, covariate-adjusted, response-adaptive trials also based on TMLE [6, 7, 38]. One of the cornerstones of the theoretical study is a new maximal inequality for martingales wrt the uniform entropy integral, proved by decoupling [9], symmetrization and chaining, which allows us to control several empirical processes indexed by random functions.

Our pivotal TMLE estimator is actually constructed as an estimator of the mean reward under the current estimate of the optimal TR. Worthy of interest on its own,

this data-adaptive statistical parameter (or similar ones) has also been considered in [3, 16–19]. Our main result is a central limit theorem for our TMLE estimator. The asymptotic variance takes the form of the variance of an efficient influence curve at a limiting distribution, allowing to discuss the efficiency of inference.

We use our TMLE estimator to infer the mean rewards under the current estimate of the optimal TR and under the optimal TR itself. Moreover, we use it to infer two additional data-adaptive statistical parameters. The first one compares the sum of the rewards actually received during the course of the experiment with the sum of the means of the rewards we would have obtained if we had used from the start the current estimate of the optimal TR to assign treatment. The second one compares the sum of the rewards actually received during the course of the experiment with the sum of the counterfactual rewards we would have obtained if we had used from the start the current estimate of the optimal TR to assign treatment.

Both additional data-adaptive statistical parameters are "cumulated pseudo-regrets". We borrow this expression from the literature on bandits. Bandits have raised a considerable interest in the machine learning community as relevant models for interactive learning schemes or recommender systems. Many articles define efficient strategies to minimize the expected cumulated pseudo-regret (also known as the "cumulated regret"); see [2] for a survey. Sometimes, the objective is to identify the arm with the largest mean reward (the best arm) as fast and accurately as possible, regardless of the number of times a sub-optimal arm is played; see [11] for an in-depth analysis of the so-called fixed-confidence setting where one looks for a strategy guaranteeing that the probability of wrongly identifying the best arm at some stopping time is no more than a fixed maximal risk while minimizing the stopping time's expectation. Here, we derive confidence intervals on the cumulated pseudo-regrets as by products of the confidence intervals that we build for the mean rewards under the current estimate of the optimal TR and under the optimal TR itself. Thus, the most relevant comparison is with the so called "contextual bandit problems"; see [15], Chapter 4, for an excellent overview.

Organization. Section 2 presents our targeted, data-adaptive sampling scheme and our pivotal estimator. Section 3 studies the convergence of the sampling scheme, *that is*, how the sequences of stochastic and TRs converge, assuming that a function of the conditional mean of the reward given treatment and baseline covariate is consistently estimated. Section 4 is devoted to the presentation of our main result, a central limit theorem for our pivotal estimator, to the comment of its assumptions and to an example. Section 5 builds upon the previous section to build confidence intervals for the mean rewards under the current estimate of the optimal TR and under the optimal TR itself, as well as confidence intervals for the two cumulated pseudo-regrets evoked in the Introduction. Section 6 presents the results of a simulation study. Section 7 closes the article with a brief discussion. All proofs are given in [8], Section A. Technical lemmas are gathered in [8], Sections B and C.

2. Targeting the optimal treatment rule and its mean reward.

2.1. *Statistical setting*. At sample size n, we will have observed the ordered vector $\mathbf{O}_n \equiv (O_1, \ldots, O_n)$, with convention $O_0 \equiv \varnothing$. For every $1 \leq i \leq n$, the data structure O_i writes as $O_i \equiv (W_i, A_i, Y_i)$. Here, $W_i \in \mathcal{W}$ consists of the baseline covariates (some of which may be continuous) of the ith patient, $A_i \in \mathcal{A} \equiv \{0, 1\}$ is the binary treatment of interest assigned to her, and $Y_i \in \mathcal{Y}$ is her primary outcome of interest. We interpret Y as a *reward*: the larger is Y, the better. We assume that the space $\mathcal{O} \equiv \mathcal{W} \times \mathcal{A} \times \mathcal{Y}$ is bounded. Without loss of generality, we may then assume that $\mathcal{Y} \equiv (0, 1)$, *that is*, that the rewards are between and bounded away from 0 and 1. Interestingly, the content of this article would still hold up to minor modifications if we assumed instead $\mathcal{Y} \equiv \{0, 1\}$.

Let μ_W be a measure on $\mathcal{W}$ equipped with a σ-field, $\mu_A = \mathrm{Dirac}(0) + \mathrm{Dirac}(1)$ be a measure on $\mathcal{A}$ equipped with its σ-field, and μ_Y be the Lebesgue measure on $\mathcal{Y}$ equipped with the Borel σ-field. Define $\mu \equiv \mu_W \otimes \mu_A \otimes \mu_Y$, a measure on $\mathcal{O}$ equipped with the product of the above σ-fields. The unknown, true likelihood of $\mathbf{O}_n$ wrt $\mu^{\otimes n}$ is given by the following factorization of the density of $\mathbf{O}_n$ wrt $\mu^{\otimes n}$:

$$\mathscr{L}_{Q_0, \mathbf{g}_n}(\mathbf{O}_n) \equiv \prod_{i=1}^{n} Q_{W,0}(W_i) \times g_i(A_i|W_i) \times Q_{Y,0}(Y_i|A_i, W_i), \tag{2.1}$$

where (i) $w \mapsto Q_{W,0}(w)$ is the density wrt μ_W of a true, unknown law on $\mathcal{W}$ (that we assume being dominated by μ_W), (ii) $\{y \mapsto Q_{Y,0}(y|a, w) : (a, w) \in \mathcal{A} \times \mathcal{W}\}$ is the collection of the conditional densities $y \mapsto Q_{Y,0}(y|a, w)$ wrt μ_Y of true, unknown laws on $\mathcal{Y}$ indexed by (a, w) (that we assume being all dominated by μ_Y), (iii) $g_i(1|W_i)$ is the known conditional probability that $A_i = 1$ given W_i and (iv) $\mathbf{g}_n \equiv (g_1, \ldots, g_n)$, the ordered vector of the n first *stochastic rules*. One reads in (2.1) (i) that $W_1, \ldots, W_n$ are independently sampled from $Q_{W,0}\, d\mu_W$, (ii) that $Y_1, \ldots, Y_n$ are conditionally sampled from $Q_{Y,0}(\cdot|A_1, W_1)\, d\mu_Y, \ldots, Q_{Y,0}(\cdot|A_n, W_n)\, d\mu_Y$, respectively and (iii) that each A_i is drawn conditionally on W_i from the Bernoulli distribution with known parameter $g_i(1|W_i)$.

In (2.1) and the subsequent text, subscript "0" stands for "truth" and refers to true, unknown features of the distribution of the data. This notational convention will prevail throughout the article.

We introduce the semiparametric collection $\mathcal{Q}$ of all elements of the form:

$$Q = \big(Q_W\, d\mu_W, Q_Y(\cdot|a, w), (a, w) \in \mathcal{A} \times \mathcal{W}\big), \quad \text{or}$$

$$Q = \left(Q_W \sum_{k=1}^{K} \mathrm{Dirac}(w_k), Q_Y(\cdot|a, w), (a, w) \in \mathcal{A} \times \mathcal{W}\right)$$

with $\{w_1, \ldots, w_K\} \subset \mathcal{W}$. Here, Q_W is a density wrt either μ_W or a discrete measure $\sum_{k=1}^{K} \mathrm{Dirac}(w_k)$ (thus, we can take the empirical measure of W as first

component of Q). Each $Q_Y(\cdot|a,w)$ is a density wrt μ_Y. In particular, $Q_0 \equiv (Q_{W,0}\,d\mu_W, Q_{Y,0}(\cdot|a,w), (a,w) \in \mathcal{A}\times\mathcal{W}) \in \mathcal{Q}$. In light of (2.1) define, for every $Q \in \mathcal{Q}$, $\mathscr{L}_{Q,\mathbf{g}_n}(\mathbf{O}_n) \equiv \prod_{i=1}^n Q_W(W_i) \times g_i(A_i|W_i) \times Q_Y(Y_i|A_i,W_i)$. The set $\{\mathscr{L}_{Q,\mathbf{g}_n} : Q \in \mathcal{Q}\}$ is a semiparametric model for the likelihood of $\mathbf{O}_n$. It contains the true, unknown likelihood $\mathscr{L}_{Q_0,\mathbf{g}_n}$.

Fix arbitrarily $Q \in \mathcal{Q}$. The conditional expectation of Y given (A,W) under Q is denoted $Q_Y(A,W) \equiv \int y Q_Y(y|A,W)\,d\mu_Y(y)$. To alleviate notation, we introduce the so-called "blip function" q_Y characterized by $q_Y(W) = Q_Y(1,W) - Q_Y(0,W)$. If $q_Y(W) \geq 0$ [resp., $q_Y(W) < 0$], then assigning treatment $A=1$ (resp., $A=0$) guarantees that the patient receives the superior treatment in the sense that her mean reward is larger in this arm than in the other one. If $q_Y(W)=0$, then the mean rewards are equal. This characterizes an optimal stochastic rule $r(Q_Y)$ given by

$$r(Q_Y)(W) \equiv \mathbf{1}\{q_Y(W) \geq 0\}. \tag{2.2}$$

It is degenerate because, given W, the assignment is deterministic. Such degenerate stochastic rules are usually referred to as *treatment rules* in the causal inference literature (already used in Section 1, this expression abbreviates to TRs). When $Q = Q_0$, we denote $Q_Y \equiv Q_{Y,0}$, $q_Y \equiv q_{Y,0}$ and $r(Q_Y) \equiv r_0$.

The parameter of interest is the mean reward under the optimal TR,

$$\psi_0 \equiv E_{Q_0}\big(Q_{Y,0}\big(r_0(W),W\big)\big) = \int Q_{Y,0}\big(r_0(w),w\big)Q_{W,0}(w)\,d\mu_W(w). \tag{2.3}$$

Let $\mathcal{G}$ be the semiparametric collection of all stochastic TRs g, which satisfy $g(1|W) = 1 - g(0|W) \in (0,1)$. From now on, for each $(Q,g) \in \mathcal{Q}\times\mathcal{G}$, we denote $P_{Q,g}$ the distribution of $O=(W,A,Y)$ obtained by drawing W from Q_W, then A from the Bernoulli distribution with parameter $g(1|W)$, then Y from the conditional distribution $Q_Y(\cdot|A,W)\,d\mu_Y$. Let $\mathcal{M} \equiv \{P_{Q,g} : Q \in \mathcal{Q}, g \in \mathcal{G}\}$. We actually see ψ_0 as the value at any $P_{Q_0,g}$ ($g \in \mathcal{G}$) of the mapping $\Psi : \mathcal{M} \to [0,1]$ characterized by

$$\Psi(P_{Q,g}) \equiv E_Q\big(Q_Y\big(r(Q_Y)(W),W\big)\big). \tag{2.4}$$

Obviously, the parameter $\Psi(P_{Q,g})$ does not depend on g. It depends linearly on the marginal distribution $Q_W\,d\mu_W$, but in a more subtle way on the conditional expectation Q_Y.

We have not specified yet what is precisely $\mathbf{g}_n \equiv (g_1,\ldots,g_n)$. Our targeted sampling scheme "targets" the optimal TR r_0 and ψ_0. By targeting r_0, we mean estimating $Q_{Y,0}$ based on past observations, and relying on the resulting estimator to collect the next block of data, as seen in (2.1), and to estimate ψ_0. Targeting ψ_0 refers to our efforts to build an estimator of ψ_0 which allows the construction of valid, narrow confidence intervals.

