University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2012 *Paper* 299

Targeted Learning of The Probability of Success of An In Vitro Fertilization Program Controlling for Time-dependent Confounders

Antoine Chambaz[*] Sherri Rose[†]
Jean Bouyer[‡] Mark J. van der Laan[**]

[*]Université Paris Descartes, achambaz@u-paris10.fr
[†]Johns Hopkins Bloomberg School of Public Health, sherrirosephd@gmail.com
[‡]Universite Paris-Sud
[**]University of California, Berkeley, laan@berkeley.edu

Targeted Learning of The Probability of Success of An In Vitro Fertilization Program Controlling for Time-dependent Confounders

Antoine Chambaz, Sherri Rose, Jean Bouyer, and Mark J. van der Laan

Abstract

<blockquote>Infertility is a global public health issue and various treatments are available. In vitro fertilization (IVF) is an increasingly common treatment method, but accurately assessing the success of IVF programs has proven challenging since they consist of multiple cycles. We present a double robust semiparametric method that incorporates machine learning to estimate the probability of success (i.e., delivery resulting from embryo transfer) of a program of at most four IVF cycles in the French Devenir Apr‘es Interruption de la FIV (DAIFI) study and several simulation studies, controlling for time-dependent confounders. We find that the probability of success in the DAIFI study is 50% (95% confidence interval [0.48, 0.53]), therefore approximately half of future participants in a program of at most four IVF cycles can expect a delivery resulting from embryo transfer.</blockquote>

(2012), **0**, 0, *pp.* 1–24

Targeted learning of the probability of success of an in vitro fertilization program controlling for time-dependent confounders

ANTOINE CHAMBAZ*[,1], SHERRI ROSE[2], JEAN BOUYER[3,4], MARK J. VAN DER LAAN[5]

[1] *Laboratoire MODAL'X, Université Paris-Ouest, 200 av. de la République, 92001 Nanterre, France*

[2] *Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 N Wolfe St, Baltimore, MD 21205, USA*

[3] *Inserm, CESP Centre for research in Epidemiology and Population Health, U1018, Reproduction and child development, F-94807, Villejuif, France*

[4] *Université Paris-Sud, UMRS 1018, F-94807, Villejuif, France*

[5] *Division of Biostatistics, University of California, Berkeley, School of Public Health, 101*

Haviland Hall, Berkeley, CA 94720, USA

*achambaz@u-paris10.fr

Summary

Infertility is a global public health issue and various treatments are available. In vitro fertilization (IVF) is an increasingly common treatment method, but accurately assessing the success of IVF programs has proven challenging since they consist of multiple cycles. We present a double robust semiparametric method that incorporates machine learning to estimate the probability of success (i.e., delivery resulting from embryo transfer) of a program of at most four IVF cycles in the French Devenir Après Interruption de la FIV (DAIFI) study and several simulation studies, controlling for time-dependent confounders. We find that the probability of success in the DAIFI study is 50% (95% confidence interval [0.48, 0.53]), therefore approximately half of future participants in

*To whom correspondence should be addressed.

a program of at most four IVF cycles can expect a delivery resulting from embryo transfer.

Key words: Confounding; Double robust estimation; In vitro fertilization; Longitudinal data.

1. Introduction

Infertility is a global concern among adults of childbearing age, one that has been labeled a public health issue by the World Health Organization. The current prevalence estimate of worldwide infertility is 9%, where infertility is defined as a couple failing to conceive within 12 months of attempting to become pregnant (Boivin *and others*, 2007). Infertility can be caused by various medical conditions, including endometriosis, damage to the fallopian tubes, polycystic ovary syndrome, abnormal sperm production, and epididymal obstruction, among others. Treatments for infertility vary based on age, gender, causes, access, and whether one or both members of the couple are affected.

In vitro fertilization (IVF) is an assisted reproductive technology that involves obtaining sperm and mature eggs from the couple (or donors), fertilizing the egg with the sperm in a laboratory environment, and then implanting the embryo(s) in the woman's uterus. The first IVF procedure resulting in a live birth occurred in 1978. IVF has become increasingly common in recent years, with over 63,000 procedures performed each year in the United States and 40,000 performed yearly in France (Adamson *and others*, 2006). Researchers have struggled to effectively quantify the success of IVF programs, and considerable debate remains. Simply calculating the number of pregnancies, deliveries, or live births per IVF procedure is not necessarily a complete summary of success, as many IVF programs consist of multiple consecutive cycles of implantation (a cycle of implantation, or simply a cycle, is a series of transfers of fresh or frozen embryos that were collected during a single oocyte pick-up). Thus, an alternative is to evaluate the entire program.

The French Devenir Après Interruption de la FIV (DAIFI) study (Soullier *and others*, 2008;

de la Rochebrochard *and others*, 2008, 2009) is a convenient choice to examine complete IVF programs, as the first four IVF cycles are entirely reimbursable under France's national health insurance system, leading to fewer dropouts due to financial reasons. In a previous article, the authors Soullier *and others* (2008) estimated the probability of success in the DAIFI study with three methods. First, they used a naive ratio of the number of deliveries following the initial IVF cycle over the total number of women enrolled (point estimate 37%, 95% confidence interval $[0.35, 0.38]$), but this method ignores dropouts. The second method was a nonparametric Kaplan–Meier survival analysis (point estimate 52%, 95% confidence interval $[0.49, 0.55]$) based on the assumption that women who abandoned the program mid-course had the same characteristics as women who did not. In particular, this second method ignored the potentially informative baseline covariates. Their final approach was multiple imputation methodology (Schafer, 1997; Little and Rubin, 2002), which is based on iteratively estimating missing data using the past (point estimate 46%, 95% confidence interval $[0.44, 0.48]$), and relies on parametric models potentially leading to bias. As one can clearly see, none of the confidence intervals overlap, and the three estimation methods in this single data set yield differing results.

A later analysis of the same DAIFI study data (Chambaz, 2011) followed the semiparametric targeted minimum loss-based estimator (TMLE) framework to develop an estimator for this problem, although it ignored the time-dependent confounding of the number of embryos (frozen or transferred) at each cycle (point estimate 51%, 95% confidence interval $[0.48, 0.53]$). This estimate was closest to the Kaplan–Meier survival analysis from the Soullier *and others* (2008) article, and the methods used by Chambaz (2011) can also be formulated in terms of a survival analysis, details of which are described in the article. TMLE is a general framework for approaching estimation problems in semiparametric models, and following this template provides estimators that have desirable statistical properties, including double robustness and efficiency (van der Laan and Rubin, 2006; van der Laan and Rose, 2011). It also allows for the incorporation of machine learning methods. Previous articles developing TMLEs in longitudinal data structures include (van der Laan, 2010; Rosenblum and van der Laan, 2010; Stitelman *and others*, 2011; Gruber and van der Laan, 2012; Schnitzer *and others*, 2012). Estimator comparisons have been presented by van der Laan and Rose (2011), among other works. Estimating-equation-based methods that are double robust (van der Laan and Robins, 2003) could also be implemented, although TMLE is generally more computationally feasible in longitudinal data while also being a substitution estimator.

Evaluating the entire IVF program therefore presents difficult statistical problems, including handling time-dependent confounding. We will address these statistical issues in this article, and estimate the probability of success (i.e., delivery resulting from embryo transfer) of a program of at most four IVF cycles in the DAIFI study using a double robust semiparametric TMLE. Unlike previous analyses, we will control for the time-dependent confounding of the number of embryos frozen or transferred at each cycle. Considering the financial, emotional, and physical burden of undergoing IVF treatment and the overall global prevalence of infertility, providing accurate measures of the success of IVF programs is an important public health question.

This article is presented with the following structure. The DAIFI study data is described in Section 2. In Section 3, we define the statistical parameter of interest and introduce important notation and assumptions. Section 4 describes our estimation strategy using TMLE. Simulation results demonstrating the utility of the TMLE procedure can be found in Section 5. The DAIFI study is analyzed in Section 6. A discussion closes the article in Section 7. Proofs and additional details of the estimation procedure, simulation protocol, and super learner implementation are presented in the Supplementary Material available at *Biostatistics* online.

2. Data description

The data consist of all women under age 42 who completed at least one IVF cycle at one of two French IVF units in Paris and Clermont-Ferrand between 1998 and 2002. Women aged 42 and over were included only if they had a normal ovarian reserve and IVF was indicated given their cause of infertility. Covariates collected include date of birth and IVF unit, as well as start date, number of oocytes harvested, number of embryos transferred or frozen, indicators of pregnancy, and successful delivery at each IVF cycle (for a detailed description, see de la Rochebrochard *and others*, 2009). Women are censored after the fourth IVF cycle.

3. Defining the statistical parameter of interest

We use throughout this article the notation $u_{i:j} = (u_i, \ldots, u_j) \in \mathbb{R}^{j-i+1}$, with convention $u_{i:j} = \emptyset$ whenever $i > j$. In particular, $1_{0:j} = (1, \ldots, 1) \in \mathbb{R}^{j+1}$ for all $j \geqslant 0$.

3.1 *Data, model, and parameter*

The observed data O has a longitudinal structure:

$$O = (W_1, W_2, C_0, L_0, A_0, C_1, L_1, A_1, C_2, L_2, A_2, C_3, L_3) \sim P_0,$$

where $W_1 \in \{0, 1\}$ indicates the IVF unit, the integer W_2 is age at first IVF cycle, $W_3 \equiv C_0 \in \mathcal{C} = \{0, \ldots, K\}$ is the number of embryos (frozen or transferred) at the first IVF cycle, $W_4 \equiv L_0 \in \{0, 1\}$ indicates whether the first IVF cycle was successful ($L_0 = 1$) or not ($L_0 = 0$). For all $1 \leqslant j \leqslant 3$, $A_{j-1} \in \{0, 1\}$ indicates whether a $(j+1)$th IVF cycle was attempted ($A_{j-1} = 1$) or not ($A_{j-1} = 0$, which also encodes dropout), $C_j \in \mathcal{C}$ is the number of embryos (frozen or transferred) at $(j + 1)$th IVF cycle, and $L_j \in \{0, 1\}$ indicates whether the $(j + 1)$th IVF cycle was successful ($L_j = 1$) or not ($L_j = 0$). The component $L_3 \equiv Y$ is the final outcome of interest. By convention, if $A_{j-1} = 0$ for some $1 \leqslant j \leqslant 3$, then $A_{j'-1} = C_{j'} = L_{j'} = 0$ for all $j \leqslant j' \leqslant 3$, and if $L_j = 1$ for some $0 \leqslant j < 3$, then $A_{j'} = C_{j'} = L_{j'} = 1$ for all $j \leqslant j' \leqslant 3$. Thus, the IVF program was successful overall if and only if $L_3 = 1$. Finally, we denote $W = W_{1:4}$: this notation conveys the notion that since C_0 and L_0 are always observed, they can be considered as baseline covariates.

The statistical parameter of interest is

$$\begin{aligned}\Psi(P_0) = E_{P_0}\Big(\sum_{\ell_{1:2}\in\{0,1\}^2, c_{1:3}\in\mathcal{C}^3} & P_0(Y = 1|C_{1:3} = c_{1:3}, A_{0:2} = 1_{0:2}, L_{1:2} = \ell_{1:2}, W) \\ & \times P_0(C_3 = c_3|C_{1:2} = c_{1:2}, A_{0:2} = 1_{0:2}, L_{1:2} = \ell_{1:2}, W) \\ & \times P_0(L_2 = \ell_2|C_{1:2} = c_{1:2}, A_{0:1} = 1_{0:1}, L_1 = \ell_1, W) \\ & \times P_0(C_2 = c_2|C_1 = c_1, A_{0:1} = 1_{0:1}, L_1 = \ell_1, W) \\ & \times P_0(L_1 = \ell_1|C_1 = c_1, A_0 = 1, W) \\ & \times P_0(C_1 = c_1|A_0 = 1, W)\Big). \end{aligned} \tag{3.1}$$

This actually defines, by substituting P for P_0 in (3.1), a mapping Ψ of the set $\mathcal{M}$ of all candidate data-generating distributions P compatible with the definition of O (including the truth, P_0) onto $\mathbb{R}$. The characterization of $\mathcal{M}$ includes the following positivity assumption: for all $P \in \mathcal{M}$ and each $0 \leqslant j \leqslant 2$, $0 < P(A_j = 1|L_{1:j}, C_{1:j}, A_{0:j-1} = 1_{0:j-1}, W)$. This assumption states that for each $0 \leqslant j \leqslant 2$, conditional on observing a woman who already went through $(j+1)$ unsuccessful IVF cycles, it cannot be certain, based on past information $(L_{1:j}, C_{1:j}, W)$, that a $(j+2)$th IVF cycle will not be attempted. We emphasize that this assumption can be tested from the data.

