Machine Learning: Day 2

Sherri Rose

Associate Professor
Department of Health Care Policy
Harvard Medical School

drsherrirose.com
@sherrirose

February 28, 2017
Goals: Day 2


2. Be introduced to the ideas behind machine learning approaches as tools for confronting the curse of dimensionality.

3. Become familiar with the properties and basic implementation of TMLE for effect estimation.
[Motivation]
Why Most Published Research Findings Are False
John P. A. Ioannidis

Summary

John P. A. Ioannidis

Do We Really Know What Makes Us Healthy?
By GARY TAUBES
Why Most Published Research Findings Are False

John P. A. Ioannidis

The New York Times
September 16, 2007
Do We Really Know What Makes Us Healthy?
By GARY TAUBES

4 high impact zones in statistical discovery with big data
September 22, 2014

FierceBigData

By: Sherri Rose, Harvard University; David Dunson, Duke University; Tyler McCormick, University of Washington; and Cynthia Rudin, MIT

Big data is transforming society with the help of statisticians, who possess in-depth experience and expertise in the art and science surrounding data. The American Statistical Association, or ASA, has recently released a white paper entitled "Discovery with Data: Leveraging Statistics with Computer Science to Transform Science and Society." This paper highlights high-impact areas where statistical science is being applied to transformative big data research questions. Statistics is, by definition, the science of learning from data, and has had a key impact in several of the most prominent fields of discovery, including the biological sciences, health care, business analytics and recommendation systems, and the social sciences. There is a strong need to work in integrated teams comprised of domain experts, statisticians, and computer scientists in order to solve these complicated problems, which require tailored solutions using the influx of big data.
Electronic Health Databases

The increasing availability of electronic medical records offers a new resource to public health researchers.

General usefulness of this type of data to answer targeted scientific research questions is an open question.

Need novel statistical methods that have desirable statistical properties while remaining computationally feasible.
Nested case-control sample (n=27,012).

- **Outcome**: death.
- **Covariates**: 184 medical flags, gender & age.

Ensembling method outperformed all other algorithms.

Generally weak signal with $R^2 = 0.11$.

Observed data structure on a subject can be represented as $O = (Y, \Delta, \Delta X)$, where $X = (W, Y)$ is the full data structure, and $\Delta$ denotes the indicator of inclusion in the second-stage sample.

How will this electronic database perform in comparison to a cohort study?

van der Laan & Rose (2011)
Super Learner: Sonoma Cohort Study

Cohort study of \( n = 2,066 \) residents of Sonoma, CA aged 54 and over.

- Outcome: death.
- Covariates: gender, age, self-rated health, leisure-time physical activity, smoking status, cardiac event history, and chronic health condition status.
- \( R^2 = 0.201 \)

Two-fold improvement with less than 10% of the subjects & less than 10% the number of covariates.

What possible conclusions can we draw?

Rose (2013)
High Dimensional ‘Big Data’ Parametric Regression

- Often dozens, hundreds, or even thousands of potential variables
High Dimensional ‘Big Data’ Parametric Regression

- Often dozens, hundreds, or even thousands of potential variables
- Impossible challenge to correctly specify the parametric regression
High Dimensional ‘Big Data’ Parametric Regression

- Often dozens, hundreds, or even thousands of potential variables
- Impossible challenge to correctly specify the parametric regression
- May have more unknown parameters than observations
High Dimensional ‘Big Data’ Parametric Regression

- Often dozens, hundreds, or even thousands of potential variables
- Impossible challenge to correctly specify the parametric regression
- May have more unknown parameters than observations
- True functional might be described by a complex function not easily approximated by main terms or interaction terms
Complications of Human Art in ‘Big Data’ Statistics

1. Fit several parametric models; select a favorite one
2. The parametric model is misspecified
3. The target parameter is interpreted as if the parametric model is correct
4. The parametric model is often data-adaptively (or worse!) built, and this part of the estimation procedure is not accounted for in the variance
Estimation is a Science

1. **Data**: realizations of random variables with a probability distribution.

2. **Statistical Model**: actual knowledge about the shape of the data-generating probability distribution.

3. **Statistical Target Parameter**: a feature/function of the data-generating probability distribution.

4. **Estimator**: an a priori-specified algorithm, benchmarked by a dissimilarity-measure (e.g., MSE) w.r.t. target parameter.
Roadmap for Effect Estimation

How does one translate the results from studies, how do we take the information in the data, and draw effective conclusions?

- Define the Research Question
  - Specify Data
  - Specify Model
    - Specify the Parameter of Interest
- Estimate the Target Parameter
- Inference
  - Standard Errors / CIs
  - Interpretation
Random variable $O$, observed $n$ times, could be defined in a simple case as $O = (W, A, Y) \sim P_0$ if we are without common issues such as missingness and censoring.