2.2. *Targeted sequential sampling and inference.* Let $\{t_n\}_{n\geq 1}$, $\{\xi_n\}_{n\geq 1}$ be two user-supplied, nonincreasing sequences with $t_1 \leq 1/2$, $\lim_n t_n \equiv t_\infty > 0$ and $\lim_n \xi_n \equiv \xi_\infty > 0$. For every $n \geq 1$, introduce the function G_n characterized over $[-1, 1]$ by

$$\begin{aligned} G_n(x) &= t_n \mathbf{1}\{x \leq -\xi_n\} + (1 - t_n)\mathbf{1}\{x \geq \xi_n\} \\ &\quad + \left(-\frac{1/2 - t_n}{2\xi_n^3}x^3 + \frac{1/2 - t_n}{2\xi_n/3}x + \frac{1}{2}\right)\mathbf{1}\{-\xi_n \leq x \leq \xi_n\}. \end{aligned} \tag{2.5}$$

For convenience, we also introduce $G_\infty \equiv G_{n_1}$ where $n_1 \geq 1$ is chosen large enough so that $t_{n_1} = t_\infty$ and $\xi_{n_1} = \xi_\infty$. Function G_n is nondecreasing and c_n-Lipschitz with

$$c_n \equiv \frac{1/2 - t_n}{2\xi_n/3} \leq \frac{1/2 - t_\infty}{2\xi_\infty/3} \equiv c_\infty.$$

A smooth approximation to $x \mapsto \mathbf{1}\{x \geq 0\}$ with values bounded away from 0 and 1, G_n will be used to derive a stochastic TR from an estimated blip function; see (2.11). This derivation mimicks the definition of the optimal TR as the indicator of the true blip function being nonnegative. This particular choice of G_n is one among many. Any other nondecreasing function $\tilde{G}_n$ such that $\tilde{G}_n(x) = t_n$ for $x \leq -\xi_n$, $\tilde{G}_n(x) = 1 - t_n$ for $x \geq \xi_n$, and $\tilde{G}_n$ κ_n-Lipschitz with κ_n upper-bounded by a finite κ_∞ could be chosen as well.

Loss functions and working models. Let $g^{\mathrm{b}} \in \mathcal{G}$ be the balanced stochastic TR wherein each arm is assigned with probability $1/2$ regardless of baseline covariates. Let $g^{\mathrm{ref}} \in \mathcal{G}$ be a stochastic TR, bounded away from 0 and 1 by choice, that serves as a reference. In addition, let L be a loss function for $Q_{Y,0}$ and $\mathcal{Q}_{1,n}$ be a working model:

$$\mathcal{Q}_{1,n} \equiv \{Q_{Y,\beta} : \beta \in B_n\} \subset \mathcal{Q}_Y \equiv \{Q_Y : Q \in \mathcal{Q}\}$$

consisting of functions $Q_{Y,\beta}$ mapping $\mathcal{A} \times \mathcal{W}$ to $[0, 1]$ [in the above display, Q_Y denotes the conditional expectation of Y given (A, W) under $Q \in \mathcal{Q}$]. One choice of L is the quasi negative-log-likelihood loss function L^{kl}. For any $Q_Y \in \mathcal{Q}_Y$ bounded away from 0 and 1, $L^{\mathrm{kl}}(Q_Y)$ satisfies

$$-L^{\mathrm{kl}}(Q_Y)(O) \equiv Y \log\big(Q_Y(A, W)\big) + (1 - Y)\log\big(1 - Q_Y(A, W)\big). \tag{2.6}$$

Another interesting loss function L for $Q_{Y,0}$ is the least-square loss function L^{ls}. It is characterized at any $Q_Y \in \mathcal{Q}_Y$ by

$$L^{\mathrm{ls}}(Q_Y)(O) \equiv \big(Y - Q_Y(A, W)\big)^2. \tag{2.7}$$

The loss function and working models will be used to estimate $Q_{Y,0}$.

Completing the description of the sampling scheme. We initialize the sampling scheme by setting $g_1 \equiv g^{\mathrm{b}}$. Consider $1 < i \leq n$. Since

$$Q_{Y,0} = \arg\min_{Q_Y \in \mathcal{Q}_Y} E_{Q_0,g}\big(L(Q_Y)(O)\big),$$

we naturally define

$$\beta_i \in \arg\min_{\beta \in B_i} \frac{1}{i-1} \sum_{j=1}^{i-1} L(Q_{Y,\beta})(O_j) \frac{g^{\mathrm{ref}}(A_j|W_j)}{g_j(A_j|W_j)} \tag{2.8}$$

and use Q_{Y,β_i} as an estimator of $Q_{Y,0}$ based on $\mathbf{O}_{i-1}$. It gives rise to q_{Y,β_i} and r_i such that

$$q_{Y,\beta_i}(W) \equiv Q_{Y,\beta_i}(1,W) - Q_{Y,\beta_i}(0,W), \tag{2.9}$$

$$r_i(W) \equiv \mathbf{1}\big\{q_{Y,\beta_i}(W) \geq 0\big\}, \tag{2.10}$$

two substitution estimators of the blip function $q_{Y,0}$ and optimal TR r_0, respectively.

For smaller sample sizes i, setting g_i equal to r_i would be hazardous. Indeed, there is no guarantee that q_{Y,β_i} estimates well $q_{Y,0}$. Say, for instance, that $q_{Y,\beta_i}(w)$ is large by mere chance for all w in a data-dependent subset S_i of $\mathcal{W}$. If we used $g_i = r_i$, then future patients with $W \in S_i$ would systematically be assigned to treatment arm $a = 1$ and the poor estimation of $q_{Y,0}$ on S_i could not be corrected, if needed. Thus, we characterize g_i by setting

$$g_i(1|W) \equiv G_i\big(q_{Y,\beta_i}(W)\big). \tag{2.11}$$

This completes the definition of the likelihood function, hence the characterization of our sampling scheme.

Note that choosing $t_1 = \cdots = t_{n_0} = 1/2$ for a limit sample size n_0 would yield $g_1 = \cdots = g_{n_0} = g^{\mathrm{b}}$, the balanced stochastic TR. Furthermore, the definitions of G_n and g_n entail straightforwardly the following lemma.

LEMMA 2.1. *Set $n \geq 1$. It holds that*

$$\inf_{w \in \mathcal{W}} g_n\big(r_n(w)|w\big) \geq 1/2, \tag{2.12}$$

$$\inf_{w \in \mathcal{W}} g_n\big(1 - r_n(w)|w\big) \geq t_n. \tag{2.13}$$

Lemma 2.1 illustrates the so-called exploration/exploitation trade-off, *that is*, the ability of the sampling scheme to exploit the accrued information (2.12) while keeping exploring in search of potential discordant new piece of information (2.13). From a different perspective, (2.12) shows that TR r_n meets the positivity assumption.

Targeted minimum loss estimator. Let $\mathcal{R}$ be the set of all TRs, *that is*, the set of all functions mapping $\mathcal{W}$ to $\{0, 1\}$. For each $g \in \mathcal{G}$ and $\rho \in \mathcal{R}$, we define a function $H_\rho(g)$ mapping $\mathcal{O}$ to $\mathbb{R}$ by setting

$$H_\rho(g)(O) \equiv \frac{\mathbf{1}\{A = \rho(W)\}}{g(A|W)}. \tag{2.14}$$

Introduce the following one-dimensional parametric model for $Q_{Y,0}$:

$$\big\{Q_{Y,\beta_n,g_n,r_n}(\epsilon) \equiv \operatorname{expit}\big(\operatorname{logit}(Q_{Y,\beta_n}) + \epsilon H_{r_n}(g_n)\big) : \varepsilon \in \mathcal{E}\big\}, \tag{2.15}$$

where $\mathcal{E} \subset \mathbb{R}$ is a closed, bounded interval containing 0 in its interior. Let ϵ_n be

$$\epsilon_n \in \arg\min_{\epsilon \in \mathcal{E}} \frac{1}{n} \sum_{i=1}^{n} L^{\mathrm{kl}}\big(Q_{Y,\beta_n,g_n,r_n}(\epsilon)\big)(O_i) \frac{g_n(A_i|W_i)}{g_i(A_i|W_i)}, \tag{2.16}$$

which indexes a minimizer of the empirical loss along the fluctuation. Define $Q^*_{Y,\beta_n,g_n,r_n} \equiv Q_{Y,\beta_n,g_n,r_n}(\epsilon_n)$ and

$$\psi_n^* \equiv \frac{1}{n} \sum_{i=1}^{n} Q^*_{Y,\beta_n,g_n,r_n}\big(r_n(W_i), W_i\big). \tag{2.17}$$

Grounded in the TMLE principle, ψ_n^* is our pivotal estimator.

3. Convergence. For every $p \geq 1$ and measurable $f : \mathcal{W} \to \mathbb{R}$, let $\|f\|_p$ be the seminorm given by

$$\|f\|_p^p \equiv \int |q_{Y,0}| \times |f|^p Q_{W,0}\, d\mu_W.$$

We introduce $g_0 \in \mathcal{G}$ given by

$$g_0(1|W) \equiv G_\infty\big(q_{Y,0}(W)\big). \tag{3.1}$$

The stochastic TR g_0 approximates the TR r_0 in the following sense:

$$\big|g_0(1|W) - r_0(W)\big| \leq t_\infty \mathbf{1}\big\{\big|q_{Y,0}(W)\big| \geq \xi_\infty\big\} + \frac{1}{2}\mathbf{1}\big\{\big|q_{Y,0}(W)\big| < \xi_\infty\big\}. \tag{3.2}$$

Therefore, if t_∞ is small and if $|q_{Y,0}(W)| \geq \xi_\infty$, then drawing A from g_0 does not differ much from drawing A from r_0. Rigorously, the distance in total variation between the Bernoulli laws with parameters $g_0(1|W)$ and $r_0(W)$ equals $2t_\infty$. On the contrary, if $|q_{Y,0}(W)| < \xi_\infty$, then the conditional laws of A given W under g_0 or r_0 may be very different. However, if ξ_∞ is small, then assigning randomly $A = 1$ or $A = 0$ has little impact on the mean value of the reward Y.

We now study the convergence of r_n to r_0 and that of g_n to g_0. In each case, the convergence is relative to two measures of discrepancy. For r_n, we consider the seminorm $\|r_n - r_0\|_p$ (any $p \geq 1$) and

$$\Delta(r_n, r_0) \equiv \big|E_{Q_0,r_n}\big(Q_{Y,0}(A, W)\big) - E_{Q_0,r_0}\big(Q_{Y,0}(A, W)\big)\big|. \tag{3.3}$$

By analogy, the measures of discrepancy for g_n are

$$\|g_n - g_0\|_p \equiv \big\|g_n(1|\cdot) - g_0(1|\cdot)\big\|_p, \tag{3.4}$$

$$\Delta(g_n, g_0) \equiv \big|E_{Q_0,g_n}\big(Q_{Y,0}(A,W)\big) - E_{Q_0,g_0}\big(Q_{Y,0}(A,W)\big)\big|. \tag{3.5}$$

Note that $\Delta(r_n, r_0)$ and $\Delta(g_n, g_0)$ are the absolute values of the differences between the mean rewards under the TRs r_n and r_0 and the stochastic TRs g_n and g_0, respectively. As such, they are targeted toward our end result, *that is*, the inference of ψ_0, as shown in the following lemma.

LEMMA 3.1. *Set $n \geq 1$. It holds that*

$$0 \leq \psi_0 - E_{Q_0,r_n}\big(Q_{Y,0}(A,W)\big) = \Delta(r_n, r_0) \leq \|r_n - r_0\|_1, \tag{3.6}$$

$$0 \leq \psi_0 - E_{Q_0,g_n}\big(Q_{Y,0}(A,W)\big) \leq \Delta(g_n, g_0) + t_\infty + \xi_\infty. \tag{3.7}$$

The next lemma shows that the convergence of q_{Y,β_n} to $q_{Y,0}$ implies that of r_n to r_0.