Interestingly, some factors involved in the definition of Ψ can be slightly simplified. Indeed, it holds almost surely that for every $P \in \mathcal{M}$ and $1 \leqslant j \leqslant 3$,

$$P(L_j = 1|C_{1:j}, A_{0:j-1}, L_{1:j-1}, W) = P(L_j = 1|C_{1:j}, L_{1:j-1}, W), \tag{3.2}$$

or, in other words, that the dependency of $P(L_j = 1|C_{1:j}, A_{0:j-1}, L_{1:j-1}, W)$ upon $A_{0:j-1}$ is entirely conveyed through $(C_{1:j}, L_{1:j-1}, W)$. We refer the interested reader to Appendix D in the Supplementary Material for the simple proof of (3.2). Consequently, the parameter of interest is also defined by: for all $P \in \mathcal{M}$,

$$\begin{aligned}\Psi(P) = E_P\Big(\sum_{\ell_{1:2}\in\{0,1\}^2, c_{1:3}\in\mathcal{C}^3} & P(Y=1|C_{1:3}=c_{1:3}, L_{1:2}=\ell_{1:2}, W) \\ & \times P(C_3=c_3|C_{1:2}=c_{1:2}, A_{0:2}=1_{0:2}, L_{1:2}=\ell_{1:2}, W) \\ & \times P(L_2=\ell_2|C_{1:2}=c_{1:2}, L_1=\ell_1, W) \\ & \times P(C_2=c_2|C_1=c_1, A_{0:1}=1_{0:1}, L_1=\ell_1, W) \\ & \times P(L_1=\ell_1|C_1=c_1, W) \\ & \times P(C_1=c_1|A_0=1, W)\Big).\end{aligned} \tag{3.3}$$

3.2 *Causal interpretation of the parameter of interest*

It is possible, at the cost of making untestable assumptions, to provide a causal interpretation of $\Psi(P)$. For this purpose, let us assume, for example, that the random phenomenon of interest obeys the following system of structural equations. There exist 13 independent random variables:

$$(U_W^1, U_W^2, U_{C_0}, U_{L_0}, U_{A_0}, U_{C_1}, U_{L_1}, \ldots, U_{A_2}, U_{C_3}, U_{L_3})$$

and 13 deterministic functions:

$$(f_W^1, f_W^2, f_{C_0}, f_{L_0}, f_{A_0}, f_{C_1}, f_{L_1}, \ldots, f_{A_2}, f_{C_3}, f_{L_3})$$

such that

$$\begin{cases} W_1 = f_W^1(U_W^1) \\ W_2 = f_W^2(W_1, U_W^2) \\ C_0 = f_{C_0}(W_2, W_1, U_{C_0}) \\ L_0 = f_{L_0}(C_0, W_2, W_1, U_{L_0}) \\ \quad \text{and for every } 1 \leqslant j \leqslant 3, \\ A_{j-1} = f_{A_{j-1}}(L_{1:j-1}, C_{1:j-1}, A_{0:j-2}, W, U_{A_{j-1}}) \\ C_j = f_{C_j}(A_{0:j-1}, L_{1:j-1}, C_{1,j-1}, W, U_{C_j}) \\ L_j = f_{L_j}(C_{1:j}, A_{0:j-1}, L_{1:j-1}, W, U_{L_j}). \end{cases}$$

The assumption of mutual independence of the sources of randomness is equivalent to assuming that there are no unmeasured confounders (the mutual independence could be slightly relaxed). It is also equivalent to assuming a causal graph whose nodes are the components of the observed data structure O, each node being the target of arrows pointing from every component of O preceding that node (chronologically).

One can intervene upon that system of structural equations by substituting the equality $A_{j-1} = 1$ to $A_{j-1} = f_{A_{j-1}}(L_{1:j-1}, C_{1:j-1}, A_{0:j-2}, W, U_{A_{j-1}})$ for each $1 \leqslant j \leqslant 3$. The intervened system describes how L_3 is randomly generated when a woman is *set* to undergo the whole IVF program. This last (chronologically speaking) random variable is denoted by $Y_{(1,1,1)}$ rather than Y in order to emphasize that it differs (in distribution) from Y. The outcome $Y_{(1,1,1)}$ is called the conterfactual outcome under the intervention $A_{0:2} = 1_{0:2}$. It is known that if $O \sim P_0$ then

$$\Pr_{P_0}(Y_{(1,1,1)} = 1) = \Psi(P_0).$$

This is an example of the G-computation formula (Robins, 1986, 1987; Pearl, 2000). In this situation, $\Psi(P_0)$ evaluates the causal effect of undergoing the whole IVF program in terms of pregnancy.

Other causal interpretations could be developed, based on the Neyman-Rubin causal model and potential outcomes thus involving the consistency and randomization assumptions (Rubin, 1974; Holland, 1986; Sekhon *and others*, 2011), or the explicit construction of counterfactuals (Gill and Robins, 2001; Yu and van der Laan, 2002) (note that the resulting causal interpretation would not hold in the real world). Finally, we emphasize that even if one is not willing to rely on the causal assumptions discussed, then the target parameter $\Psi(P_0)$ still represents an effect of interest (not necessarily a causal one, but at least one that is produced by intervening on the distribution of the data), which aims at getting as close as possible to a causal effect as the data allow.

4. Targeted minimum loss-based estimation

Suppose that we observe n independent copies $O^{(1)}, \ldots, O^{(n)}$ of the observed data structure $O \sim P_0$. We denote P_n as the empirical measure.

4.1 *Pathwise differentiability and the efficient influence curve*

We can say that Ψ is pathwise differentiable (i.e., identifiable and smooth enough to allow central-limit-theorem-based inference) at any $P \in \mathcal{M}$ with respect to the maximal tangent space $L_0^2(P)$. This means that there exists a unique $D^\star(P) \in L_0^2(P)$, called the efficient influence curve of Ψ at P, such that if $\{P(\varepsilon) : \varepsilon \in \mathbb{R}\} \subset \mathcal{M}$ is a fluctuation of P in direction $s \in L_0^2(P)$, then $\frac{\partial}{\partial \varepsilon}\Psi(P(\varepsilon))|_{\varepsilon=0} = E_P[D^\star(P)(O)s(O)]$.

The efficient influence curve $D^\star$ tells us a lot about the nature of Ψ from a statistical viewpoint, which is why we devote effort to elaborating it in some detail in this section. In particular, the efficient influence curve serves as a benchmark for the estimation of $\Psi(P_0)$ expressed in terms of statistical efficiency. Indeed, it is known that the asymptotic variance of all (regular) estimators of $\Psi(P_0)$ has a lower bound given by the variance $\mathrm{Var}_{P_0} D^\star(P_0)(O)$ of the efficient influence curve at P_0. This is a consequence of the convolution theorem (van der Vaart, 1998, Theorem 25.20). The efficient influence curve also proves useful as a tool due to a robustness property discussed at the end of this subsection.

Recall that $\{P(\varepsilon) : \varepsilon \in \mathbb{R}\} \subset \mathcal{M}$ is a fluctuation of P in direction $s \in L_0^2(P)$. For instance, if $\{P(\varepsilon) : \varepsilon \in \mathbb{R}\}$ is a one-dimensional parametric model such that $P(\varepsilon) = P$ at $\varepsilon = 0$ and

$\frac{\partial}{\partial \varepsilon} \log P(\varepsilon)(O)|_{\varepsilon=0} = s(O)$. Here it is implicitly understood that this statement is relative to the (minus log-likelihood) loss function L characterized by $L(P) : O \mapsto L(P)(O) = -\log P(O)$ for all $P \in \mathcal{M}$.

For the purpose of describing $D^\star(P)$, let us introduce the binary description $(C_{j,1}, \ldots, C_{j,K})$ of C_j for $1 \leqslant j \leqslant 3$ (formally characterized by $C_{j,k} = \mathbf{1}\{C_j = k\}$ for every $1 \leqslant k < K$ and $C_{j,K} = \mathbf{1}\{C_j \geqslant K\}$), and the following sets of parents: for all $1 \leqslant j \leqslant 3$ and $1 \leqslant k \leqslant K$,

$$\begin{aligned}
\mathrm{Pa}(C_{j,k}) &= \{C_{j,1:k-1}, A_{0:j-1}, L_{1:j-1}, C_{1:j-1}, W\},\\
\mathrm{Pa}(L_j) &= \{C_{1:j}, A_{0:j-1}, L_{1:j-1}, W\},\\
\mathrm{Pa}(A_{j-1}) &= \{L_{1:j-1}, C_{1:j-1}, A_{0:j-2}, W\}.
\end{aligned}$$

Note that it is possible to exclude $A_{0:j-1}$ from $\mathrm{Pa}(L_j)$, a side consequence of the proof of (3.2). Similarly, it is possible to exclude $A_{0:j-2}$ from $\mathrm{Pa}(A_{j-1})$. Our description of $D^\star(P)$ consists of expressing it as a linear projection of a function $D(P)$ characterized by

$$D(P)(O) = (Y - \Psi(P)) \frac{\mathbf{1}\{A_{0:2} = 1_{0:2}\}}{G_2(P)(O)},$$

where, for every $0 \leqslant J \leqslant 2$,

$$\begin{aligned}
G_J(P)(O) &= \prod_{j=0}^{J} g_j(P)(O), \quad \text{with}\\
g_j(P)(O) &= P(A_j = 1|\mathrm{Pa}(A_j)), \quad \text{for each } 0 \leqslant j \leqslant 2.
\end{aligned}$$

We denote $g(P) = (g_j(P) : 0 \leqslant j \leqslant 2)$. This interpretation of $D^\star(P)$ as a projection (justified, for example in Appendix A, Section 7 of van der Laan and Rose, 2011) is summarized by the following relation:

$$D^\star(P) = \Pi(D(P)|\mathcal{T}(Q_W)) + \sum_{j=1}^{3} \sum_{k=1}^{K} \Pi(D(P)|\mathcal{T}(Q_{C_{j,k}})) + \sum_{j=1}^{3} \Pi(D(P)|\mathcal{T}(Q_{L_j})), \tag{4.4}$$

where $\Pi(D(P)|\mathcal{T}(Q_W))(O) = E_P[D(P)(O)|W]$, and for all $1 \leqslant j \leqslant 3$ and $1 \leqslant k \leqslant K$,

$$\begin{aligned}
\Pi(D(P)|\mathcal{T}(Q_{C_{j,k}}))(O) &= E_P[D(P)(O)|C_{j,k}, \mathrm{Pa}(C_{j,k})] - E_P[D(P)(O)|\mathrm{Pa}(C_{j,k})],\\
\Pi(D(P)|\mathcal{T}(Q_{L_j}))(O) &= E_P[D(P)(O)|L_j, \mathrm{Pa}(L_j)] - E_P[D(P)(O)|\mathrm{Pa}(L_j)].
\end{aligned}$$