LEMMA 3.2. *Set $p \geq 1$. If $\|q_{Y,\beta_n} - q_{Y,0}\|_2 = o_P(1)$, then $\|r_n - r_0\|_p = o_P(1)$ hence $\Delta(r_n, r_0) = o_P(1)$.*

Similarly, the convergence of q_{Y,β_n} to $q_{Y,0}$ implies that of g_n to g_0.

LEMMA 3.3. *Set $p \geq 1$. It holds that $0 \leq \Delta(g_n, g_0) \leq \|g_n - g_0\|_p$. Moreover, if $\|q_{Y,\beta_n} - q_{Y,0}\|_2 = o_P(1)$, then $\|g_n - g_0\|_p = o_P(1)$ hence $\Delta(g_n, g_0) = o_P(1)$.*

4. Asymptotia.

4.1. *Notation.* Consider a class $\mathcal{F}$ of functions mapping a measured space $\mathcal{X}$ to $\mathbb{R}$ and $\phi : \mathbb{R} \to \mathbb{R}$. Recall that $\mathcal{F}$ is said separable if there exists a countable collection $\mathcal{F}'$ of functions such that each element of $\mathcal{F}$ is the pointwise limit of a sequence of elements of $\mathcal{F}'$. If $\phi \circ f$ is well defined for each $f \in \mathcal{F}$, then we note $\phi(\mathcal{F}) \equiv \{\phi \circ f : f \in \mathcal{F}\}$. In particular, we introduce the sets $\mathcal{G}_{1,n} \equiv \{G_n(q_Y) : Q_Y \in \mathcal{Q}_{1,n}\}$, $r(\mathcal{Q}_{1,n}) \equiv \{r(Q_Y) : Q_Y \in \mathcal{Q}_{1,n}\}$ (all $n \geq 1$) and $\mathcal{G}_1 \equiv \bigcup_{n\geq 1} \mathcal{G}_{1,n}$.

Set $\delta > 0$, μ a probability measure on $\mathcal{X}$, and let F be an envelope function for $\mathcal{F}$, *that is*, a function such that $|f(x)| \leq F(x)$ for every $f \in \mathcal{F}$, $x \in \mathcal{X}$. We denote $N(\delta, \mathcal{F}, \|\cdot\|_{2,\mu})$ the δ-covering number of $\mathcal{F}$ wrt $\|\cdot\|_{2,\mu}$, *that is*, the minimum number of $L^2(\mu)$-balls of radius δ needed to cover $\mathcal{F}$. The corresponding uniform entropy integral wrt F for $\mathcal{F}$ evaluated at δ is $J_F(\delta, \mathcal{F}) \equiv \int_0^\delta \sqrt{\log \sup_\mu N(\varepsilon \|F\|_{2,\mu}, \mathcal{F}, \|\cdot\|_{2,\mu})}\, d\varepsilon$, where the supremum is taken over all probability measures μ on the measured space $\mathcal{X}$ for which $\|F\|_{2,\mu} > 0$.

In general, given a known $g \in \mathcal{G}$ and an observation O drawn from $P_{Q_0,g}$, $Z \equiv g(A|W)$ is a deterministic function of g and O. Note that Z should be interpreted

as a weight associated with O and will be used as such. Therefore, we can augment O with Z, *i.e.*, substitute (O, Z) for O, while still denoting $(O, Z) \sim P_{Q_0,g}$. In particular, during the course of our trial, conditionally on $\mathbf{O}_{i-1}$, the stochastic TR g_i is known and we can substitute $(O_i, Z_i) = (O_i, g_i(A_i|W_i)) \sim P_{Q_0,g_i}$ for O_i drawn from P_{Q_0,g_i}. The inverse weights $1/g_i(A_i|W_i)$ are bounded because $\mathcal{G}_1$ is uniformly bounded away from 0 and 1.

The empirical distribution of $\mathbf{O}_n$ is denoted P_n. For a measurable function $f : \mathcal{O} \times [0,1] \to \mathbb{R}^d$, we use the notation $P_n f \equiv n^{-1} \sum_{i=1}^n f(O_i, Z_i)$. Likewise, for any fixed $P_{Q,g} \in \mathcal{M}$, $P_{Q,g} f \equiv E_{Q,g}(f(O, Z))$ and, for each $i = 1, \ldots, n$,

$$P_{Q_0,g_i} f \equiv E_{Q_0,g_i}\big[f(O_i, Z_i)|\mathbf{O}_{i-1}\big],$$

$$P_{Q_0,\mathbf{g}_n} f \equiv \frac{1}{n} \sum_{i=1}^n E_{Q_0,g_i}\big[f(O_i, Z_i)|\mathbf{O}_{i-1}\big].$$

The supremum norm of a function $f : \mathcal{O} \times [0,1] \to \mathbb{R}^d$ is denoted $\|f\|_\infty$. When $d = 1$, we denote $\|f\|^2_{2,P_{Q_0,g^{\mathrm{ref}}}} \equiv P_{Q_0,g^{\mathrm{ref}}} f^2$. If f is only a function of W, then $\|f\|_2 = \||q_{Y,0}|^{1/2} f\|_{2,P_{Q_0,g^{\mathrm{ref}}}}$.

For every $Q_{Y,\beta} \in \mathcal{Q}_1 \equiv \bigcup_{n\geq 1} \mathcal{Q}_{1,n}$, the blip function $Q_{Y,\beta}(1,\cdot) - Q_{Y,\beta}(0,\cdot)$ is denoted $q_{Y,\beta}$ by analogy with (2.9). We will often deal with seminorms $\|f\|_2$ with $f = Q_Y - Q_{Y,\beta_0}$ for some $Q_Y \in \mathcal{Q}_Y$ and $\mathcal{Q}_{Y,\beta_0} \in \mathcal{Q}_1$. A consequence of the trivial inequality $(a-b)^2 \leq 2(ua^2 + (1-u)b^2)/\min(u, 1-u)$ (valid for all $a, b \in \mathbb{R}$, $0 < u < 1$), the following bound will prove useful:

$$\begin{aligned} \|q_Y - q_{Y,\beta_0}\|_2 &\leq 2\big\||q_{Y,0}|^{1/2}/g^{\mathrm{ref}}\big\|_\infty \times \|Q_Y - Q_{Y,\beta_0}\|_{2,P_{Q_0,g^{\mathrm{ref}}}} \\ &\leq 2\big\|1/g^{\mathrm{ref}}\big\|_\infty \times \|Q_Y - Q_{Y,\beta_0}\|_{2,P_{Q_0,g^{\mathrm{ref}}}}. \end{aligned} \tag{4.1}$$

The constant $2\|1/g^{\mathrm{ref}}\|_\infty$ is minimized at $g^{\mathrm{ref}} = g^{\mathrm{b}}$, with $2\|1/g^{\mathrm{b}}\|_\infty = 4$.

4.2. *Central limit theorem.* Our main result is a central limit theorem for ψ_n^*. It relies on the following assumptions, upon which we comment in Section 4.3.

A1. The conditional distribution of Y given (A, W) under Q_0 is not a degenerate law. Moreover, $P_{Q_0}(|q_{Y,0}(W)| > 0) = 1$.

Existence and convergence of projections.

A2. For each $n \geq 1$, there exists $Q_{Y,\beta_{n,0}} \in \mathcal{Q}_{1,n}$ satisfying

$$P_{Q_0,g^{\mathrm{ref}}} L(Q_{Y,\beta_{n,0}}) = \inf_{Q_{Y,\beta} \in \mathcal{Q}_{1,n}} P_{Q_0,g^{\mathrm{ref}}} L(Q_{Y,\beta}).$$

Moreover, there exists $Q_{Y,\beta_0} \in \mathcal{Q}_1$ such that, for all $\delta > 0$,

$$P_{Q_0,g^{\mathrm{ref}}} L(Q_{Y,\beta_0}) < \inf_{\{Q_Y \in \mathcal{Q}_1 : \|Q_Y - Q_{Y,\beta_0}\|_{2,P_{Q_0,g^{\mathrm{ref}}}} \geq \delta\}} P_{Q_0,g^{\mathrm{ref}}} L(Q_Y).$$

Finally, it holds that $q_{Y,\beta_0} = q_{Y,0}$.

A3. For all $\rho \in \mathcal{R}$ and $\epsilon \in \mathcal{E}$, introduce

$$Q_{Y,\beta_0,g_0,\rho}(\epsilon) \equiv \operatorname{expit}\bigl(\operatorname{logit}(Q_{Y,\beta_0}) + \epsilon H_\rho(g_0)\bigr), \tag{4.2}$$

where $H_\rho(g_0)$ is given by (2.14) with $g = g_0$. For every $\rho \in \mathcal{R}$, there exists a unique $\epsilon_0(\rho) \in \mathcal{E}$ such that

$$\epsilon_0(\rho) \in \arg\min_{\epsilon\in\mathcal{E}} P_{Q_0,g_0} L^{\mathrm{kl}}\bigl(Q_{Y,\beta_0,g_0,\rho}(\epsilon)\bigr). \tag{4.3}$$

Reasoned complexity.

A4. The classes $\mathcal{Q}_{1,n}$, $L(\mathcal{Q}_{1,n})$ and $r(\mathcal{Q}_{1,n})$ are separable. Moreover, the following entropy conditions hold: $J_1(1, \mathcal{Q}_{1,n}) = o(\sqrt{n})$, $J_1(1, r(\mathcal{Q}_{1,n})) = o(\sqrt{n})$, $J_{F_n}(1, L(\mathcal{Q}_{1,n})) = o(\sqrt{n})$, where each F_n is an envelope function for $L(\mathcal{Q}_{1,n})$.

A4*. Let $\{\delta_n\}_{n\geq 1}$ be a sequence of positive numbers. If $\delta_n = o(1)$, then $J_1(\delta_n, \mathcal{Q}_{1,n}) = o(1)$ and $J_1(\delta_n, r(\mathcal{Q}_{1,n})) = o(1)$.

Margin condition.

A5. There exist $\gamma_1, \gamma_2 > 0$ such that, for all $t \geq 0$,

$$P_{Q_0}\bigl(0 < \bigl|q_{Y,0}(W)\bigr| \leq t\bigr) \leq \gamma_1 t^{\gamma_2}.$$

We first focus on the convergence of the sequences of stochastic TRs g_n and empirical TR r_n. By Lemmas 3.2 and 3.3, it suffices to consider the convergence of q_{Y,β_n}. By (4.1), we may consider the convergence of Q_{Y,β_n}. By adapting the classical scheme of analysis of M-estimators and exploitation of A5, we derive the following result.

PROPOSITION 4.1. *Under* A2 *and* A4, *both* $\|Q_{Y,\beta_n} - Q_{Y,\beta_0}\|_{2,P_{Q_0,g^{\mathrm{ref}}}} = o_P(1)$ *and* $\|q_{Y,\beta_n} - q_{Y,0}\|_2 = o_P(1)$. *Hence, for any* $p \geq 1$, $\|r_n - r_0\|_p = o_P(1)$, $\|g_n - g_0\|_p = o_P(1)$, $\Delta(r_n, r_0) = o_P(1)$, $\Delta(g_n, g_0) = o_P(1)$ *by Lemmas* 3.2 *and* 3.3. *If* A1 *and* A5 *are also met, then* $\|r_n - r_0\|_{2,P_{Q_0,g^{\mathrm{ref}}}} = o_P(1)$ *and* $\|g_n - g_0\|_{2,P_{Q_0,g^{\mathrm{ref}}}} = o_P(1)$ *as well.*

Define now the data-adaptive parameter

$$\psi_{r_n,0} \equiv E_{Q_0}\bigl(Q_{Y,0}(r_n(W), W)\bigr) = E_{Q_{0,r_n}}\bigl(Q_{Y,0}(A, W)\bigr). \tag{4.4}$$

By (3.6) in Lemma 3.1 and Lemma 3.2, we have the following corollary to Proposition 4.1.

COROLLARY 4.1. *Under* A2 *and* A4, $0 \leq \psi_0 - \psi_{r_n,0} = o_P(1)$.

We now turn to the convergence of ψ_n^*. Its asymptotic behavior can be summarized in these terms.

THEOREM 4.1. *Suppose that* A1, A2, A3, A4, A4* *and* A5 *are met. It holds that* $\psi_n^* - \psi_{r_n,0} = o_P(1)$. *Thus, by Corollary* 4.1, $\psi_n^* - \psi_0 = o_P(1)$ *as well. Moreover,* $\sqrt{n/\Sigma_n}(\psi_n^* - \psi_{r_n,0})$ *is approximately standard normally distributed, where* Σ_n *is the explicit estimator given in* (4.14).

Theorem 4.1 is a toned down version of Theorem 4.2 that we state and comment on in Section 4.5. Section 4.3 discusses their assumptions and Section 4.4 presents an example. Theorems 4.1 and 4.2 allow the construction of confidence intervals for several parameters of interest, as shown in Section 5.

4.3. *Commenting on the assumptions.* Assumption A1 consists in two statements. The first one is a simple condition guaranteeing that the limit variance of $\sqrt{n}(\psi_n^* - \psi_{r_n,0})$ is positive. The second one is more stringent. In the terminology of [25], it states that Q_0 is not exceptional. If Q_0 were exceptional, then the set $\{w \in \mathcal{W} : q_{Y,0}(w) = 0\}$ would have positive probability under Q_0. To a patient falling in this set, the optimal TR $r(Q_{Y,0}) \equiv r_0$ recommends to assign treatment $A = 1$ instead of treatment $A = 0$. This arbitrary choice has no consequence whatsoever in terms of conditional mean of the reward given treatment and baseline covariates.

However, it is well documented that exceptional laws are problematic. For the estimation of the optimal TR r_0, one reason is that an estimator will typically not converge to a fixed limit on $\{w \in \mathcal{W} : q_{Y,0}(w) = 0\}$ [19, 25, 26]. Another reason is that the mean reward under the optimal TR seen as a functional, Ψ, is pathwise differentiable at Q_0 if and only if, Q_0-almost surely, either $|q_{Y,0}(W)| > 0$ or the conditional distributions of Y given $(A = 1, W)$ and $(A = 0, W)$ under Q_0 are a degenerate law [19], Theorem 1. This explains why it is also assumed that the true law is not exceptional in [18, 20, 34]. Other approaches have been considered to circumvent the need to make this assumption: relying on m-out-of-n bootstrap [3] [at the cost of a $\sqrt{m} = o(\sqrt{n})$-rate of convergence and need to fine-tune m], or changing the parameter of interest by focusing on the mean reward under the optimal TR conditional on patients for whom the best treatment has a clinically meaningful effect (truncation) [12, 16, 17].

To the best of our knowledge, only [19] addresses the inference of the original parameter at a $\sqrt{n}$-rate of convergence without assuming that the true law is not exceptional. Moreover, if the true law is not exceptional, then the estimator is asymptotically efficient among all regular and asymptotically linear estimators. Developed in the i.i.d. setting, the estimator of [19] does not require that the estimator of r_0 converge as the sample size grows. It relies on a clever iteration of a two-step procedure consisting in (i) estimating well-chosen nuisance parameters, including r_0, on a small chunk of data, then (ii) constructing an estimator targeted to the mean reward under the current estimate of r_0 with the nuisance parameters obtained in (i). The final estimator is a weighted average of the resulting

chunk-specific estimators. Adapting this procedure to our setting where data are dependent would be very challenging.

Assumptions A2 states the existence of L-projections $Q_{Y,\beta_{n,0}}$ of $Q_{Y,0}$ onto each working model $\mathcal{Q}_{1,n}$ and their convergence to a limit L-projection $Q_{Y,\beta_0} \in \mathcal{Q}_1 \equiv \bigcup_{n\geq 1} \mathcal{Q}_{1,n}$. More importantly, it states that the blip function q_{Y,β_0} associated with Q_{Y,β_0} equals the true blip function $q_{Y,0}$ associated with $Q_{Y,0}$.

For any fixed TR $\rho \in \mathcal{R}$, the limit L-projection Q_{Y,β_0} can be fluctuated in a direction $H_\rho(g_0)$ characterized by ρ and $Q_{Y,0}$; see (2.14), (3.1) and (4.2). Assumption A3 states that there exists a unique L^{kl}-projection of $Q_{Y,0}$ onto this ρ-specific one-dimensional parametric model fluctuating Q_{Y,β_0}. In particular, when $\rho = r_n$, the estimator of r_0 at sample size n, $Q_{Y,0}$ is uniquely L^{kl}-projected onto, say, $Q^*_{Y,0,r_n}$. One of the keys to our approach is the equality $E_{Q_0}(Q^*_{Y,0,r_n}(r_n(W), W)) = \psi_{r_n,0} \equiv E_{Q_0}(Q_{Y,0}(r_n(W), W))$ even if $Q_{Y,0}$ and $Q^*_{Y,0,r_n}$ differ. Proven in step 3 of the proof of [8], Proposition A.1, which states that ψ^*_n is a consistent estimator of $\psi_{r_n,0}$ [*i.e.*, $\psi^*_n - \psi_{r_n,0} = o_P(1)$], this robustness property is a by product of the robustness of the efficient influence curve of the mean reward under r_n treated as a fixed TR; see [8], Lemma C.1.

Expressed in terms of separability and conditions on uniform entropy integrals, A4 and A4* restrict the complexities of the working models $\mathcal{Q}_{1,n}$ and resulting classes $r(\mathcal{Q}_{1,n})$ and $L(\mathcal{Q}_{1,n})$. Imposing separability is a convenient way to ensure that some delicate measurability conditions are met. Assumption A4* partially strengthens A4 because choosing $\delta_n \equiv 1/\sqrt{n}$ (all $n \geq 1$) in A4* implies $J_1(1, \mathcal{F}_n) = o(\sqrt{n})$ for both $\mathcal{F}_n \equiv \mathcal{Q}_{1,n}$ and $\mathcal{F}_n \equiv r(\mathcal{Q}_{1,n})$ by a simple change of variable. Section 4.4 presents an example of sequence $\{\mathcal{Q}_{1,n}\}_{n\geq 1}$ of working models which meets A4 and A4*. Its construction involves VC-classes of functions, which are archetypal examples of classes with well-behaved uniform entropy integrals. Restricting the complexities of the working models $\mathcal{Q}_{1,n}$, $r(\mathcal{Q}_{1,n})$ and $L(\mathcal{Q}_{1,n})$ in terms of bracketing entropy is tempting because of the great diversity of examples of classes of functions which behave well in these terms. Unfortunately, this is not a viable alternative, since bounds on the bracketing numbers of $\mathcal{Q}_{1,n}$ do not imply bounds on those of $r(\mathcal{Q}_{1,n})$. As a result, we have to prove a new maximal inequality for martingales wrt the uniform entropy integral to control several empirical processes indexed by random functions; see [8], Lemma B.3.

Inspired from the seminal article [21], assumptions similar to A5 are known as "margin assumptions" in the literature. They describe how the data-distribution concentrates on adverse events, *that is*, on events that make inference more difficult. We have already discussed the fact that inferring the optimal TR and its mean reward is less challenging when the law of the absolute value of $|q_{Y,0}(W)|$ puts no mass on $\{0\}$. It actually occurs that the less mass this law puts *around* $\{0\}$, the less challenging is the inference. Assumption A5 formalizes tractable concentrations. It has already proven useful in the i.i.d. setting; see [18], Lemma 1, and [19], Condition (16). By Markov's inequality, A5 is implied by the following, clearer assumption:

A5.** There exists $\gamma_2 > 0$ such that

$$\gamma_1 \equiv E_{Q_0}\big(|q_{Y,0}(W)|^{-\gamma_2}\mathbf{1}\{|q_{Y,0}(W)| > 0\}\big) < \infty.$$

4.4. *An example*. In this section, we construct a sequence $\{\mathcal{Q}_{1,n}\}_{n\geq 1}$ of working models which meets A4 and A4*, see Proposition 4.2. Let $\mathcal{F}^-$ be a separable class of measurable functions from $\mathcal{W}$ to $[-1, 1] \setminus \{0\}$ such that $\{\{w \in \mathcal{W} : f^-(w) \geq t\} : f^- \in \mathcal{F}^-, t \in [-1, 1]\}$ is a VC-class of sets. By definition, $\mathcal{F}^-$ is a VC-major class [33], Sections 2.6.1 and 2.6.4. Thus, Corollary 2.6.12 in [33] guarantees the existence of two constants $K^- > 0$ and $\alpha^- \in [0, 1)$ such that, for every $\varepsilon > 0$,

$$\log \sup_{\mu} N\big(\varepsilon \|1\|_{2,\mu}, \mathcal{F}^-, \|\cdot\|_{2,\mu}\big) \leq K^-\left(\frac{1}{\varepsilon}\right)^{2\alpha^-}. \tag{4.5}$$

Let $\mathcal{F}^+$ be a separable class of measurable functions from $\mathcal{W}$ to $[0, 2]$ such that, for two constants $K^+ > 0$, $\alpha^+ \in [0, 1)$ and for every $\varepsilon > 0$,

$$\log \sup_{\mu} N\big(\varepsilon \|2\|_{2,\mu}, \mathcal{F}^+, \|\cdot\|_{2,\mu}\big) \leq K^+\left(\frac{1}{\varepsilon}\right)^{2\alpha^+}. \tag{4.6}$$

For instance, $\mathcal{F}^+$ may be a VC-hull class of functions, *that is*, a subset of the pointwise sequential closure of the symmetric convex hull of a VC-class of functions [33], Section 2.6.3. (The suprema in (4.5) and (4.6) are taken over all probability measures μ on the measured space $\mathcal{W}$.)

We now use $\mathcal{F}^-$ and $\mathcal{F}^+$ to define the sequence $\{\mathcal{Q}_{1,n}\}_{n\geq 1}$ of working models. Let $\mathcal{F}^- = \bigcup_{n\geq 1} \mathcal{F}_n^-$ and $\mathcal{F}^+ = \bigcup_{n\geq 1} \mathcal{F}_n^+$ be rewritten as the limits of two increasing sequences of sets $\{\mathcal{F}_n^-\}_{n\geq 1}$ and $\{\mathcal{F}_n^+\}_{n\geq 1}$. Set $n \geq 1$ and define

$$B_n \equiv \big\{(f^-, f^+) \in \mathcal{F}_n^- \times \mathcal{F}_n^+ : 0 \leq f^+ + f^-, f^+ - f^- \leq 2\big\}.$$

For each $\beta \equiv (f^-, f^+) \in B_n$, introduce $Q_{Y,\beta}$ mapping $\mathcal{A} \times \mathcal{W}$ to $[0, 1]$ characterized by

$$Q_{Y,\beta}(A, W) = \frac{A}{2}\big(f^+(W) + f^-(W)\big) + \frac{(1 - A)}{2}\big(f^+(W) - f^-(W)\big). \tag{4.7}$$

We define the nth working model as $\mathcal{Q}_{1,n} \equiv \{Q_{Y,\beta} : \beta \in B_n\}$. It is separable because $\mathcal{F}^-$ and $\mathcal{F}^+$ are separable.