In the above equations, $\Pi(h|\mathcal{T})$ denotes the projection of $h \in L_0^2(P)$ onto $\mathcal{T}$, $\mathcal{T}$ being one of $\mathcal{T}(Q_W) = \{h \in L_0^2(P) : h(O) = h(W), E_P[h(W)] = 0\}$, $\mathcal{T}(Q_{C_{j,k}}) = \{h \in L_0^2(P) : h(O) = h(C_{j,k}, \mathrm{Pa}(C_{j,k})), E_P[h(C_{j,k}, \mathrm{Pa}(C_{j,k}))|\mathrm{Pa}(C_{j,k})] = 0\}$ and $\mathcal{T}(Q_{L_j}) = \{h \in L_0^2(P) : h(O) = h(L_j, \mathrm{Pa}(L_j)), E_P[h(L_j, \mathrm{Pa}(L_j))|\mathrm{Pa}(L_j)] = 0\}$. Here Q_W, $Q_{C_{j,k}}$, and Q_{L_j} respectively stand for the marginal distribution of W, conditional distribution of $C_{j,k}$ given $\mathrm{Pa}(C_{j,k})$, and conditional distribution of L_j given $\mathrm{Pa}(L_j)$. We find it convenient to set $Q(P) = (Q_W(P), Q_{C_{j,k}}(P), Q_{L_j}(P) : 1 \leqslant j \leqslant 3, 1 \leqslant k \leqslant K)$, the collection of these distributions associated to $P \in \mathcal{M}$. In particular, the last term in (4.4) is characterized by

$$\Pi(D(P)|\mathcal{T}(Q_{L_3}))(O) = \Big(L_3 - P(L_3 = 1|\mathrm{Pa}(L_3))\Big)\frac{\mathbf{1}\{A_{0:2} = 1_{0:2}\}}{G_2(P)(O)}. \tag{4.5}$$

We now discuss in more detail the robustness property provided by the efficient influence curve. Let us first note that obviously, $\Psi(P) = \Psi(P_0)$ holds when $Q(P) = Q(P_0)$. Furthermore, (4.4) makes clear that $D^\star$ is robust in the sense of the following lemma:

Lemma 4.1 Set $P \in \mathcal{M}$. If $E_{P_0}[D^\star(P)(O)] = 0$, in which case we say that P solves the efficient influence curve equation, then $\Psi(P) = \Psi(P_0)$ when $g(P) = g(P_0)$ (implicitly: even if $Q(P) \neq Q(P_0)$).

Thus, we take advantage of the pathwise differentiability of Ψ and robustness of $D^\star$ in order to estimate $\Psi(P_0)$. More specifically, we build a substitution estimator $\Psi(P_n^*)$ of $\Psi(P_0)$ where the estimator P_n^* of P_0 is targeted toward $\Psi(P_0)$, i.e., constructed in such a way that we can exploit the robustness of $D^\star$.

4.2 *TMLE procedure*

The TMLE methodology is a two-step estimation procedure. First, we derive an initial estimator $P_n^0 \equiv (Q(P_n^0), g(P_n^0))$ of the entire data-generating distribution P_0. This yields an initial substitution estimator $\Psi(P_n^0)$ of $\Psi(P_0)$. Second, we fluctuate P_n^0 in the direction of $D^\star(P_n^0)$, bending P_n^0 into P_n^* and $\Psi(P_n^0)$ into the final substitution estimator $\Psi(P_n^*)$, the TMLE of $\Psi(P_0)$.

4.2.1 *First step.* Regarding the construction of the estimator $Q(P_n^0)$ of $Q(P_0)$, we estimate the true marginal distribution of W, $Q_W(P_0)$, by its empirical counterpart. In other words, we set $Q_W(P_n^0) = P_{n,W} \equiv \frac{1}{n}\sum_{i=1}^n \mathrm{Dirac}(W^{(i)})$. Note that the remaining components of $Q(P_0)$ are conditional distributions of binary random variables. Similarly, estimating $g(P_0)$ is equivalent to estimating three conditional distributions of binary random variables. Thus, for the purpose of completing the construction of P_n^0, one can simply carry out a series of logistic regressions. Or, one can alternatively rely on super learning, an estimation methodology based on the aggregation of several estimators into a single algorithm with the smallest cross-validated risk (van der Laan *and others*, 2007; van der Laan and Rose, 2011). Of course, P_n^0 is built in such a way that the encoding conventions of Section 3.1 are fully exploited. We also make sure that, for each $0 \leqslant j \leqslant 2$, $g_j(P_n^0)(O) \geqslant c > 0$ has a lower bound given by a (possibly small) positive constant, except when it equals 0 exactly due to the latter encoding convention. As discussed above, $\Psi(P_n^0)$ is a natural estimator of $\Psi(P_0)$ to consider once $Q(P_n^0)$ is derived. Moreover, it is a consistent estimator if $Q(P_n^0)$ consistently estimates $Q(P_0)$, but not necessarily otherwise. The second step of the procedure involves taking into account the information provided by $g(P_n^0)$ which, in light of Lemma 4.1, is overlooked by $\Psi(P_n^0)$.

4.2.2 *Second step.* This second step decomposes into a finite series of successive updates of P_n^k into P_n^{k+1} (starting from $k = 0$). Each update amounts to *(a)* building a fluctuation $\{P_n^k(\varepsilon) : \varepsilon \in \mathbb{R}\}$ of the current P_n^k in the direction of a certain component of the efficient influence curve $D^\star(P_n^k)$ then determining the optimal stretch ε_n^k along that fluctuation and setting $P_n^{k+1} = P_n^k(\varepsilon_n^k)$, and *(b)* preparing the next update. Note that targeting the direction of interest and determining the optimal stretch is understood relative to a loss function that we carefully choose at each update.

(We refer the interested reader to Appendix A in the Supplementary Material for the detailed presentation of this updating procedure. The exhaustive description presented therein makes it very easy to write down the corresponding algorithm and code.)

Eventually, this second step bends the initial estimator P_n^0 of P_0 into P_n^*. Its differences relative to P_n^0 stem from the fact that P_n^* targets $\Psi(P_0)$ whereas P_n^0 does not. This is notably reflected in the following lemma, whose proof is naturally encapsulated in the detailed presentation of the updating procedure that we provide in Appendix A of the Supplementary Material:

Lemma 4.2 It holds that $P_n D^\star(P_n^*) = 0$.

Because $P_n D^\star(P_n^*)$ estimates $E_{P_0}[D^\star(P_n^*)(O)]$, we say that P_n^* solves the *empirical* efficient influence curve equation in view of Lemma 4.1. Lemma 4.2 is the cornerstone of the study of the asymptotic behavior of the TMLE $\Psi(P_n^*)$.

4.2.3 *Asymptotics.* The theory of estimating equations (see Chapter 25 in van der Vaart, 1998; van der Laan and Robins, 2003, and references therein) and Lemma 4.2 pave the way to describing the asymptotic behavior of the TMLE $\Psi(P_n^*)$. We refer the interested to Appendix A and (van der Laan and Rose, 2011, Section 18) for a formal presentation. In summary, under a set of conditions often referred to as regularity conditions, the TMLE $\Psi(P_n^*)$ is a consistent estimator of the truth $\Psi(P_0)$ and it satisfies a central limit theorem. The latter set of conditions include the requirement that $Q(P_n^*)$ and $g(P_n^*)$ both must converge to some Q_1 and g_1 with either $Q_1 = Q(P_0)$ or $g_1 = g(P_0)$ representing the truth. It also includes the requirement that $D^\star(P_n^*)$ must belong to a P_0-Donsker class with probability tending to one, and that a second-order term involving a product of distances between $Q(P_n^*)$ and Q_1 on the one hand and $g(P_n^*)$ and g_1 on the other hand is $o_P(1/\sqrt{n})$. If both $Q_1 = Q(P_0)$ and $g_1 = g(P_0)$ then the TMLE is asymptotically efficient and its asymptotic variance is consistently estimated by $P_n D^\star(P_n^*)^2$, the estimated variance of the efficient influence curve at P_n^*. Furthermore, if $g(P_n^*)$ is a maximum-likelihood estimator of $g(P_0)$ based on a correctly specified parametric model, then $P_n D^\star(P_n^*)^2$ is a conservative estimator of the asymptotic variance of $\Psi(P_n^*)$.

5. Simulation study

We present a simulation study that aims to illustrate the TMLE methodology, evaluate its performance numerically, and investigate the theoretical properties discussed in the previous section. We generate a simulation scheme by specifying the system of structural equations presented in Section 3.2. Complete details of the simulation study data generation are presented in Appendix B of the Supplementary Material. We denote P^s as the resulting data-generating distribution. The true value of the parameter of interest $\Psi(P^s)$ can be estimated with great precision by Monte Carlo; using a simulated dataset of one million observations under the intervention $A_{0:2} = 1_{0:2}$ yields $\Psi(P^s) \approx 0.72$.

We repeat $B = 1000$ times the following steps: *(i)* simulate a dataset with sample size $n = 3000$ under P^s, and *(ii)* estimate $\Psi(P^s)$ with $\Psi(P_{n,b}^*)$, the bth TMLE based on this bth simulated dataset. We also keep track of the initial estimator $\Psi(P_{n,b}^0)$ based on the same dataset. We then repeat that entire procedure four times, by relying either on a small library or a large library with the super learner to estimate the initial density, and by relying either on a correctly specified maximum likelihood estimator for $g(P^s)$ or the super learner. The estimation of $Q(P^s)$ always relies on the super learning procedure. See Table 1 for results. It is worth emphasizing that we cannot guarantee that the regularity conditions required in Section 4.2.3 to derive the asymptotic behavior of the TMLE are met.

The specifics of the super learning procedure that we use are reported in Appendix C of the Supplementary Material. We only emphasize here that when we estimate a component $Q_{L_j}(P^s)$ for $j = 2, 3$, we actually regress L_j on the set $\mathrm{Pa}(L_j) \setminus \{C_{1:j-1}\}$ of its parents deprived of

$C_{1:j-1}$. Similarly, when we estimate the component $Q_{A_2}(P^s)$, we actually regress A_2 on the set $\mathrm{Pa}(A_2) \setminus \{C_1\}$ of its parents deprived of C_1, and when we estimate a component $Q_{C_{j,k}}(P^s)$ for some $2 \leqslant j \leqslant 3$ and $1 \leqslant k \leqslant K$, we actually regress $C_{j,k}$ on the set $\mathrm{Pa}(C_{j,k}) \setminus \{C_{j',k'} : 1 \leqslant j' < j, 1 \leqslant k' \leqslant K\}$. This considerably diminishes the computational burden. Furthermore, and in the same spirit, we summarize each C_j $(1 \leqslant j \leqslant 3)$ by the binary description $(C_{j,1}, C_{j,2}, C_{j,3}) = (\mathbf{1}\{C_j = 1\}, \mathbf{1}\{C_j = 2\}, \mathbf{1}\{C_j \geqslant 3\})$. Overall, this makes the super learning procedure less apt at estimating the various components of P^s.

Four important features arise from Table 1. First, and as expected, the TMLE performs better than the initial estimator in terms of bias and root mean-squared error (rMSE) in all situations. One reaches the same conclusion visually when one looks at Figure 1. Second, the initial estimator performs better when it relies on the small super learner library than the large one. This illustrates the fact that building a better estimator of P_0 (if one agrees on the superiority of the super learner with a larger library upon the super learner with a smaller library relative to the task of estimating P_0) does not necessarily translate into building a better estimator of $\Psi(P_0)$, which is precisely one of the motivations of the TMLE procedure. Third, and surprisingly, the TMLE performs equally well whether it relies on a correct g-specification or not in terms of bias and rMSE, for both super-learning libraries. However, the conclusions are very different when one considers the empirical coverage. This coverage is guaranteed by the confidence intervals based on the central limit theorem for the TMLE and the estimation of the asymptotic variance by the estimated variance of the efficient influence curve at P_n^*. Indeed, the TMLE performs better in terms of empirical coverage when it relies on the larger super learner library than on the smaller one, with better results when the g-specification is correct. Yet, with an empirical coverage of 89.4% relying on the large super learning library and incorrect g-specification, the TMLE does not achieve the required 95% coverage in the situation we care for the most. When the g-specification is correct, this might be a pure second order term issue that would have disappeared for larger sample size. Judging by Figure 1, the approximate normality seems satisfactory, and the issue is raised through the estimation of the asymptotic variance of the TMLE. Note that although $n = 3001$ seems quite a large sample size, the estimation of $g_3(P_0)$ only relies on 397 observations (see Table 3).