Because $q_{Y,\beta} \equiv Q_{Y,\beta}(1, \cdot) - Q_{Y,\beta}(0, \cdot) = f^-$ for every $\beta \equiv (f^-, f^+) \in B_n$, it holds that

$$\begin{aligned} r(\mathcal{Q}_{1,n}) &\equiv \big\{\mathbf{1}\{q_{Y,\beta}(\cdot) \geq 0\} : \beta \in B_n\big\} \\ &= \big\{\mathbf{1}\{f^-(\cdot) \geq 0\} : f^- \in \mathcal{F}_n^-\big\} \subset \big\{\mathbf{1}\{f^-(\cdot) \geq 0\} : f^- \in \mathcal{F}^-\big\} \end{aligned}$$

which, by construction, is a fixed subset of a VC-class of functions, hence a VC-class of functions itself. Moreover, $r(\mathcal{Q}_{1,n})$ is separable because $\mathcal{F}^-$ is separable and elements of $\mathcal{F}^-$ take only positive or negative values. These properties and (4.5), (4.6) are the main arguments in the proof of the following result.

PROPOSITION 4.2. *The sequence* $\{\mathcal{Q}_{1,n}\}_{n\geq 1}$ *of working models satisfies* A4 (*with* $L = L^{\mathrm{ls}}$ *the least-square loss*) *and* A4*.

4.5. *Asymptotic linear expansion, resulting central limit theorem.* Theorem 4.1 is a summary of Theorem 4.2 below, whose main result is the asymptotic linear expansion (4.15). The statement of Theorem 4.2 requires additional notation.

Let $Q^*_{Y,0}$, $d^*_{W,0}$, $d^*_{Y,0}$ and Σ_0 be given by

$$Q^*_{Y,0}(A,W) \equiv Q_{Y,\beta_0,g_0,r_0}\big(\epsilon_0(r_0)\big)(A,W), \tag{4.8}$$

$$d^*_{W,0}(W) \equiv Q^*_{Y,0}\big(r_0(W),W\big) - E_{Q_0}\big(Q^*_{Y,0}\big(r_0(W),W\big)\big), \tag{4.9}$$

$$d^*_{Y,0}(O,Z) \equiv \frac{\mathbf{1}\{A = r_0(W)\}}{Z}\big(Y - Q^*_{Y,0}(A,W)\big), \tag{4.10}$$

$$\Sigma_0 \equiv P_{Q_0,g_0}\big(d^*_{W,0} + d^*_{Y,0}\big)^2. \tag{4.11}$$

Analogously, recall that $Q^*_{Y,\beta_n,g_n,r_n} \equiv Q_{Y,\beta_n,g_n,r_n}(\epsilon_n)$ and let $d^*_{W,n}$, $d^*_{Y,n}$ and Σ_n be given by

$$d^*_{W,n}(W) \equiv Q^*_{Y,\beta_n,g_n,r_n}\big(r_n(W),W\big) - \psi^*_n, \tag{4.12}$$

$$d^*_{Y,n}(O,Z) \equiv \frac{\mathbf{1}\{A = r_n(W)\}}{Z}\big(Y - Q^*_{Y,\beta_n,g_n,r_n}(A,W)\big), \tag{4.13}$$

$$\Sigma_n \equiv P_n\big(d^*_{W,n} + d^*_{Y,n}\big)^2. \tag{4.14}$$

Note that $d^*_{W,n}$, $d^*_{Y,n}$ and Σ_n are empirical counterparts to $d^*_{W,0}$, $d^*_{Y,0}$ and Σ_0.

THEOREM 4.2. *Suppose that* A1, A2, A3, A4, A4* *and* A5 *are met. It holds that* $\psi^*_n - \psi_{r_n,0} = o_P(1)$. *Thus, by Corollary* 4.1, $\psi^*_n - \psi_0 = o_P(1)$ *as well. Moreover,* $\Sigma_n = \Sigma_0 + o_P(1)$ *with* $\Sigma_0 > 0$ *and*

$$\psi^*_n - \psi_{r_n,0} = (P_n - P_{Q_0,\mathbf{g}_n})\big(d^*_{Y,0} + d^*_{W,0}\big) + o_P(1/\sqrt{n}). \tag{4.15}$$

Consequently, $\sqrt{n/\Sigma_n}(\psi^*_n - \psi_{r_n,0})$ *converges in law to the standard normal distribution.*

Consider (4.9). It actually holds that the term $E_{Q_0}(Q^*_{Y,0}(r_0(W),W))$ equals $\psi_0 \equiv E_{Q_0}(Q_{Y,0}(r_0(W),W))$ (see step one of the proof of Corollary A.1 in [8], Section A.2). This proximity between $Q_{Y,0}$ and $Q^*_{Y,0}$ follows from the careful fluctuation of Q_{Y,β_0}. The proof of Theorem 4.2 mainly consists in showing the consistency of ψ^*_n and (4.15). It is delicate because it hinges on the control of empirical processes indexed by random functions which, similar to $Q^*_{Y,\beta_n,g_n,r_n} \equiv Q_{Y,\beta_n,g_n,r_n}(\epsilon_n)$, are defined stepwise ($\beta_n$ yields g_n and r_n, and altogether they yield ϵ_n).

Set $Q_0^* \equiv (Q_{W,0}\, d\mu_W, Q_{Y,0}^*(\cdot|a,w), (a,w) \in \mathcal{A}\times\mathcal{W}) \in \mathcal{Q}$. The influence function $d_{Y,0}^* + d_{W,0}^*$ in (4.15) is closely related to the efficient influence curve $D_{r_0}(Q_0^*, g_0)$ at $P_{Q_0^*,g_0}$ of the mapping $\Psi_{r_0} : \mathcal{M} \to [0,1]$ characterized by

$$\Psi_{r_0}(P_{Q,g}) \equiv E_Q\big(Q_Y\big(r_0(W), W\big)\big), \tag{4.16}$$

the mean reward under Q of the TR r_0 (possibly different from the optimal TR $r(Q_Y)$ under Q) treated as known and fixed. Specifically, in light of Lemma C.1 in [8], Section C,

$$d_{Y,0}^*(O,Z) + d_{W,0}^*(W) = D_{r_0}\big(Q_0^*, g_0\big)(O)$$

when $Z = g_0(A|W)$. Consequently, $\Sigma_0 = P_{Q_0,g_0} D_{r_0}(Q_0^*, g_0)^2$.

If $Q_{Y,\beta_0} = Q_{Y,0}$ (a stronger condition than equality $q_{Y,\beta_0} = q_{Y,0}$ in A2), then $Q_{Y,0}^* = Q_{Y,0}$ (because $\epsilon_0(r_0)$ from A3 equals zero), hence $Q_0^* = Q_0$ and, finally, the remarkable equality $\Sigma_0 = P_{Q_0,g_0} D_{r_0}(Q_0, g_0)^2$: the asymptotic variance of $\sqrt{n}(\psi_n^* - \psi_{r_n,0})$ coincides with the generalized Cramér–Rao lower bound for the asymptotic variance of any regular and asymptotically linear estimator of $\Psi_{r_0}(P_{Q_0,g_0})$ when sampling independently from P_{Q_0,g_0} (see [8], Lemma C.1). Otherwise, the discrepancy between Σ_0 and $P_{Q_0,g_0} D_{r_0}(Q_0, g_0)^2$ will vary depending on that between Q_{Y,β_0} and $Q_{Y,0}$, hence in particular on the user-supplied sequence $\{\mathcal{Q}_{1,n}\}_{n\geq 1}$ of working models. Studying this issue in depth is very difficult, if at all possible, and beyond the scope of this article.

5. Confidence regions. We explore how Theorems 4.1 and 4.2 enable the construction of confidence intervals for various possibly data-adaptive parameters: the mean rewards under the optimal TR and under its current estimate in Section 5.1; the empirical cumulative pseudo-regret in Section 5.2; the counterfactual cumulative pseudo-regret in Section 5.3.

Set a confidence level $\alpha \in (0, 1/2)$. Let $\xi_\alpha < 0$ and $\xi_{1-\alpha/2} > 0$ be the corresponding α- and $(1-\alpha/2)$-quantiles of the standard normal distribution.

5.1. *Confidence intervals for the mean rewards under the optimal treatment rule and under its current estimate.* Theorems 4.1 and 4.2 yield straightforwardly a confidence interval for the mean reward under the current best estimate of the optimal TR, $\psi_{r_n,0}$.

PROPOSITION 5.1. *Under the assumptions of Theorems 4.1 or 4.2, the probability of the event*

$$\psi_{r_n,0} \in \left[\psi_n^* \pm \xi_{1-\alpha/2}\sqrt{\frac{\Sigma_n}{n}}\right]$$

converges to $(1-\alpha)$ *as n goes to infinity.*

We need to strengthen A5 to guarantee that the confidence interval in Proposition 5.1 can also be used to infer the mean reward under the optimal TR, ψ_0. Consider thus the following:

A5*. There exist $\gamma_1 > 0$, $\gamma_2 \geq 1$ such that, for all $t \geq 0$,

$$P_{Q_0}\big(0 < \big|q_{Y,0}(W)\big| \leq t\big) \leq \gamma_1 t^{\gamma_2}.$$

Just like A5 is a consequence of A5**, A5* is a consequence of A5** where one substitutes the condition $\gamma_2 > 0$ for the stronger condition $\gamma_2 \geq 1$.

PROPOSITION 5.2. *Under* A5**, *there exists a constant* $c > 0$ *such that*

$$0 \leq \psi_0 - \psi_{r_n,0} \leq c\|q_{Y,\beta_n} - q_{Y,0}\|_2^{2(1+\gamma_2)/(3+\gamma_2)}. \tag{5.1}$$

Set $\gamma_3 \equiv 1/4 + 1/2(1+\gamma_2) \in (1/4, 1/2]$. *By* (5.1), *if* $\|Q_{Y,\beta_n} - Q_{Y,\beta_0}\|_{2,P_{Q_0,g^{\mathrm{ref}}}} = o_P(1/n^{\gamma_3})$, *then* $\|q_{Y,\beta_n} - q_{Y,0}\|_2 = o_P(1/n^{\gamma_3})$, *which implies* $0 \leq \psi_0 - \psi_{r_n,0} = o_P(1/\sqrt{n})$.

Therefore, *if the assumptions of Theorems* 4.1 *or* 4.2 *are also met*, *then the probability of the event*

$$\psi_0 \in \left[\psi_n^* \pm \xi_{1-\alpha/2}\sqrt{\frac{\Sigma_n}{n}}\right]$$

converges to $(1-\alpha)$ *as* n *goes to infinity.*

The definition of γ_3 in Proposition 5.2 justifies the requirement $\gamma_2 \geq 1$ in A5*. Indeed, $\gamma_3 \leq 1/2$ is equivalent to $\gamma_2 \geq 1$. Moreover, it holds that $\gamma_3 = 1/2$ [so that $\|q_{Y,\beta_n} - q_{Y,0}\|_2 = o_P(1/n^{\gamma_3})$ can be read as a parametric rate of convergence] if and only if $\gamma_2 = 1$.

5.2. *Lower confidence bound for the empirical cumulative pseudo-regret.* We call

$$\mathscr{E}_n \equiv \frac{1}{n}\sum_{i=1}^{n}\big(Y_i - Q_{Y,0}\big(r_n(W_i), W_i\big)\big) \tag{5.2}$$

the "empirical cumulative pseudo-regret" at sample size n. A data-adaptive parameter, it is the difference between the average of the *actual* rewards garnered so far, $n^{-1}\sum_{i=1}^{n} Y_i$, and the average of the *mean* rewards under the current estimate r_n of the optimal TR r_0 in the successive contexts drawn so far during the course of the experiment,

$$n^{-1}\sum_{i=1}^{n} Q_{Y,0}\big(r_n(W_i), W_i\big).$$

The former is a known quantity, so the challenge is to infer the latter. Moreover, we are mainly interested in obtaining a lower confidence bound.

Define $\Sigma_0^{\mathscr{E}}$ and $\Sigma_n^{\mathscr{E}}$ respectively equal to

$$E_{Q_0,g_0}\big(d^*_{W,0}(W) - \big(Q_{Y,0}\big(r_0(W), W\big) - \psi_0\big) + d^*_{Y,0}(O, Z)\big)^2, \tag{5.3}$$

$$\frac{1}{n}\sum_{i=1}^{n}\big(d^*_{W,n}(W_i) - \big(Q_{Y,\beta_n}\big(r_n(W_i), W_i\big) - \psi_n^0\big) + d^*_{Y,n}(O_i, Z_i)\big)^2, \tag{5.4}$$

with $\psi_n^0 \equiv n^{-1}\sum_{i=1}^n Q_{Y,\beta_n}(r_n(W_i), W_i)$. Note that $\Sigma_n^{\mathscr{E}}$ is an empirical counterpart to $\Sigma_0^{\mathscr{E}}$. The key to the derivation of the lower confidence bound presented below is (4.15), from which we deduce another asymptotic linear expansion for $\sqrt{n}(\psi_n^* + \mathscr{E}_n - n^{-1}\sum_{i=1}^n Y_i)$.

PROPOSITION 5.3. *Under the assumptions of Theorems* 4.1 *or* 4.2, *the probability of the event*

$$\mathscr{E}_n \geq \frac{1}{n}\sum_{i=1}^{n} Y_i - \psi_n^* + \xi_\alpha\sqrt{\frac{\Sigma_n^{\mathscr{E}}}{n}}$$

converges to $(1-\alpha)$ *as* n *goes to infinity.*

5.3. *Lower confidence bound for the counterfactual cumulative pseudo-regret.* In this section, we cast our probabilistic model in a causal model. We postulate the existence of counterfactual rewards $Y_n(1)$ and $Y_n(0)$ of assigning treatment $a = 1$ and $a = 0$ to the nth patient (all $n \geq 1$). They are said counterfactual because it is impossible to observe them jointly. The observed nth reward writes $Y_n = A_n Y_n(1) + (1 - A_n)Y_n(0)$.

We call

$$\mathscr{C}_n \equiv \frac{1}{n}\sum_{i=1}^{n}\big(Y_i - Y_i\big(r_n(W_i)\big)\big) \tag{5.5}$$

the "counterfactual cumulative pseudo-regret" at n. It is the difference between the average of the *actual* rewards garnered so far, $n^{-1}\sum_{i=1}^n Y_i$, and the average of the *counterfactual* rewards under the current estimate r_n of the optimal TR r_0 in the successive contexts drawn so far during the course of the experiment, $n^{-1}\sum_{i=1}^n Y_i(r_n(W_i))$. Once more, the former is a known quantity, so the challenge is to infer the latter. Moreover, we are mainly interested in obtaining a lower confidence bound.

For simplicity, we adopt the so called "nonparametric structural equations" approach [22]. So, we actually postulate the existence of a sequence $\{U_n\}_{n\geq 1}$ of i.i.d. random variables independent from $\{O_n\}_{n\geq 1}$ with values in $\mathcal{U}$ and that of a deterministic measurable function $\mathbb{Q}_{Y,0}$ mapping $\mathcal{A}\times\mathcal{W}\times\mathcal{U}$ to $\mathcal{Y}$ such that, for every $n \geq 1$ and both $a = 0, 1$,

$$Y_n(a) = \mathbb{Q}_{Y,0}(a, W_n, U_n).$$

The notation $\mathbb{Q}_{Y,0}$ is motivated by the following property. Let $(A, W, U) \in \mathcal{A} \times \mathcal{W} \times \mathcal{U}$ be distributed from $\mathbb{P}$ in such a way that (i) A is conditionally independent from U given W, and (ii) with $Y \equiv A\mathbb{Q}_{Y,0}(1, W, U) + (1 - A)\mathbb{Q}_{Y,0}(0, W, U)$, the conditional distribution of Y given (A, W) is $Q_{Y,0}(\cdot|A, W)\, d\mu_Y$. Then, for each $a \in \mathcal{A}$,

$$\begin{aligned} E_{\mathbb{P}}\big(\mathbb{Q}_{Y,0}(a, W, U)|W\big) &= E_{\mathbb{P}}\big(\mathbb{Q}_{Y,0}(a, W, U)|A = a, W\big) \\ &= E_{\mathbb{P}}(Y|A = a, W) \\ &= Q_{Y,0}(a, W). \end{aligned} \tag{5.6}$$

Although $\mathscr{C}_n$ is by nature a counterfactual data-adaptive parameter, it is possible to construct a conservative lower confidence bound yielding a confidence interval whose asymptotic coverage is no less than $(1 - \alpha)$.

PROPOSITION 5.4. *Under the assumptions of Theorems 4.1 or 4.2, the probability of the event*

$$\mathscr{C}_n \geq \frac{1}{n}\sum_{i=1}^{n} Y_i - \psi_n^* + \xi_\alpha \sqrt{\frac{\Sigma_n^{\mathscr{E}}}{n}}$$

converges to $(1 - \alpha') \geq (1 - \alpha)$ *as* n *goes to infinity.*

The key to this result is threefold. First, the asymptotic linear expansion (4.15) still holds in the above causal model where each observation (O_n, Z_n) is augmented with U_n (every $n \geq 1$). Second, the expansion yields a confidence interval with asymptotic level $(1-\alpha)$. Unfortunately, its asymptotic width depends on features of the causal distribution which are not identifiable from the real-world (as opposed to causal) distribution. Third, and fortunately, $\Sigma_n^{\mathscr{E}}$ is a conservative estimator of the limit width. We refer the reader to the proof of Proposition 5.4 in [8], Section A.3, for details. It draws inspiration from [1], where the same trick was first devised to estimate the so-called sample average treatment effect.

Linear contextual bandit problems. Consider the following contextual bandit problem: an agent is sequentially presented a context $w_t \in \mathbb{R}^d$, has to choose an action $a_t \in \{0, 1\}$, and receives a random reward $y_t = f(a_t, w_t) + \varepsilon_t$, with f an unknown real-valued function and ε_t a centered, typically sub-Gaussian noise. The agent aims at maximizing the cumulated sum of rewards. The contextual bandit problem is linear if there exists $\theta \equiv (\theta_0, \theta_1) \in \mathbb{R}^{2d}$ such that $f(a, w) \equiv w^\top \theta_a$ for all $(a, w) \in \{0, 1\} \times \mathbb{R}^d$. At time t, the best action is $a_t^* \equiv \arg\max_{a=0,1} w_t^\top \theta_a$ and maximizing the cumulated sum of rewards is equivalent to minimizing the cumulated pseudo-regret $R_T^\theta \equiv \sum_{t=1}^T w_t^\top (a_t^* \theta_{a_t^*} - a_t \theta_{a_t})$.

We refer to [15], Chapter 4, for an overview of the literature dedicated to this problem, which bears evident similitudes with our problem of interest. Optimistic

algorithms consist in constructing a frequentist region of confidence for θ and choosing that action a_{t+1} maximizing $a \mapsto \max_{\vartheta} w_{t+1}^{\top}\vartheta_a$ where ϑ ranges over the confidence region. The Bayes-UCB algorithm and its variants follow the same idea with Bayesian regions of confidence substituted for the frequentist ones. As for the celebrated Thompson sampling algorithm, it consists in drawing $\tilde{\theta}$ from the posterior distribution of θ and choosing that action a_{t+1} maximizing $a \mapsto w_{t+1}^{\top}\tilde{\theta}_a$. Each time estimating θ (which is essentially equivalent to estimating the optimal TR and its mean reward) is a means to an end.

Various frequentist analyses of such algorithms have been proposed. It notably appears that the cumulated pseudo-regret R_T^{θ} typically scales in $\tilde{O}(\sqrt{T})$ with high probability, where $\tilde{O}$ ignores logarithmic factors in T. This is consistent with the form of the lower confidence bounds that we obtain, as by products rather than main objectives and under milder assumptions on $f/Q_{Y,0}$, for our empirical and counterfactual cumulated pseudo-regrets.

6. Simulation study. The simulation study is conducted in `R` [29], using the package `tsml.cara.rct` designed for this purpose [5]. The package comes with a "vignette" showing how it works. In particular, the vignette contains a step-by-step guide to carrying out a simulation study.

6.1. *Setup*. We now present the results of a simulation study. Under Q_0, the baseline covariate W decomposes as $W \equiv (U, V) \in [0, 1] \times \{1, 2, 3\}$, where U and V are independent random variables respectively drawn from the uniform distribution on $[0, 1]$ and such that $P_{Q_0}(V = 1) = \frac{1}{2}$, $P_{Q_0}(V = 2) = \frac{1}{3}$ and $P_{Q_0}(V = 3) = \frac{1}{6}$. Moreover, Y is conditionally drawn given (A, W) from the Beta distribution with a constant variance set to 0.01 and a mean $Q_{Y,0}(A, W)$ satisfying

$$Q_{Y,0}(1, W) \equiv \frac{1}{2}\left(1 + \frac{3}{4}\cos(\pi U V)\right), \qquad Q_{Y,0}(0, W) \equiv \frac{1}{2}\left(1 + \frac{1}{2}\sin(3\pi U/V)\right).$$

The conditional means and associated blip function $q_{Y,0}$ are represented in the left plots of Figure 2 (to be found in [8] due to space constraints). We compute the numerical values of the following parameters: $\psi_0 \approx 0.6827$ (true parameter); $\mathrm{Var}_{P_{Q_0,g^{\mathrm{b}}}} D(Q_0, g^{\mathrm{b}})(O) \approx 0.1916^2$ (the variance under $P_{Q_0,g^{\mathrm{b}}}$ of the efficient influence curve of Ψ at $P_{Q_0,g^{\mathrm{b}}}$, *that is*, under Q_0 with equiprobability of being assigned $A = 1$ or $A = 0$); $\mathrm{Var}_{P_{Q_0,g_0}} D(Q_0, g_0)(O) \approx 0.1666^2$ (the variance under P_{Q_0,g_0} of the efficient influence curve of Ψ at P_{Q_0,g_0}, *i.e.*, under Q_0 and the approximation g_0 to the optimal TR r_0); and $\mathrm{Var}_{P_{Q_0,r_0}} D(Q_0, r_0)(O) \approx 0.1634^2$ (the variance under P_{Q_0,r_0} of the efficient influence curve of Ψ at P_{Q_0,r_0}, *that is*, under Q_0 and the optimal TR r_0).