In order to address this issue, we propose the use of the bootstrap to estimate the asymptotic variance. This also enables us to consider bootstrapped confidence intervals. We emphasize that there is no theoretical guarantee that this should work. The bootstrap has also been used for the purpose of correcting flawed confidence intervals by Stitelman and van der Laan (2010), to compensate for a lack of normality. Other solutions can be thought of here (including the targeted estimation of the asymptotic variance), but they are beyond the scope of this article.

Therefore, we now repeat $B = 50$ times the following steps: *(i)* simulate a dataset of sample size $n = 3000$ under P^s, hence an empirical measure $P_{n,b}$, *(ii)* estimate $\Psi(P^s)$ with $\Psi(P^*_{n,b})$, the bth TMLE based on this bth simulated dataset, and *(iii)* resample independently $M = 100$ times a dataset of $n = 3000$ independent variables drawn from $P_{n,b}$ and estimate $\Psi(P^s)$ with $\Psi(P^*_{n,b,m})$, the mth TMLE based on this mth bootstrapped dataset. We keep track of all $\Psi(P^*_{n,b}), \Psi(P^*_{n,b,m})$ ($1 \leqslant b \leqslant B$ and $1 \leqslant m \leqslant M$). We repeat this entire procedure two times, by relying either on the same small or large libraries for the super learner as above. In both cases, $g(P^s)$ is estimated by super learning. For each $1 \leqslant b \leqslant B$, we compute the empirical standard deviation of $\{\Psi(P^*_{n,b,m}) : 1 \leqslant m \leqslant M\}$, from which we deduce a 95% confidence interval by relying on the central limit theorem. We also compute the confidence interval based on the 2.5th and 97.5th percentiles of $\{\Psi(P^*_{n,b,m}) : 1 \leqslant m \leqslant M\}$.

Two important features arise from Table 2. First, and in agreement with the previous simulation, the TMLE performs better in terms of coverage when it relies on the large super learner library than on the smaller one. This result is independent of the way confidence intervals are built (based on the central limit theorem with an asymptotic variance estimated by nonparametric bootstrap or on quantiles). The same conclusion holds when one considers the mean widths of the confidence intervals. Second, this simulation shows that it is more reliable coveragewise to build the confidence intervals based on the central limit theorem with an asymptotic variance estimated by nonparametric bootstrap than on quantiles. On average, the former method yields slightly wider confidence intervals than the latter method. Overall, only one of the four configurations (small super learning library and confidence intervals based on quantiles) empirically fails to guarantee the prescribed coverage. Indeed, the probability that a binomial random variable with parameter $(50, 95\%)$ be smaller than 43 equals approximately 1.2%.

6. DAIFI Study

We observe $n = 3001$ women and report in Table 3 the empirical probabilities of $A_j = 1$ (each $j = 0, 1, 2$), the empirical probabilities of $L_j = 1$ (each $j = 0, 1, 2, 3$), and the empirical conditional means of C_j given $A_{0:j-1} = 1_{0:j-1}, L_{j-1} = 0$ (each $j = 0, 1, 2, 3$). The latter conditional means correspond to the mean number of embryos transferred or frozen in women who have undergone j unsuccessful IVF cycles at the next attempt. They only slightly differ from each other.

We apply the TMLE methodology as described in Sections 4.2 and 5. Following the conclusions of our simulation study, we rely on the central limit theorem to construct a 95% confidence interval, with an asymptotic variance estimated by bootstrap ($M = 100$ iterations). The results (point estimate, confidence interval) do not depend on the choice of super learner library (small or large), nor on the choice of the number of binary indicators $C_{j,1}, \ldots, C_{j,K}$ used to summarize each C_j, $j = 1, 2, 3$ ($K \in \{3, \ldots, 8\}$). The point estimate equals $\Psi(P_n^*) = 0.50$ with $[0.48, 0.53]$ as 95% confidence interval. In conclusion, future participants in a program of at most four IVF cycles can be confidently informed that approximately half of them may subsequently succeed in having a child.

7. Discussion

Our goal was to develop an estimator to study probability of success during a course of at most four cycles of IVF in France, taking into account the time-dependent confounding of the number of embryos transferred at each cycle. For this, we developed and implemented a TMLE, which has desirable statistical properties discussed in this article, including double robustness. TMLE also allows the incorporation of machine learning methods or ensembling algorithms such as super learning for the estimation of the relevant components of the likelihood. Our simulation study explored the theoretical properties of the described TMLE, with noteworthy results. Firstly, using the estimated variance of the efficient influence curve does not provide the desired empirical coverage, and our simulation study suggests the use of the nonparametric bootstrap for variance estimation and confidence intervals based on the central limit theorem. Unsurprisingly, we also find that the TMLE has improved performance with regard to coverage when a larger super learner library is used. Finite sample theory shows that one should make the library as large as possible (van der Laan *and others*, 2007).

Our estimate of the success of a four cycle IVF program in the DAIFI study was equal to 50% (95% confidence interval $[0.48, 0.53]$). This result was similar to previous analyses that did not control for time-dependent confounding, namely Chambaz (2011) with a result of 51% (95% confidence interval [0.48, 0.53]) and the Kaplan–Meier survival analysis of Soullier *and others* (2008) with a point estimate of 52% (95% confidence interval $[0.49, 0.55]$). Biologically, the number of embryos transferred at each cycle should have an effect on its success. Therefore, among multiple explanations, it is possible that other measured variables, particularly in the analysis of Chambaz (2011), such as age or the number of embryos transferred at the first cycle, had larger impacts on the outcome and the additional adjustment for number of embryos resulted in a trivial difference in the estimate. While this result was unexpected, it could not have been anticipated a priori given our subject matter background, and the TMLE that adjusts for time-dependent confounders presented in this article is preferred to the one developed by Chambaz (2011). In other applications, adjusting for time-dependent confounders is likely to improve the accuracy of the estimates.

Our result that 50% of future program participants will succeed in having a child in the course of at most four IVF cycles is quite high. Consider the following: if the successive menstrual cycles of a woman were independent and stationary, then having a child in the course of at most four menstrual cycles would have a 50% chance to occur if the probability of succeeding at each cycle were equal to 16%. Furthermore, the average population fecundability (i.e., the probability that a woman who is neither pregnant nor in postpartum amenorrhea will conceive during a given menstrual cycle with unprotected intercourse) range between 17% and 30%, depending on the way it is computed (Leridon, 1977; Heckman and Walker, 1990).

8. Supplementary Material

The reader is referred to the online Supplementary Materials for technical appendices and details of the simulation scheme. Supplementary Material is available online at

Funding

This work was supported by the France-Berkeley Fund, the National Science Foundation [DMS-1103901 to S. Rose], and the National Institutes of Health [R01-AI074345-05 to M. J. van der Laan].

Acknowledgments

The authors wish to thank E. de la Rochebrochard (INED, INSERM, CESP, Université Paris-Sud, UMRS 1018), the Cochin and the Clermont-Ferrand IVF units for sharing the DAIFI dataset.

Conflict of Interest: None declared.

References

Adamson, G. D., de Mouzon, J., Lancaster, P., Nygren, K-G., Suulivan, E. and Zegers-Hochscild, F. (2006). World collaborative report on in vitro fertilization, 2000. *Fertility and Sterility* **85**(1586–1622).

Boivin, J., Bunting, L, Collins, J. A. and Nygren, K-G. (2007). International estimates of infertility prevalence and treatment-seeking: potential need and demand for infertility medical care. *Human Reproduction* **22**(1506–1512).

Chambaz, A. (2011). Probability of success of an in vitro fertilization program. In: *Targeted learning*, Springer Ser. Statist. New York: Springer, pp. 419–434.

de la Rochebrochard, E., Quelen, C., Peikrishvili, R., Guibert, J. and Bouyer, J. (2009). Long-term outcome of parenthood project during in vitro fertilization and after discontinuation of unscuccessful in vitro fertilization. *Fertility and Sterility* **92**, 149–156.

de la Rochebrochard, E., Soullier, N., Peikrishvili, R., Guibert, J. and Bouyer, J. (2008). High in vitro fertilization discontinuation rate in france. *International Journal of Gynaecology and Obstetrics* **103**, 74–75.

Gill, R. and Robins, J. M. (2001). Causal inference in complex longitudinal studies: continuous case. *Ann. Stat.* **29**(6), 1785–1811.

Gruber, S. and van der Laan, M.J. (2012). Targeted minimum loss based estimation of causal effects of multiple time point interventions. *Int J Biostat* **8**(1).

Heckman, J. J. and Walker, J. R. (1990). Estimating fecundability from data on waiting times to first conception. *Journal of the American Statistical Association* **85**(410), 283–295.

Holland, P. W. (1986). Statistics and causal inference (with discussion). *J. Am. Stat. Assoc.* **81**(396), 945–970.

Leridon, H. (1977). *Human Fertility: The Basic Components*. University of Chicago Press.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical analysis with missing data*, Second edition., Wiley Series in Probability and Statistics. Hoboken, NJ: Wiley-Interscience [John Wiley & Sons].

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge.

Robins, J. M. (1986). A new approach to causal inference in mortality studies with sustained exposure periods. Application to control of the healthy worker survivor effect. *Math. Mod.* **7**, 1393–1512.

Robins, J. M. (1987). Addendum to: "A new approach to causal inference in mortality studies with a sustained exposure period. Application to control of the healthy worker survivor effect" [Math. Mod. 7:1393–1512 (1986)]. *Comput. Math. Appl.* **14**(9-12), 923–945.

Rosenblum, M. and van der Laan, M.J. (2010). Simple examples of estimating causal effects using targeted maximum likelihood estimation. *Technical Report* 262, Division of Biostatistics, University of California, Berkeley.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *J. Educ. Psychol* **66**, 688–701.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*, Volume 72, Monographs on Statistics and Applied Probability. London: Chapman & Hall.

Schnitzer, ME, Moodie, E.E.M. and Platt, R.W. (2012). Targeted maximum likelihood estimation for marginal time-dependent treatment effects under density misspecification. *Biostatistics* **doi: 10.1093/biostatistics/kxs024**, In press.

Sekhon, J. S., Gruber, S., Porter, K. E. and van der Laan, M. J. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data*, Chapter Propensy-score-based estimators and C-TMLE. Springer Verlag, pp. 343–364.

Soullier, N., Bouyer, J., J-L., Pouly, Guibert, J. and de la Rochebrochard. (2008). Estimating the success of an in vitro fertilization programme using multiple imputation. *Human Reproduction* **23**, 187–192.

Stitelman, O. M., De Gruttola, V. and van der Laan, M. J. (2011). A general implementation of TMLE for longitudinal data applied to causal inference in survival analysis. *Technical Report* 281, Division of Biostatistics, University of California, Berkeley. to appear in Int J Biostat.

Stitelman, O. M. and van der Laan, M. J. (2010). Collaborative targeted maximum likelihood for time to event data. *Int J Biostat* **6**(1), Article 21.

van der Laan, M.J. and Rubin, Daniel B. (2006). Targeted maximum likelihood learning. *Int J Biostat* **2**(1), Article 11.

van der Laan, M. J. (2010). Targeted maximum likelihood based causal inference. I. *Int. J. Biostat.* **6**(2), Art. 2, 44.

van der Laan, M. J., Polley, E. C. and Hubbard, A. E. (2007). Super learner. *Stat. Appl. Genet. Mol. Biol.* **6**, Art. 25, 23 pp. (electronic).

van der Laan, M. J. and Robins, J. M. (2003). *Unified methods for censored longitudinal data and causality*, Springer Series in Statistics. New York: Springer-Verlag.

van der Laan, M. J. and Rose, S. (2011). *Targeted learning*, Springer Series in Statistics. New York: Springer. Causal inference for observational and experimental data.

van der Vaart, A. W. (1998). *Asymptotic statistics*, Volume 3, Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.