The sequences $\{t_n\}_{n\geq 1}$ and $\{\xi_n\}_{n\geq 1}$ are chosen constant, with values $t_\infty = 10\%$ and $\xi_\infty = 1\%$, respectively. We choose $g^{\mathrm{ref}} = g^{\mathrm{b}}$ as reference. The targeting steps are performed when sample size is a multiple of 100, at least 200 and

no more than 1000, when sampling is stopped. At such a sample size n, the working model $\mathcal{Q}_{1,n}$ consists of functions $Q_{Y,\beta}$ mapping $\mathcal{A} \times \mathcal{W}$ to $[0,1]$ such that, for each $a \in \mathcal{A}$ and $v \in \{1,2,3\}$, $\operatorname{logit} Q_{Y,\beta}(a,(U,v))$ is a linear combination of $1, U, U^2, \ldots, U^{d_n}$ and $\mathbf{1}\{(l-1)/\ell_n \leq U < l/\ell_n\}$ $(1 \leq l \leq \ell_n)$ with $d_n = 3 + \lfloor n/500 \rfloor$ and $\ell_n = \lceil n/250 \rceil$. The resulting global parameter β belongs to $\mathbb{R}^{6(d_n+\ell_n+1)}$ (in particular, $\mathbb{R}^{60}$ at sample size $n = 1000$). Working model $\mathcal{Q}_{1,n}$ is fitted wrt $L = L^{\mathrm{kl}}$ using the `cv.glmnet` function from package `glmnet` [10], with weights given in (2.8) and the option `"lambda.min"`. This means imposing (data-adaptive) upper-bounds on the ℓ^1- and ℓ^2-norms of parameter β (via penalization), hence the search for a sparse optimal parameter β_n.

6.2. *Results*. We repeat $N = 1000$ times, independently, the procedure described in Section 2.2 and the construction of confidence intervals for $\psi_{r_n,0}$ and confidence lower-bounds for the empirical and counterfactual cumulative pseudo-regrets described in Section 5. Table 2 (to be found in [8] due to space constraints) reports four empirical summary measures computed across simulations for each parameter among $\psi_{r_n,0}$, ψ_0, $\mathscr{E}_n$ and $\mathscr{C}_n$. In rows [a], we have the empirical coverages. In rows [b] and [c], we have the p-values of the binomial tests of 95%-coverage at least or 94%-coverage at least (null hypotheses) against their one-sided alternatives. In rows [d], we have the mean values of the possibly data-adaptive parameters. In rows [e], we have the mean values of Σ_n (for $\psi_{r_n,0}$), mean values of $|\mathscr{E}_n - (n^{-1}\sum_{i=1}^n Y_i - \psi_n^* + \xi_\alpha\sqrt{\Sigma_n^{\mathscr{E}}/n})|/|\mathscr{E}_n|$ (for $\mathscr{E}_n$), mean values of $|\mathscr{C}_n - (n^{-1}\sum_{i=1}^n Y_i - \psi_n^* + \xi_\alpha\sqrt{\Sigma_n^{\mathscr{E}}/n})|/|\mathscr{C}_n|$ (for $\mathscr{C}_n$).

It appears that the empirical coverage of the confidence intervals for the data-adaptive parameter $\psi_{r_n,0}$ and the fixed parameter ψ_0 is very satisfying. Although 14 out of 18 empirical proportions of coverage lie below 95%, the simulation study does not reveal a coverage smaller than 94%, even without adjusting for multiple testing. For a sample size larger than 400, the simulation study does not reveal a coverage smaller than the nominal 95%, even without adjusting for multiple testing.

The asymptotic variance of ψ_n^* seems to stabilize below 0.1850^2. This is slightly smaller than $\mathrm{Var}_{P_{Q_0,g^{\mathrm{b}}}} D(Q_0, g^{\mathrm{b}})(O) \approx 0.1916^2$ $(1916/1850 \approx 1.04)$ and a little larger than $\mathrm{Var}_{P_{Q_0,g_0}} D(Q_0, g_0)(O) \approx 0.1666^2$ $(1850/1666 \approx 1.11)$. In theory, the asymptotic variance of ψ_n^* can converge to the variance $\mathrm{Var}_{P_{Q_0,g_0}} D(Q_0, g_0)(O)$ if Q_{Y,β_n} converges to $Q_{Y,0}$. Rigorously speaking, this cannot be the case here given the working models we rely on. This is nonetheless a quite satisfying finding: we estimate $\psi_{r_n,0}$ and ψ_0 more efficiently than if we had achieved their efficient estimation based on i.i.d. data sampled under Q_0 and the balanced TR g^{b} and, in addition, do so in such a way that most patients [those for whom $r_n(W) = r_0(W)$] are much more likely (90% versus 50%) to be assigned their respective optimal treatments.

The empirical coverage provided by the lower confidence bounds on the data-adaptive parameters $\mathscr{E}_n$ and $\mathscr{C}_n$ is excellent. Actually, the empirical proportions of coverage for $\mathscr{E}_n$, all larger than 96.5%, suggest that either $\mathscr{E}_n$ or the asymptotic variance of its estimator is slightly overestimated (or both are). Naturally, there is no evidence whatsoever of an effective coverage smaller than 95% for $\mathscr{E}_n$. The empirical proportions of coverage for $\mathscr{C}_n$, all larger than 98.9% and often equal to 100%, illustrate the fact that the lower confidence bounds are conservative by construction.

Finally, the mean values of $|\mathscr{E}_n - (n^{-1}\sum_{i=1}^n Y_i - \psi_n^* + \xi_\alpha\sqrt{\Sigma_n^{\mathscr{E}}/n})|/|\mathscr{E}_n|$ and $|\mathscr{C}_n - (n^{-1}\sum_{i=1}^n Y_i - \psi_n^* + \xi_\alpha\sqrt{\Sigma_n^{\mathscr{E}}/n})|/|\mathscr{C}_n|$ quickly stabilize around 1.30. They quantify how close the lower confidence bounds are to the parameters they lower bound, at the scale of the parameters themselves (which, by nature, are bound to get close to zero, if not to converge to it).

Comparison with a traditional randomized clinical trial. As per suggestion of a reviewer, we also conduct the same simulation study except for the fact that we now set $t_\infty = 50\%$, and thus simulate a traditional randomized clinical trial (RCT) with independent sampling from P_{Q_0,g^b}. Its results in terms of empirical coverage are better for $n \leq 400$ and as good as those reported in [8], Table 2, for $n \geq 500$ (not shown). This suggests that by not adapting the TRs, we reach more quickly the asymptotic regime. The two main and most interesting differences concern the variance of the estimators and values of the pseudo-regrets. The ratios at each sample size $n \geq 200$ of the mean value across the N simulations of Σ_n^2 under the targeted sequential design of Section 6.1 (see the fourth row of [8], Table 2) to its counterpart under i.i.d. sampling range between 89.70% ($n = 300$) and 94.26% ($n = 400$). These percentages should be compared with the ratio of true variances $0.1666^2/0.1916^2 \simeq 75.60\%$, which provide an asymptotic, loose lower bound (see Section 6.1 and the comment on the asymptotic variance of ψ_n^* in the present section). In words, one needs recruiting fewer patients under the targeted sequential design than under the traditional balanced RCT to reach a given precision in estimation. By design, but nevertheless remarkably, the gain in efficiency goes with a gain in empirical cumulated pseudo-regret: the ratios at each sample size $n \geq 200$ of the mean value across the N simulations of $\mathscr{E}_n$ under the targeted sequential design of Section 6.1 (see the 14th row of [8], Table 2) to its counterpart under i.i.d. sampling decrease from 65.25% ($n = 200$) to 22.54% ($n = 1000$). By adapting the TRs, more patients are assigned their optimal treatments under the targeted sequential design than under the traditional balanced RCT, hence the decrease in pseudo-regrets (the same holds for the counterfactual cumulated pseudo-regret).

6.3. *Illustration.* Figures 1 and 2 (the latter in [8] due to space constraints) illustrate the data-adaptive inference of the optimal TR, its mean reward and the related pseudo-regrets with a visual summary of one additional run of the procedure described in Sections 2.2 and 5. We see in the top plot of Figure 1 that each

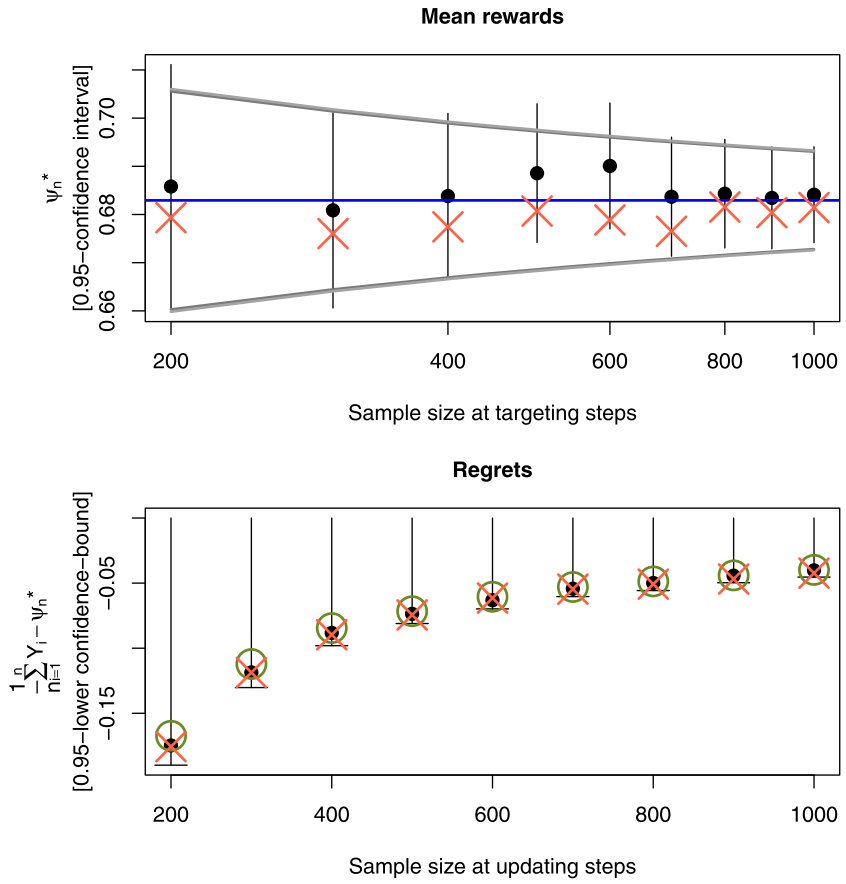

FIG. 1. *Illustrating the data-adaptive inference of the optimal treatment rule* (*TR*), *its mean reward and the related pseudo-regrets* (*see also Figure* 2). Top plot. *The blue horizontal line represents the value of the mean reward under the optimal TR*, ψ_0. *The grey curves represent the mapping* $n \mapsto \psi_0 \pm \xi_{97.5\%}\sigma_0/\sqrt{n}$, *where* $\sigma_0 = 0.1634$ *is the square root of* $\mathrm{Var}_{P_{Q_0,r_0}} D(Q_0, r_0)(O)$; *thus, at a given sample size* n, *the length of the vertical segment joining the two curves equals the length of a confidence interval based on a regular, asymptotically efficient estimator of* ψ_0. *The pink crosses represent the successive values of the data-adaptive parameters* $\psi_{r_n,0}$. *The black dots represent the successive values of* ψ_n^*, *and the vertical segments centered at them represent the successive* 95%*-confidence intervals for* $\psi_{r_n,0}$ *and, under additional assumptions, for* ψ_0 *as well*. Bottom plot. *The pink crosses and green circles represent the successive values of the empirical and counterfactual cumulative pseudo-regrets* $\mathscr{E}_n$ *and* $\mathscr{C}_n$. *The black dots represent the successive values of* $n^{-1}\sum_{i=1}^{n} Y_i - \psi_n^*$, *and the vertical segments represent the successive* 95%*-lower confidence bounds on* $\mathscr{E}_n$ *and* $\mathscr{C}_n$.