Yu, Z. and van der Laan, M. J. (2002). Construction of counterfactuals and the G-computation formula. *Technical Report* 122, Division of Biostatistics, University of California, Berkeley.

Table 1. Simulation results (1/2). The true value of the parameter of interest is $\Psi(P^s) \approx 0.72$. "rMSE" stands for root mean-squared error and "coverage" for the empirical coverage of the 95% confidence intervals based on the central limit theorem and the estimated variance of the efficient influence curve at P_n^*. All values are multiplied by 100.

	small super learner library				large super learner library			
	g correct		g incorrect		g correct		g incorrect	
estimator	init.	TMLE	init.	TMLE	init.	TMLE	init.	TMLE
bias	-1.80	0.42	-2.01	0.38	-2.58	-0.47	-2.40	-0.39
rMSE	2.23	1.57	2.36	1.58	2.62	1.31	2.47	1.30
coverage	-	85.2	-	81.0	-	91.8	-	89.4

Table 2. Simulation results (2/2). We report the empirical coverages and mean widths of the 95% confidence intervals (CIs) based on the central limit theorem and the estimation of the asymptotic variance by nonparametric bootstrap and those of the confidence interval based on the 2.5th and 97.5th percentiles of the bootstrapped TMLEs. All values are multiplied by 100.

	small super learner library		large super learner library	
CIs based on	CLT	quantiles	CLT	quantiles
coverage	96.0	86.0	100.0	92.0
mean width	6.79	6.39	5.37	5.06

[*Received August 1, 2010; revised October 1, 2010; accepted for publication November 1, 2010*]

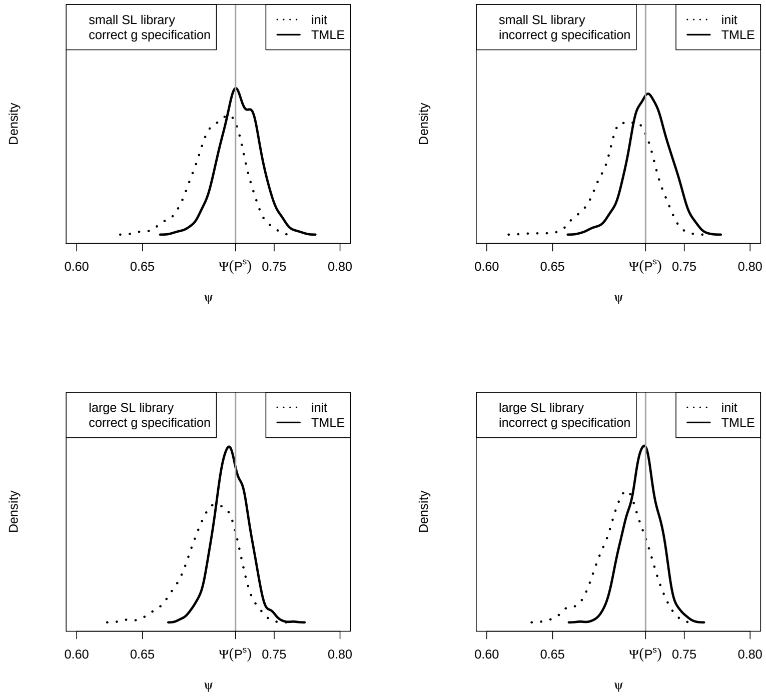

Fig. 1. Kernel estimates of the densities of the initial and targeted minimum loss estimators. In each graph, the vertical line indicates the true values of the parameter of interest $\Psi(P^s) \approx 0.72$. The first and second rows respectively correspond to the use of a small or large library for the super learning procedure. The first column corresponds to the estimation of $g(P^s)$ by maximum likelihood on a correctly specified parametric model, while the second column corresponds to the estimation of $g(P^s)$ by super learning.

Table 3. Empirical probabilities of $A_j = 1$ (each $j = 0, 1, 2$), empirical probabilities of $L_j = 1$ (each $j = 0, 1, 2, 3$), and empirical conditional means of C_j given $A_{0:j-1} = 1_{0:j-1}, L_{0:j-1} = 0_{0:j-1}$ (each $j = 0, 1, 2, 3$), as computed on the DAIFI data. In the fifth column, we report between parentheses the number of women involved in the computation of the corresponding empirical conditional mean.

	empirical probabilities of		empirical conditional mean of	# women
IVF cycle j	$A_j = 1$	$L_j = 1$	C_j given $A_{0:j-1} = 1_{0:j-1}, L_{0:j-1} = 0_{0:j-1}$	
0	75%	22%	3.31	(3001)
1	59%	32%	3.09	(1624)
2	49%	35%	2.95	(813)
3	-	37%	3.23	(397)

(2012), **0**, 0, *pp.* 1–15

Targeted learning of the probability of success of an in vitro fertilization program controlling for time-dependent confounders: Supplementary materials

ANTOINE CHAMBAZ[*,1], SHERRI ROSE[2], JEAN BOUYER[3,4], MARK J. VAN DER LAAN[5]

[1] *Laboratoire MODAL'X, Université Paris-Ouest, 200 av. de la République, 92001 Nanterre, France*

[2] *Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 N Wolfe St, Baltimore, MD 21205, USA*

[3] *Inserm, CESP Centre for research in Epidemiology and Population Health, U1018, Reproduction and child development, F-94807, Villejuif, France*

[4] *Université Paris-Sud, UMRS 1018, F-94807, Villejuif, France*

[5] *Division of Biostatistics, University of California, Berkeley, School of Public Health, 101 Haviland Hall, Berkeley, CA 94720, USA*

*antoine.chambaz@u-paris10.fr

APPENDIX

A. TMLE Updating Procedure in Detail

The TMLE updating procedure involves computing the terms of the right-hand side sum in (4.4). To do so, we actually start from the "last term", $\Pi(D(P)|\mathcal{T}_{Q_{L_3}})(O)$, see Section A.1, then derive the previous terms recursively going backwards, see Sections A.2 and A.3. The following almost exhaustive description makes it very easy to write down the corresponding algorithm and code.

[*] To whom correspondence should be addressed.

A.1 *First update*

For the sake of notational consistency, let us characterize Δ_3 by $\Delta_3(O) = L_3$. Remember that we know (at least numerically) the projection $\Pi(D(P_n^0)|\mathcal{T}(Q_{L_3}))$ (just substitute P_n^0 for P in (4.5)). The first update goes as follows. We define $H_n^0(O) = 1$ and characterize $P_n^0(\varepsilon)$ (for any $\varepsilon \in \mathbb{R}$) as being the data-generating distribution for O such that *(i)* $\mathrm{Pa}(L_3)$ has the same distribution under $P_n^0(\varepsilon)$ as under P_n^0, and *(ii)* the conditional distribution of L_3 given $\mathrm{Pa}(L_3)$ under $P_n^0(\varepsilon)$ is the Bernoulli law with parameter

$$\mathrm{expit}\,\left(\mathrm{logit}\, P_n^0(L_3 = 1|\mathrm{Pa}(L_3)) + \varepsilon H_n^0(O)\right)$$

(by convention, $\mathrm{logit}\,(0) = -\infty$, $\mathrm{logit}\,(1) = +\infty$, $\mathrm{expit}\,(-\infty) = 0$ and $\mathrm{expit}\,(+\infty) = 1$). One can easily check that $\{P_n^0(\varepsilon) : \varepsilon \in \mathbb{R}\}$ fluctuates P_n^0 (*i.e.*, $P_n^0(0) = P_n^0$) in the direction of $\Pi(D(P_n^0)|\mathcal{T}(Q_{L_3}))$ for the weighted log-likelihood loss function L_n^0 characterized by $L_n^0(P)(O) = -\frac{\mathbf{1}\{A_{0:2}=1_{0:2}\}}{G_2(P_n^0)(O)} \log P(O)$. Indeed,

$$\frac{\partial}{\partial\varepsilon} L_n^0(P_n^0(\varepsilon))(O)|_{\varepsilon=0} = -\Pi(D(P_n^0)|\mathcal{T}(Q_{L_3}))(O).$$

The optimal stretch ε_n^0 along that fluctuation of P_n^0 is that which minimizes the empirical loss, *i.e.*, $\varepsilon_n^0 = \arg\min_{\varepsilon\in\mathbb{R}} P_n L_n^0(P_n^0(\varepsilon))$. We conclude this first update by setting $P_n^1 = P_n^0(\varepsilon_n^0)$.

For the sake of preparing the next fluctuation, we derive (numerically) $E_{P_n^1}[\Delta_3(O)|\mathrm{Pa}(L_3)] = P_n^1(L_3 = 1|C_{3,K}, \mathrm{Pa}(C_{3,K}))$, from which we deduce

$$\begin{aligned} E_{P_n^1}[\Delta_3(O)|\mathrm{Pa}(C_{3,K})] &= E_{P_n^1}[E_{P_n^1}[\Delta_3(O)|\mathrm{Pa}(L_3)]|\mathrm{Pa}(C_{3,K})] \\ &= \sum_{c_3\in\{0,1\}} P_n^1(C_{3,K} = c_3|\mathrm{Pa}(C_{3,K}))P_n^1(L_3 = 1|C_{3,K} = c_3, \mathrm{Pa}(C_{3,K})), \end{aligned}$$

hence in turn the projection

$$H_n^1(O) \equiv \Pi(\Delta_3|\mathcal{T}(Q_{C_{3,K}}))(O) = E_{P_n^1}[\Delta_3(O)|C_{3,K}, \mathrm{Pa}(C_{3,K})] - E_{P_n^1}[\Delta_3(O)|\mathrm{Pa}(C_{3,K})],$$

which satisfies

$$\Pi(D(P_n^1)|\mathcal{T}(Q_{C_{3,K}}))(O) = H_n^1(O)\frac{\mathbf{1}\{A_{0:2} = 1_{0:2}\}}{G_2(P_n^1)(O)}.$$

The discussion of four important points are in order:

- this fluctuation obviously only fluctuates the conditional distribution of L_3 given $\mathrm{Pa}(L_3)$; it does so in the direction of the corresponding component of the current estimate of the efficient influence curve, $D^\star(P_n^0)$;
- for each $0 \leqslant j \leqslant 2$, if $\mathbf{1}\{A_{0:j} = 1_{0:j}\} = 1$ then $G_j(P_n^0)(O) > 0$; otherwise we say that the weight $\frac{\mathbf{1}\{A_{0:j}=1_{0:j}\}}{G_j(P_n^0)(O)}$ equals 0 by convention;
- deriving what P_n^1 is merely amounts to fitting a weighted logistic regression of L_3 on $H_n^0(O)$ with an offset equal to $\operatorname{logit} P_n^0(L_3 = 1|\mathrm{Pa}(L_3))$, using only those observations for which $A_{0:2}^{(i)} = 1_{0:2}$ and the corresponding weights $1/G_2(P_n^0)(O^{(i)})$;
- it holds that $P_n \Pi(D(P_n^1)|\mathcal{T}(Q_{L_3})) = 0$.

A.2 *Successive 2nd to* $(K+1)$*th updates*

We describe now the next K successive updates. Starting from $k = 1$, we characterize $P_n^k(\varepsilon)$ (for any $\varepsilon \in \mathbb{R}$) as being the data-generating distribution of O such that *(i)* $\mathrm{Pa}(C_{3,K+1-k})$ has the same distribution under $P_n^k(\varepsilon)$ as under P_n^k, *(ii)* the conditional distribution of $C_{3,K+1-k}$ given $\mathrm{Pa}(C_{3,K+1-k})$ is the Bernoulli law with parameter

$$\operatorname{expit}\left(\operatorname{logit} P_n^k(C_{3,K+1-k} = 1|\mathrm{Pa}(C_{3,K+1-k})) + \varepsilon H_n^k(O)\right),$$

and *(iii)* all remaining variables have the same conditional distributions given their parents under $P_n^k(\varepsilon)$ as under P_n^k. One can easily check that $\{P_n^k(\varepsilon) : \varepsilon \in \mathbb{R}\}$ fluctuates P_n^k (*i.e.*, $P_n^k(0) = P_n^k$) in the direction of $\Pi(D(P_n^k)|\mathcal{T}(Q_{C_{3,K+1-k}}))$ for the weighted log-likelihood loss function L_n^k characterized by $L_n^k(P)(O) = -\frac{\mathbf{1}\{A_{0:2}=1_{0:2}\}}{G_2(P_n^k)(O)} \log P(O)$. Indeed,

$$\frac{\partial}{\partial\varepsilon} L_n^k(P_n^k(\varepsilon))(O)|_{\varepsilon=0} = -\Pi(D(P_n^k)|\mathcal{T}(Q_{C_{3,K+1-k}}))(O).$$

The optimal stretch ε_n^k along that fluctuation of P_n^k is that which minimizes the empirical loss, *i.e.*, $\varepsilon_n^k = \arg\min_{\varepsilon\in\mathbb{R}} P_n L_n^k(P_n^k(\varepsilon))$, and we conclude this update by setting $P_n^{k+1} = P_n^k(\varepsilon_n^k)$.

For the sake of preparing the next fluctuation, we now derive $E_{P_n^{k+1}}[\Delta_3(O)|\mathrm{Pa}(C_{3,K+1-k})]$. If $k < K$ then we can deduce from there

$$E_{P_n^{k+1}}[\Delta_3(O)|\mathrm{Pa}(C_{3,K-k})] = E_{P_n^{k+1}}[E_{P_n^{k+1}}[\Delta_3(O)|\mathrm{Pa}(C_{3,K+1-k})]|\mathrm{Pa}(C_{3,K-k})],$$

hence in turn the projection $\Pi(\Delta_3|\mathcal{T}(Q_{C_{3,K-k}}))$, which can be written as

$$\Pi(\Delta_3|\mathcal{T}(Q_{C_{3,K-k}}))(O) = E_{P_n^{k+1}}[\Delta_3(O)|C_{3,K-k}, \mathrm{Pa}(C_{3,K-k})] - E_{P_n^{k+1}}[\Delta_3(O)|\mathrm{Pa}(C_{3,K-k})].$$

Then we set $H_n^{k+1}(O) = \Pi(\Delta_3|\mathcal{T}(Q_{C_{3,K-k}}))(O)$, which satisfies

$$\Pi(D(P_n^{k+1})|\mathcal{T}(Q_{C_{3,K-k}}))(O) = H_n^{k+1}(O)\frac{\mathbf{1}\{A_{0:2} = 1_{0:2}\}}{G_2(P_n^{k+1})(O)}.$$

Now, everything is in place to carry on and undertake the next fluctuation: if $k < K$ then we increment $k \leftarrow k + 1$ and repeat the procedure described in the current Section A.2, otherwise we proceed to Section A.3. Again four important points are in order:

- the latter kth fluctuation obviously only fluctuates the conditional distribution of $C_{3,K+1-k}$ given $\mathrm{Pa}(C_{3,K+1-k})$; it does so in the direction of the corresponding component of the current estimate of the efficient influence curve, $D^\star(P_n^k)$;
- we emphasize that $G_2(P_n^k) = G_2(P_n^0)$ so that there is no ambiguity in the definition of the weight $\frac{\mathbf{1}\{A_{0:2}=1_{0:2}\}}{G_2(P_n^k)(O)}$;
- deriving what is P_n^{k+1} from P_n^k merely amounts to fitting a weighted logistic regression of $C_{3,K+1-k}$ on $H_n^k(O)$ with an offset equal to $\mathrm{logit}\, P_n^k(C_{3,K+1-k} = 1|\mathrm{Pa}(C_{3,K+1-k}))$, using only those observations for which $A_{0:2}^{(i)} = 1_{0:2}$ and the corresponding weights $1/G_2(P_n^0)(O^{(i)})$;
- eventually, we have $P_n\Pi(D(P_n^{K+1})|\mathcal{T}(Q_{L_3})) = P_n\Pi(D(P_n^{K+1})|\mathcal{T}(Q_{C_{3,K+1-k}})) = 0$ for all $1 \leqslant k \leqslant K$.

A.3 *Successive* $(K+2)th$ *to* $3(K+1)th$ *updates*

The next $2(K+1)$ updates are very similar to the $(K+1)$ updates that we described in Sections A.1 and A.2, because we leave unchanged (i.e., do not fluctuate) the initial estimation of the conditional distribution of A_2 given $\mathrm{Pa}(A_2)$ and, later, of A_1 given $\mathrm{Pa}(A_1)$. In other words, all the data-generating distributions P_n^k that we consider below are such that $g(P_n^k) = g(P_n^0)$. Consequently, there will be no ambiguity in the forthcoming definitions of the weights $\frac{\mathbf{1}\{A_{0:j}=1_{0:j}\}}{G_j(P_n^k)(O)}$ for each $j = 0, 1$.

A.3.1 $(K+2)th$ *update.* We characterize $\Delta_2(P)$ for all $P \in \mathcal{M}$ by setting $\Delta_2(P)(O) = E_P[\Delta_3(O)|A_2 = 1, \mathrm{Pa}(A_2)]$ (the subscript '2' refers to the fact that $\Delta_2(P)(O)$ depends on O only through L_2 and $\mathrm{Pa}(L_2)$). We know $\Delta_2(P_n^{K+1})$ because we have derived earlier the conditional expectation $E_{P_n^{K+1}}[\Delta_3(O)|\mathrm{Pa}(C_{3,1}) \equiv (A_2, L_2, \mathrm{Pa}(L_2))]$. By the tower rule, it holds that

$$\begin{aligned} &E_{P_n^{K+1}}[D(P_n^{K+1})(O)|L_2, \mathrm{Pa}(L_2)] \\ &\quad = E_{P_n^{K+1}}\left(\frac{\mathbf{1}\{A_{0:2}=1_{0:2}\}}{G_2(P_n^{K+1})(O)} E_{P_n^{K+1}}\left[\Delta_3(O) - \Psi(P_n^{K+1}) \middle| A_2, L_2, \mathrm{Pa}(L_2)\right] \middle| L_2, \mathrm{Pa}(L_2)\right) \\ &\qquad\qquad = \left(\Delta_2(P_n^{K+1})(O) - \Psi(P_n^{K+1})\right) \frac{\mathbf{1}\{A_{0:1}=1_{0:1}\}}{G_1(P_n^{K+1})(O)}. \end{aligned} \tag{A.1}$$

Thus deriving

$$H_n^{K+1}(O) \equiv \Pi(\Delta_2(P_n^{K+1})|\mathcal{T}(Q_{L_2}))(O) = \Delta_2(P_n^{K+1})(O) - E_{P_n^{K+1}}[\Delta_2(P_n^{K+1})(O)|\mathrm{Pa}(L_2)]$$

yields the next component of the efficient influence curve, because

$$\Pi(D(P_n^{K+1})|\mathcal{T}(Q_{L_2}))(O) = H_n^{K+1}(O)\frac{\mathbf{1}\{A_{0:1}=1_{0:1}\}}{G_1(P_n^{K+1})(O)}.$$

Interestingly, it is not necessary to compute $\Psi(P_n^{K+1})$ to derive $\Pi(D(P_n^{K+1})|\mathcal{T}(Q_{L_2}))$, even though the parameter appears in (A.1).

Now we fluctuate the conditional distribution of L_2 given $\mathrm{Pa}(L_2)$ only, and do so in the direction of the corresponding component of the current estimate of the efficient influence curve,

$D^\star(P_n^{K+1})$. Similarly to Section A.1, we characterize $P_n^{K+1}(\varepsilon)$ (for any $\varepsilon \in \mathbb{R}$) as being the data-generating distribution for O such that *(i)* $\mathrm{Pa}(L_2)$ has the same distribution under $P_n^{K+1}(\varepsilon)$ as under P_n^{K+1}, *(ii)* the conditional distribution of L_2 given $\mathrm{Pa}(L_2)$ is the Bernoulli law with parameter

$$\operatorname{expit}\left(\operatorname{logit} P_n^{K+1}(L_2 = 1|\mathrm{Pa}(L_2)) + \varepsilon H_n^{K+1}(O)\right),$$

and *(iii)* all remaining variables have the same conditional distributions given their parents under $P_n^{K+1}(\varepsilon)$ as under P_n^{K+1}. One can easily check that $\{P_n^{K+1}(\varepsilon) : \varepsilon \in \mathbb{R}\}$ fluctuates P_n^{K+1} (*i.e.*, $P_n^{K+1}(0) = P_n^{K+1}$) in the direction of $\Pi(D(P_n^{K+1})|\mathcal{T}(Q_{L_2}))$ for the weighted log-likelihood loss function L_n^{K+1} characterized by $L_n^{K+1}(P)(O) = -\frac{\mathbf{1}\{A_{0:1}=1_{0:1}\}}{G_1(P_n^{K+1})(O)} \log P(O)$. Indeed,

$$\frac{\partial}{\partial\varepsilon} L_n^{K+1}(P_n^{K+1}(\varepsilon))(O)|_{\varepsilon=0} = -\Pi(D(P_n^{K+1})|\mathcal{T}(Q_{L_2}))(O).$$

The optimal stretch ε_n^{K+1} along that fluctuation of P_n^{K+1} is that which minimizes the empirical loss, *i.e.*, $\varepsilon_n^{K+1} = \arg\min_{\varepsilon\in\mathbb{R}} P_n L_n^{K+1}(P_n^{K+1}(\varepsilon))$, and we conclude this update by setting $P_n^{K+2} = P_n^{K+1}(\varepsilon_n^{K+1})$.

For the sake of preparing the next fluctuation, we derive $E_{P_n^{K+2}}[\Delta_2(P_n^{K+2})(O)|\mathrm{Pa}(L_2)]$, from which we deduce

$$E_{P_n^{K+2}}[\Delta_2(P_n^{K+2})(O)|\mathrm{Pa}(C_{2,K})] = E_{P_n^{K+2}}[E_{P_n^{K+2}}[\Delta_2(P_n^{K+2})(O)|\mathrm{Pa}(L_2)]|\mathrm{Pa}(C_{2,K})],$$

hence in turn the projection

$$\begin{aligned} H_n^{K+2}(O) &\equiv \Pi(\Delta_2(P_n^{K+2})|\mathcal{T}(Q_{C_{2,K}}))(O) \\ &= E_{P_n^{K+2}}[\Delta_2(P_n^{K+2})(O)|C_{2,K}, \mathrm{Pa}(C_{2,K})] - E_{P_n^{K+2}}[\Delta_2(P_n^{K+2})(O)|\mathrm{Pa}(C_{2,K})], \end{aligned}$$

which satisfies

$$\Pi(D(P_n^{K+2})|\mathcal{T}(Q_{C_{2,K}}))(O) = H_n^{K+2}(O)\frac{\mathbf{1}\{A_{0:1} = 1_{0:1}\}}{G_1(P_n^{K+2})(O)}.$$

We emphasize that:

- deriving P_n^{K+2} from P_n^{K+1} merely amounts to fitting a weighted logistic regression of L_2 on $H_n^{K+1}(O)$ with an offset equal to logit $P_n^{K+1}(L_2 = 1|\mathrm{Pa}(L_2))$, using only those observations for which $A_{0:1}^{(i)} = 1_{0:1}$ and the corresponding weights $1/G_1(P_n^0)(O^{(i)})$;

- it holds that $P_n\Pi(D(P_n^{K+2})|\mathcal{T}(Q_{L_j})) = P_n\Pi(D(P_n^{K+2})|\mathcal{T}(Q_{C_{3,K+1-k}})) = 0$ for all $1 \leqslant k \leqslant K$ and $2 \leqslant j \leqslant 3$.