95%-confidence interval contains both its corresponding data-adaptive parameter $\psi_{r_n,0}$ and ψ_0. Moreover, the difference between the length of the 95%-confidence interval at sample size n and that of the vertical segment joining the two grey curves at this sample size gets smaller as n grows, showing that the variance of

ψ_n^* gets closer to the optimal variance $\mathrm{Var}_{P_{Q_0,r_0}} D(Q_0, r_0)(O)$. Finally, the bottom plot also reveals that the empirical and counterfactual cumulated pseudo-regrets $\mathscr{C}_n$ and $\mathscr{E}_n$ go to zero and that each 95%-lower confidence-bound is indeed below its corresponding pseudo-regrets.

7. Discussion. We develop a targeted, data-adaptive sampling scheme and TMLE estimator to build confidence intervals on the mean reward under the current estimate of the optimal TR and the optimal TR itself. As a by product, we also obtain lower confidence bounds on two cumulated pseudo-regrets. A simulation study illustrates the theoretical results. One of the cornerstones of the study is a new maximal inequality for martingales wrt the uniform entropy integral which allows the control of several empirical processes indexed by random functions.

We acknowledge that assuming $q_{Y,\beta_0} = q_{Y,0}$ in A2 is a stringent condition. This equality is mandatory only in the context of Proposition 5.2, where we give sufficient conditions guaranteeing that the TMLE ψ_n^* can be used to derive a confidence interval for ψ_0, the mean reward under the optimal TR r_0. In fact, we can strip out condition $q_{Y,\beta_0} = q_{Y,0}$ from A2, replace g_0 in A3 with $g_1 \in \mathcal{G}$ given by $g_1(1|W) \equiv G_\infty(q_{Y,\beta_0}(W))$ [compare with (3.1)], replace $q_{Y,0}$ in A5, A5*, A5** with q_{Y,β_0} and add that the ratio $|q_{Y,0}/q_{Y,\beta_0}|$ can be defined and has a finite (essential) supremum norm. Then, all our results still hold with the substitution of q_{Y,β_0} for $q_{Y,0}$, g_1 for g_0, TR $r_1 \equiv r(Q_{Y,\beta_0})$ for the optimal TR r_0, *and* that of $\psi_1 \equiv E_{Q_{0,r_1}}(Q_{Y,0}(A, W))$, the mean reward under TR r_1, for ψ_0.

Our analysis is asymptotic in the number of patients enrolled in the trial. We do not consider the issue of determining when the asymptotic regime is reached or when the trial should be stopped from either a theoretical or a numerical viewpoint. Devising a theoretical answer to this delicate question is certainly very challenging, if only because the complexity of $\mathcal{W}$, that of the true blip function $q_{Y,0}$ and related optimal TR r_0, and the choice of working models $\mathcal{Q}_n$ would play each its intricate role in the study. We should start by developing a group-sequential testing procedure [14] on top of the statistical analysis that we carry out, by following the same steps as in [6]. Indeed, a group-sequential testing procedure may end the trial at random stopping times based on data accrued so far, either because it is already possible to reject the null for its alternative at the pre-specified type I and II errors, or because there is no hope that that will be the case later if the trial continued.

We assume here that there is no stratum of the baseline covariates where treatment is neither beneficial nor harmful, *i.e.*, that nonexceptionality holds [25]. In future work, we will extend our result to handle exceptionality, building upon [19] (where observations are sampled independently). Extension to more than two treatments and to the inference of an optimal dynamic TR (where treatment assignment consists in successive assignments at successive time points) and its mean reward will also be considered.

Acknowledgements. The authors thank the Associate Editor and referees for their valuable suggestions. Antoine Chambaz thanks Henrik Bengtsson (UCSF) for his generous programming insight.

SUPPLEMENTARY MATERIAL

Technical lemmas, proofs, notation index (DOI: 10.1214/16-AOS1534SUPP; .pdf). The supplemental article contains the proofs of the results stated in the article, some technical lemmas used in the proofs, a notation index and a table and figure summarizing the results of the simulation study described in Section 6.

REFERENCES

[1] BALZER, L. B., PETERSEN, M. L., VAN DER LAAN, M. J. and THE SEARCH COLLABORATION (2016). Targeted estimation and inference for the sample average treatment effect in trials with and without pair-matching. *Stat. Med.* **35** 3717–3732. MR3538043

[2] BUBECK, S. and CESA-BIANCHI, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Found. Trends Mach. Learn.* **5** 1–122.

[3] CHAKRABORTY, B., LABER, E. B. and ZHAO, Y-Q. (2014). Inference about the expected performance of a data-driven dynamic treatment regime. *Clin. Trials* **11** 408–417.

[4] CHAKRABORTY, B. and MOODIE, E. E. M. (2013). *Statistical Methods for Dynamic Treatment Regimes*: *Reinforcement Learning*, *Causal Inference*, *and Personalized Medicine*. Springer, New York. MR3112454

[5] CHAMBAZ, A. (2016). tsml.cara.rct: Targeted sequential minimum loss CARA RCT design and inference. R package version 0.1.0.

[6] CHAMBAZ, A. and VAN DER LAAN, M. J. (2014). Inference in targeted group-sequential covariate-adjusted randomized clinical trials. *Scand. J. Stat.* **41** 104–140. MR3181135

[7] CHAMBAZ, A., VAN DER LAAN, M. J. and ZHENG, W. (2015). Targeted covariate-adjusted response-adaptive lasso-based randomized controlled trials. In *Modern Adaptive Randomized Clinical Trials*: *Statistical*, *Operational*, *and Regulatory Aspects* (O. Sverdlov, ed.) 345–368. CRC Press, Boca Raton, FL.

[8] CHAMBAZ, A., ZHENG, W. and VAN DER LAAN, M. J. (2017). Supplement to "Targeted sequential design for targeted learning inference of the optimal treatment rule and its mean reward." DOI:10.1214/16-AOS1534SUPP.

[9] DE LA PEÑA, V. H. and GINÉ, E. (1999). *Decoupling*: *From Dependence to Independence*. Springer, New York. MR1666908

[10] FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22.

[11] GARIVIER, A. and KAUFMANN, E. (2016). Optimal best arm identification with fixed confidence. *Proc. Mach. Learn. Res.* **49** 998–1027.

[12] GOLDBERG, Y., SONG, R., ZENG, D. and KOSOROK, M. R. (2014). Comment on "Dynamic treatment regimes: Technical challenges and applications". *Electron. J. Stat.* **8** 1290–1300.

[13] IBRAGIMOV, I. A. and HAS′MINSKIĬ, R. Z. (1981). *Statistical Estimation*: *Asymptotic Theory*. *Applications of Mathematics* **16**. Springer, New York–Berlin. Translated from the Russian by Samuel Kotz. MR0620321

[14] JENNISON, C. and TURNBULL, B. W. (2000). *Group Sequential Methods with Applications to Clinical Trials*. Chapman & Hall/CRC, Boca Raton, FL. MR1710781

[15] KAUFMANN, E. (2014). Analyse de stratégies bayésiennes et fréquentistes pour l'allocation séquentielle de ressources. Ph.D. thesis, TELECOM ParisTech. Available at http://chercheurs.lille.inria.fr/ekaufman/TheseEmilie.pdf.

[16] LABER, E. B., LIZOTTE, D. J., QIAN, M., PELHAM, W. E. and MURPHY, S. A. (2014). Dynamic treatment regimes: Technical challenges and applications. *Electron. J. Stat.* **8** 1225–1272.

[17] LABER, E. B., LIZOTTE, D. J., QIAN, M., PELHAM, W. E. and MURPHY, S. A. (2014). Rejoinder of "Dynamic treatment regimes: Technical challenges and applications" [MR3263119; MR3263120; MR3263121; MR3263122; MR3263118]. *Electron. J. Stat.* **8** 1312–1321. MR3263123

[18] LUEDTKE, A. R. and VAN DER LAAN, M. J. (2015). Targeted learning of the mean outcome under an optimal dynamic treatment rule. *J. Causal Inference* **3** 61–95.

[19] LUEDTKE, A. R. and VAN DER LAAN, M. J. (2016). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Ann. Statist.* **44** 713–742. MR3476615

[20] LUEDTKE, A. R. and VAN DER LAAN, M. J. (2016). Super-learning of an optimal dynamic treatment rule. *Int. J. Biostat.* **12** 305–332. MR3505699

[21] MAMMEN, E. and TSYBAKOV, A. B. (1999). Smooth discrimination analysis. *Ann. Statist.* **27** 1808–1829.

[22] PEARL, J. (2000). *Causality*: *Models*, *Reasoning and Inference*, *Vol.* 29. Cambridge Univ. Press, Cambridge.

[23] PFANZAGL, J. (1982). *Contributions to a General Asymptotic Statistical Theory*. *Lecture Notes in Statistics* **13**. Springer, New York–Berlin.

[24] QIAN, M. and MURPHY, S. A. (2011). Performance guarantees for individualized treatment rules. *Ann. Statist.* **39** 1180–1210.

[25] ROBINS, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In *Proc. Second Seattle Symp. Biostat* (D. Y. Lin and P. Heagerty, eds.) 189–326.

[26] ROBINS, J. M. and ROTNITZKY, A. (2014). Discussion of "Dynamic treatment regimes: Technical challenges and applications." *Electron. J. Stat.* **8** 1273–1289.

[27] RUBIN, D. B. and VAN DER LAAN, M. J. (2012). Statistical issues and limitations in personalized medicine research with clinical trials. *Int. J. Biostat.* **8** Article 1.

[28] STANLEY, K. (2007). Design of randomized controlled trials. *Circulation* **115** 1164–1169.

[29] R CORE TEAM (2016). *R*: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at https://www.R-project.org/.

[30] VAN DER LAAN, M. J. and ROBINS, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer, New York. MR1958123

[31] VAN DER LAAN, M. J. and ROSE, S. (2011). *Targeted Learning*: *Causal Inference for Observational and Experimental Data*. Springer, New York.

[32] VAN DER LAAN, M. J. and RUBIN, D. (2006). Targeted maximum likelihood learning. *Int. J. Biostat.* **2** Art. 11. MR2306500

[33] VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence*. Springer, Berlin.

[34] ZHANG, B., TSIATIS, A., DAVIDIAN, M., ZHANG, M. and LABER, E. (2012). A robust method for estimating optimal treatment regimes. *Biometrics* **68** 1010–1018.

[35] ZHANG, B., TSIATIS, A., DAVIDIAN, M., ZHANG, M. and LABER, E. (2012). Estimating optimal treatment regimes from a classification perspective. *Stat* **68** 103–114.

[36] ZHAO, Y., ZENG, D., LABER, E. B. and KOSOROK, M. R. (2015). New statistical learning methods for estimating optimal dynamic treatment regimes. *J. Amer. Statist. Assoc.* **110** 583–598.

[37] ZHAO, Y., ZENG, D., RUSH, A. J. and KOSOROK, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *J. Amer. Statist. Assoc.* **107** 1106–1118.

[38] ZHENG, W., CHAMBAZ, A. and VAN DER LAAN, M. J. (2015). Drawing valid targeted inference when covariate-adjusted response-adaptive RCT meets data-adaptive loss-based estimation, with an application to the LASSO. Technical Report 339, Univ. California, Berkeley, Division of Biostatistics Working Paper Series.

A. CHAMBAZ
MODAL'X
UPL, UNIVERSITÉ PARIS NANTERRE
F92000 NANTERRE
FRANCE
E-MAIL: achambaz@u-paris10.fr

W. ZHENG
M. J. VAN DER LAAN
DIVISION OF BIOSTATISTICS
AND
CENTER FOR TARGETED MACHINE
LEARNING AND CAUSAL INFERENCE
UNIVERSITY OF CALIFORNIA, BERKELEY
101 HAVILAND HALL
BERKELEY, CALIFORNIA 94720
USA
E-MAIL: wenjing.zheng@berkeley.edu
laan@berkeley.edu