A.3.2 $(K+3)$*th to* $2(K+1)$*th updates.* We now carry out the next K updates similarly to Section A.2. Define for conciseness $\kappa(\ell, k) = (3-\ell)(K+1)+k$ for all $1 \leqslant k \leqslant K$ and $1 \leqslant \ell \leqslant 2$. Set $\ell = 2$. Starting from $k = 1$, we characterize $P_n^{\kappa(\ell,k)}(\varepsilon)$ (for any $\varepsilon \in \mathbb{R}$) as being the data-generating distribution of O such that *(i)* $\mathrm{Pa}(C_{\ell,K+1-k})$ has the same distribution under $P_n^{\kappa(\ell,k)}(\varepsilon)$ as under $P_n^{\kappa(\ell,k)}$, *(ii)* the conditional distribution of $C_{\ell,K+1-k}$ given $\mathrm{Pa}(C_{\ell,K+1-k})$ is the Bernoulli law with parameter

$$\operatorname{expit}\left(\operatorname{logit} P_n^{\kappa(\ell,k)}(C_{\ell,K+1-k} = 1|\mathrm{Pa}(C_{\ell,K+1-k})) + \varepsilon H_n^{\kappa(\ell,k)}(O)\right),$$

and *(iii)* all remaining variables have the same conditional distributions given their parents under $P_n^{\kappa(\ell,k)}(\varepsilon)$ as under $P_n^{\kappa(\ell,k)}$. One can easily check that $\{P_n^{\kappa(\ell,k)}(\varepsilon) : \varepsilon \in \mathbb{R}\}$ fluctuates $P_n^{\kappa(\ell,k)}$ (*i.e.*, $P_n^{\kappa(\ell,k)}(0) = P_n^{\kappa(\ell,k)}$) in the direction of $\Pi(D(P_n^{\kappa(\ell,k)})|\mathcal{T}(Q_{C_{\ell,K+1-k}}))$ for the weighted log-likelihood loss function $L_n^{\kappa(\ell,k)}$ characterized by $L_n^{\kappa(\ell,k)}(P)(O) = -\frac{\mathbf{1}\{A_{0:(\ell-1)}=1_{0:(\ell-1)}\}}{G_{\ell-1}(P_n^{\kappa(\ell,k)})(O)} \log P(O)$. Indeed,

$$\frac{\partial}{\partial\varepsilon} L_n^{\kappa(\ell,k)}(P_n^{\kappa(\ell,k)}(\varepsilon))(O)|_{\varepsilon=0} = -\Pi(D(P_n^{\kappa(\ell,k)})|\mathcal{T}(Q_{C_{\ell,K+1-k}}))(O).$$

The optimal stretch $\varepsilon_n^{\kappa(\ell,k)}$ along that fluctuation of $P_n^{\kappa(\ell,k)}$ is that which minimizes the empirical loss, *i.e.*, $\varepsilon_n^{\kappa(\ell,k)} = \arg\min_{\varepsilon\in\mathbb{R}} P_n L_n^{\kappa(\ell,k)}(P_n^{\kappa(\ell,k)}(\varepsilon))$, and we conclude this update by setting $P_n^{\kappa(\ell,k)+1} = P_n^{\kappa(\ell,k)}(\varepsilon_n^{\kappa(\ell,k)})$.

We now derive $E_{P_n^{\kappa(\ell,k)+1}}[\Delta_j(P_n^{\kappa(\ell,k)+1})(O)|\mathrm{Pa}(C_{\ell,K+1-k})]$ for the sake of preparing the next fluctuation. If $k < K$ then we can deduce from there

$$\begin{aligned} &E_{P_n^{\kappa(\ell,k)+1}}[\Delta_j(P_n^{\kappa(\ell,k)+1})(O)|\mathrm{Pa}(C_{\ell,K-k})] \\ &\qquad = E_{P_n^{\kappa(\ell,k)+1}}[E_{P_n^{\kappa(\ell,k)+1}}[\Delta_j(P_n^{\kappa(\ell,k)+1})(O)|\mathrm{Pa}(C_{\ell,K+1-k})]|\mathrm{Pa}(C_{\ell,K-k})], \end{aligned}$$

hence in turn the projection $\Pi(\Delta_j(P_n^{\kappa(\ell,k)+1})|\mathcal{T}(Q_{C_{\ell,K-k}}))$, which can be written as

$$\begin{aligned} &\Pi(\Delta_j(P_n^{\kappa(\ell,k)+1})|\mathcal{T}(Q_{C_{\ell,K-k}}))(O) \\ &= E_{P_n^{\kappa(\ell,k)+1}}[\Delta_j(P_n^{\kappa(\ell,k)+1})(O)|C_{\ell,K-k}, \mathrm{Pa}(C_{\ell,K-k})] - E_{P_n^{\kappa(\ell,k)+1}}[\Delta_j(P_n^{\kappa(\ell,k)+1})(O)|\mathrm{Pa}(C_{\ell,K-k})]. \end{aligned}$$

Then we set $H_n^{\kappa(\ell,k)+1}(O) = \Pi(\Delta_j(P_n^{\kappa(\ell,k)+1})|\mathcal{T}(Q_{C_{\ell,K-k}}))(O)$, which satisfies

$$\Pi(D(P_n^{\kappa(\ell,k)+1})|\mathcal{T}(Q_{C_{\ell,K-k}}))(O) = H_n^{\kappa(\ell,k)+1}(O)\frac{\mathbf{1}\{A_{0:(\ell-1)} = 1_{0:(\ell-1)}\}}{G_{\ell-1}(P_n^{\kappa(\ell,k)+1})(O)}.$$

Now, everything is in place to carry on and undertake the next fluctuation: if $k < K$ then we increment $k \leftarrow k+1$ and repeat the procedure described in the current Section A.3.2, otherwise we proceed to Section A.3.3. We emphasize that:

- deriving what is $P_n^{\kappa(\ell,k)+1}$ from $P_n^{\kappa(\ell,k)}$ merely amounts to fitting a weighted logistic regression of $C_{\ell,K+1-k}$ on $H_n^{\kappa(\ell,k)}(O)$ with an offset equal to

  $$\mathrm{logit}\, P_n^{\kappa(\ell,k)}(C_{\ell,K+1-k} = 1|\mathrm{Pa}(C_{\ell,K+1-k})),$$

  using only those observations for which $A_{0:1}^{(i)} = 1_{0:1}$ and the weights $1/G_{\ell-1}(P_n^0)(O^{(i)})$;

- eventually, we have $P_n\Pi(D(P_n^{2(K+1)})|\mathcal{T}(Q_{L_j})) = P_n\Pi(D(P_n^{2(K+1)})|\mathcal{T}(Q_{C_{j,K+1-k}})) = 0$ for all $1 \leqslant k \leqslant K$ and $2 \leqslant j \leqslant 3$.

A.3.3 $(2K+3)$*th update.* We characterize $\Delta_1(P)$ for all $P \in \mathcal{M}$ by setting $\Delta_1(P)(O) = E_P[\Delta_2(P)(O)|A_1 = 1, \mathrm{Pa}(A_1)]$ (the subscript '1' refers to the fact that $\Delta_1(P)(O)$ depends on O only through L_1 and $\mathrm{Pa}(L_1)$). We know $\Delta_1(P_n^{2K+2})$ because we have derived earlier the conditional expectation $E_{P_n^{2K+2}}[\Delta_2(P_n^{2K+2})(O)|\mathrm{Pa}(C_{2,1}) \equiv (A_1, \mathrm{Pa}(A_1))]$. Following the lines of Section A.3.1, we derive $H_n^{2K+2}(O) \equiv \Pi(\Delta_1(P_n^{2K+2})|\mathcal{T}(Q_{L_1}))(O)$ which yields the next component of the efficient influence curve, because

$$\Pi(D(P_n^{2K+2})|\mathcal{T}(Q_{L_1}))(O) = H_n^{2K+2}(O)\frac{\mathbf{1}\{A_0 = 1\}}{g_0(P_n^{2K+2})(O)}.$$

Now we fluctuate the conditional distribution of L_1 given $\mathrm{Pa}(L_1)$ only, and do so in the direction of the corresponding component of the current estimate of the efficient influence curve, $D^\star(P_n^{2K+2})$. Similarly to Sections A.1 and A.3.1, we characterize $P_n^{2K+2}(\varepsilon)$ (for any $\varepsilon \in \mathbb{R}$) as being the data-generating distribution for O such that *(i)* $\mathrm{Pa}(L_1)$ has the same distribution under $P_n^{2K+2}(\varepsilon)$ as under P_n^{2K+2}, *(ii)* the conditional distribution of L_1 given $\mathrm{Pa}(L_1)$ is the Bernoulli law with parameter

$$\mathrm{expit}\left(\mathrm{logit}\, P_n^{2K+2}(L_1 = 1|\mathrm{Pa}(L_1)) + \varepsilon H_n^{2K+2}(O)\right),$$

and *(iii)* all remaining variables have the same conditional distributions given their parents under $P_n^{2K+2}(\varepsilon)$ as under P_n^{2K+2}. One can easily check that $\{P_n^{2K+2}(\varepsilon) : \varepsilon \in \mathbb{R}\}$ fluctuates P_n^{2K+2} (*i.e.*, $P_n^{2K+2}(0) = P_n^{2K+2}$) in the direction of $\Pi(D(P_n^{2K+2})|\mathcal{T}(Q_{L_1}))$ for the weighted log-likelihood loss function L_n^{2K+2} characterized by $L_n^{2K+2}(P)(O) = -\frac{\mathbf{1}\{A_0=1\}}{g_0(P_n^{2K+2})(O)} \log P(O)$. Indeed,

$$\frac{\partial}{\partial\varepsilon} L_n^{2K+2}(P_n^{2K+2}(\varepsilon))(O)|_{\varepsilon=0} = -\Pi(D(P_n^{2K+2})|\mathcal{T}(Q_{L_1}))(O).$$

The optimal stretch ε_n^{2K+2} along that fluctuation of P_n^{2K+2} is that which minimizes the empirical loss, *i.e.*, $\varepsilon_n^{2K+2} = \arg\min_{\varepsilon\in\mathbb{R}} P_n L_n^{2K+2}(P_n^{2K+2}(\varepsilon))$, and we conclude this update by setting $P_n^{2K+3} = P_n^{2K+2}(\varepsilon_n^{2K+2})$.

For the sake of preparing the next fluctuation, we derive $E_{P_n^{2K+3}}[\Delta_1(P_n^{2K+3})(O)|\mathrm{Pa}(L_1)]$, from which we deduce $E_{P_n^{2K+3}}[\Delta_1(P_n^{2K+3})(O)|\mathrm{Pa}(C_{1,K})]$ hence in turn the projection $H_n^{2K+3}(O) \equiv \Pi(\Delta_1(P_n^{2K+3})|\mathcal{T}(Q_{C_{1,K}}))(O)$, which satisfies

$$\Pi(D(P_n^{2K+3})|\mathcal{T}(Q_{C_{1,K}}))(O) = H_n^{2K+3}(O)\frac{\mathbf{1}\{A_0 = 1\}}{g_0(P_n^{2K+3})(O)}.$$

We emphasize that:

- deriving P_n^{2K+3} from P_n^{2K+2} merely amounts to fitting a weighted logistic regression of L_1 on $H_n^{2K+2}(O)$ with an offset equal to logit $P_n^{2K+2}(L_1 = 1|\mathrm{Pa}(L_1))$, using only those observations for which $A_0^{(i)} = 1$ and the corresponding weights $1/g_0(P_n^0)(O^{(i)})$;
- we have $P_n\Pi(D(P_n^{2K+3})|\mathcal{T}(Q_{L_{j'}})) = P_n\Pi(D(P_n^{2K+3})|\mathcal{T}(Q_{C_{j,K+1-k}})) = 0$ for all $1 \leqslant k \leqslant K$, $2 \leqslant j \leqslant 3$, and $1 \leqslant j' \leqslant 3$.

A.3.4 $(2K+4)$*th to last updates.* The next K updates are accurately described in Section A.3.2 when one sets $\ell = 1$ there. We do not update $Q_W(P_n^0)$, the marginal distribution of W. Actually, even if one tried to in the same spirit as we have carried out the previous updates, one would end up *not* updating $Q_W(P_n^0)$. Eventually, the whole updating procedure bends P_n^0 into $P_n^* \equiv P_n^{3(K+1)}$. Its differences relative to the initial estimator stem from the fact that P_n^* targets $\Psi(P_0)$ whereas P_n^0 does not. This is notably reflected in the following equalities: for all $1 \leqslant k \leqslant K$ and $1 \leqslant j \leqslant 3$,

$$P_n\Pi(D(P_n^*)|\mathcal{T}(Q_{L_j})) = P_n\Pi(D(P_n^*)|\mathcal{T}(Q_{C_{j,K+1-k}})) = 0. \tag{A.2}$$

We can add that, since on one hand $\Psi(P) = E_P(Y\mathbf{1}\{A_{0:2} = 1_{0:2}\}/G_2(P)(O))$ and on the other hand $Q_W(P_n^*) = Q_W(P_n^0) = P_{n,W}$, the empirical distribution of W, it also automatically holds that

$$P_n\Pi(D(P_n^*)|\mathcal{T}(Q_W)) = 0. \tag{A.3}$$

Combining (A.2) and (A.3) yields the result stated in Lemma 4.2.

A.3.5 *Comments.* Consider first the computation of the TMLE $\Psi(P_n^*)$ which proves, perhaps surprisingly, very easy at this stage of the procedure. Characterize $\Delta_0(P)$ for all $P \in \mathcal{M}$ by setting $\Delta_0(P)(W) = E_P[\Delta_1(P)(O)|A_0 = 1, W]$. We know $\Delta_0(P_n^*)$ because we have derived earlier the conditional expectation $E_{P_n^*}[\Delta_1(P_n^*)(O)|\mathrm{Pa}(C_{1,1}) \equiv (A_0, W)]$. Furthermore, the "first term" of the right-hand side of (4.4) evaluated at $P = P_n^*$ satisfies (A.3) and

$$\Pi(D(P_n^*)|\mathcal{T}_{Q_W})(O) = E_{P_n^*}\left(\frac{\mathbf{1}\{A_0 = 1\}}{g_0(P_n^*)(O)} E_{P_n^*}\left[\Delta_1(P_n^*)(O) - \Psi(P_n^*)\middle| A_0, W\right]\right) = \Delta_0(P_n^*)(W) - \Psi(P_n^*).$$

Thus, it straightforwardly holds that $\Psi(P_n^*) = \frac{1}{n}\sum_{i=1}^n \Delta_0(P_n^*)(W^{(i)})$.

Second, we would like to comment on the reason why we systematically relied on *weighted* logistic regressions on covariates $H_n^k(O)$ with weights of the form $\mathbf{1}\{A_{0:j} = 1_{0:j}\}/G_j(P_n^k)(O)$ instead of unweighted logistic regressions on covariates of the form $H_n^k(O)\mathbf{1}\{A_{0:j} = 1_{0:j}\}/G_j(P_n^k)(O)$. By using this trick, we actually guarantee that the numerical values of $G_j(P_n^k)(O)$ are required only at our observations $O^{(1)}, \ldots, O^{(n)}$, therefore limiting the overall computational burden.

B. The Simulation Scheme

We describe here the data-generating distribution P^s used for the simulation study. First, $W = (W_1, W_2, C_0, L_0)$ is drawn from the empirical distribution of W based on the DAIFI data set. If $L_0 = 1$, then all the remaining components are fully determined by our set of conventions. Otherwise, the next components of O are successively drawn conditionally on their past and following the chronological ordering. Specifically, A_0 is conditionally drawn from the Bernoulli law with parameter $\mathrm{expit}\,(1 + 0.05C_0)$ if $W_1 = 0$ and $\mathrm{expit}\,(1 + 0.1C_0)$ if $W_1 = 1$. If $A_0 = 0$ then all the remaining components are fully determined by our set of conventions. Otherwise, C_1 is conditionally drawn from the binomial law with parameter $(n_1, p_1) = (\min(19, \max(1, (C_0 + 1)/p_1)), \frac{0.5+C_0}{3+2\max(4,C_0)})$ if $W_1 = 0$ or $(n_1, p_1) = (\min(19, \max(1, C_0 + 1/p_1)), \frac{1/3+C_0}{2.67+2\max(4,C_0)})$ if $W_1 = 1$. Then L_1 is conditionally drawn from the Bernoulli law with parameter $\mathrm{expit}\,(0.26\log(W_1) - 0.48\log(2 + C_1^3))$ if $W_1 = 0$ and $\mathrm{expit}\,(0.52\log(W_1) - \log(2 + 0.01C_1^3))$ if $W_1 = 1$.

If $L_1 = 1$, then all the remaining components are fully determined by our set of conventions. Otherwise, A_1 is conditionally drawn from the Bernoulli law with parameter $\mathrm{expit}\,(2\min(1, 0.25C_1))$

if $W_1 = 0$ and $\text{expit}\,(\min(1, 0.25C_1))$ if $W_1 = 1$. If $A_1 = 0$ then all the remaining components are fully determined by our set of conventions. Otherwise, C_2 is conditionally drawn from the binomial law with parameter $(n_2, p_2) = (\min(19, \max(1, (C_1 + 1)/p_2)), \frac{0.75+C_1}{3.5+2\max(4,C_1)})$ if $W_1 = 0$ or $(n_2, p_2) = (\min(19, \max(1, C_1 + 1/p_2)), \frac{0.67+C_1}{3.33+2\max(4,C_1)})$ if $W_1 = 1$. Then L_2 is conditionally drawn from the Bernoulli law with parameter $\text{expit}\,(1 - 0.29\sqrt{W_1} - 0.72\log(2 + C_2^3))$ if $W_1 = 0$ and $\text{expit}\,(2 - 0.15\sqrt{W_1} - 2\log(2 + 0.01C_2^3))$ if $W_1 = 1$.

If $L_2 = 1$, then all the remaining components are fully determined by our set of conventions. Otherwise, A_2 is conditionally drawn from the Bernoulli law with parameter $\text{expit}\,(0.1 - 0.05\min(1, 0.25C_2))$ if $W_1 = 0$ and $\text{expit}\,(0.2 - 0.2\min(1, 0.25C_2))$ if $W_1 = 1$. If $A_2 = 0$ then all the remaining components are fully determined by our set of conventions. Otherwise, C_3 is conditionally drawn from the binomial law with parameter $(n_3, p_3) = (\min(19, \max(1, (C_2 + 1)/p_3)), \frac{0.83+C_2}{3.67+2\max(4,C_2)})$ if $W_1 = 0$ or $(n_3, p_3) = (\min(19, \max(1, C_2 + 1/p_3)), \frac{0.8+C_2}{3.6+2\max(4,C_2)})$ if $W_1 = 1$. Then, L_3 is conditionally drawn from the Bernoulli law with parameter $\text{expit}\,(-4 - W_1^2/2116 + 0.72\log(2 + C_2^3))$ if $W_1 = 0$ and $\text{expit}\,(-5 + W_1^2/1058 - 4\log(2 + 0.01C_2^3))$ if $W_1 = 1$.

Note that the resulting distribution P^s actually obeys a system of structural equations as discussed in Section 3.2 of the main manuscript. The functional forms are made intricate in order to convince the reader that the super learner library is misspecified for the estimation of both $Q(P^s)$ and $g(P^s)$, see Section C. Furthermore, the true value of the parameter of interest $\Psi(P^s)$ can be estimated with great precision by Monte Carlo. Using a simulated dataset of one million observations under the intervention $A_{0:2} = 1_{0:2}$ yields $\Psi(P^s) \approx 0.72$.

C. Specifics of the Super Learning Procedure

Super learning is a cross-validation based aggregation method that builds a predictor as a convex combination of base predictors (van der Laan *and others*, 2007; van der Laan and Rose, 2011). The weights of the convex combination are chosen so as to minimize the prediction error. This error is expressed in terms of the nonnegative least squares (NNLS) loss function (Lawson and Hanson, 1995) and estimated by V-fold cross-validation. Heuristically the resulting predictor is by construction at least as good as the best of the base predictors. This statement has a rigorous form implying oracle inequalities (van der Laan *and others*, 2007; van der Laan and Rose, 2011). The algorithmic challenge is easily overcome, thanks to the `R`-package `SuperLearner` by Polley and van der Laan (2011) and the library of `R` packages (R Development Core Team, 2010) built by the statistical community. We choose $V = 5$. As for the base predictors, which are classifiers since we only consider binary random variables, they involve (in alphabetical order):

- Elastic nets (large library only): we use the `glmnet` `R` package by Friedman *and others* (2010), with its default values and the tuning parameter `alpha` set to 1 and 0.5.

- Generalized additive models: we use the `gam` `R` package by Hastie (2011), with its default values and the tuning parameter `deg.gam` set to 2 and 3.

- Generalized linear models (large library only): we use the `glm` `R` function.

- Random forests: we use the `randomForest` `R` package by Liaw and Wiener (2002), with its default values.

D. Miscellanea

Proof of equality (3.2). Set $1 \leqslant j \leqslant 3$. The right-hand side term of (3.2) equals

$$
\begin{aligned}
& P(L_j = 1 | C_{0:j}, L_{0:j-1}, W) \\
& = E[P(L_j = 1 | C_{0:j}, A_{0:j-1}, L_{0:j-1}, W) | C_{0:j}, L_{0:j-1}, W] \\
& = \sum_{a_{0:j-1} \in \{0,1\}^j} P(L_j = 1 | C_{0:j}, A_{0:j-1} = a_{0:j-1}, L_{0:j-1}, W) \times P(A_{0:j-1} = a_{0:j-1} | C_{0:j}, L_{0:j-1}, W) \\
& = P(L_j = 1 | C_{0:j}, A_{0:j-1} = 1_{0:j-1}, L_{0:j-1}, W) \times P(A_{0:j-1} = 1_{0:j-1} | C_{0:j}, L_{0:j-1}, W) \\
& = \mathbf{1}\{C_j \geqslant 1\} P(L_j = 1 | C_{0:j}, A_{0:j-1} = 1_{0:j-1}, L_{0:j-1}, W) \\
& = \mathbf{1}\{C_j \geqslant 1, A_{0:j-1} = 1_{0:j-1}\} P(L_j = 1 | C_{0:j}, A_{0:j-1} = 1_{0:j-1}, L_{0:j-1}, W).
\end{aligned}
$$

Furthermore, the left-hand side of (3.2) satisfies

$$
\begin{aligned}
& P(L_j = 1 | C_{0:j}, A_{0:j-1}, L_{0:j-1}, W) \\
& = \mathbf{1}\{A_{0:j-1} = 1_{0:j-1}\} P(L_j = 1 | C_{0:j}, A_{0:j-1} = 1_{0:j-1}, L_{0:j-1}, W) \\
& = \mathbf{1}\{C_j \geqslant 1, A_{0:j-1} = 1_{0:j-1}\} P(L_j = 1 | C_{0:j}, A_{0:j-1} = 1_{0:j-1}, L_{0:j-1}, W).
\end{aligned}
$$

Clearly, the two expressions coincide. □

References

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**(1), 1–22.

Hastie, T. (2011). *Generalized additive models*. R package version 1.04.1.

Lawson, C. L. and Hanson, R. J. (1995). *Solving least squares problems*, Volume 15. Society for Industrial Mathematics.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News* **2**(3), 18–22.

Polley, E. and van der Laan, M. J. (2011). *SuperLearner*. R package version 2.0-4.

R Development Core Team. (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

van der Laan, M. J., Polley, E. C. and Hubbard, A. E. (2007). Super learner. *Stat. Appl. Genet. Mol. Biol.* **6**, Art. 25, 23 pp. (electronic).

van der Laan, M. J. and Rose, S. (2011). *Targeted learning*, Springer Series in Statistics. New York: Springer. Causal inference for observational and experimental data.

[*Received August 1, 2010; revised October 1, 2010; accepted for publication November 1, 2010*]