- $W$: vector of covariates
- $A$: exposure or treatment
- $Y$: outcome

This data structure makes for effective examples, but data structures found in practice are frequently more complicated.
Data: Censoring & Missingness

Define $O = (W, A, \tilde{T}, \Delta) \sim P_0$.

- $T$: time to event $Y$
- $C$: censoring time
- $\tilde{T} = \min(T, C)$: represents the $T$ or $C$ that was observed first
- $\Delta = I(T \leq \tilde{T}) = I(C \geq T)$: indicator that $T$ was observed at or before $C$

Define $O = (W, A, \Delta, \Delta Y) \sim P_0$.

- $\Delta$: Indicator of missingness
General case: Observe $n$ i.i.d. copies of random variable $O$ with probability distribution $P_0$.

The data-generating distribution $P_0$ is also known to be an element of a statistical model $\mathcal{M}$: $P_0 \in \mathcal{M}$.

A **statistical model** $\mathcal{M}$ is the set of possible probability distributions for $P_0$; it is a collection of probability distributions.

If all we know is that we have $n$ i.i.d. copies of $O$, this can be our statistical model, which we call a nonparametric statistical model.
A statistical model can be augmented with additional (nontestable causal) assumptions, allowing one to enrich the interpretation of $\Psi(P_0)$. This does not change the statistical model.
Target Parameters

Define the parameter of the probability distribution \( P \) as function of \( P : \psi(P) \).

\[ \psi_{RD} = \psi_{RD}(P) = E_W[E(Y \mid A = 1, W) - E(Y \mid A = 0, W)] = E(Y_1) - E(Y_0) = P(Y_1 = 1) - P(Y_0 = 1) \]

\[ \psi_{RR} = \frac{P(Y_1 = 1)}{P(Y_0 = 1)} \]

and

\[ \psi_{OR} = \frac{P(Y_1 = 1)P(Y_0 = 0)}{P(Y_1 = 0)P(Y_0 = 1)}. \]

\( Y \) is the outcome, \( A \) the exposure, and \( W \) baseline covariates.
Effect Estimation vs. Prediction

Both **effect** and **prediction** research questions are inherently **estimation** questions, but they are distinct in their goals.
Effect Estimation vs. Prediction

Both **effect** and **prediction** research questions are inherently **estimation** questions, but they are distinct in their goals.

**Effect:** Interested in estimating the effect of exposure on outcome adjusted for covariates.
Effect Estimation vs. Prediction

Both effect and prediction research questions are inherently estimation questions, but they are distinct in their goals.

Effect: Interested in estimating the effect of exposure on outcome adjusted for covariates.

Prediction: Interested in generating a function to input covariates and predict a value for the outcome.
[(Causal) Effect Estimation]
Learning from Data

Just what type of studies are we conducting? The often quoted “ideal experiment” is one that cannot be conducted in real life.
Assume a structural causal model (SCM) (Pearl 2009), comprised of endogenous variables $X = (X_j : j)$ and exogenous variables $U = (U_{X_j} : j)$.

- Each $X_j$ is a deterministic function of other endogenous variables and an exogenous error $U_j$.
- The errors $U$ are never observed.
- For each $X_j$ we characterize its parents from among $X$ with $Pa(X_j)$. 

Causal Model
$$X_j = f_{X_j}(Pa(X_j), U_{X_j}), \ j = 1 \ldots, J,$$

The functional form of $f_{X_j}$ is often unspecified.

An SCM can be fully parametric, but we do not do that here as our background knowledge does not support the assumptions involved.
Causal Model

We could specify the following SCM:

\[
W = f_W(U_W), \\
A = f_A(W, U_A), \\
Y = f_Y(W, A, U_Y),
\]

Recall that we assume for the full data:

1. for each \( X_j \), \( X_j = f_j(Pa(X_j), U_X) \) depends on the other endogenous variables only through the parents \( Pa(X_j) \),
2. the exogenous variables have a particular joint distribution \( P_U \); \( U_A \perp U_Y | W \).

In our simple study, \( X = (W, A, Y) \), and \( Pa(A) = W \). We know this due to the time ordering of the variables.
Figure: Causal graphs with various assumptions about the distribution of $P_U$
A Note on Causal Assumptions

We could alternatively use the Neyman–Rubin Causal Model and assume
- randomization ($A \perp Y_a \mid W$) and
- stable unit treatment value assumption (SUTVA; no interference between subjects and consistency assumption).
Positivity Assumption

We need that each possible exposure level occurs with some positive probability within each stratum of $W$.

For our data structure $(W, A, Y)$ we are assuming:

$$ P_0(A = 1 \mid W = w) > 0 \text{ and } P_0(A = 0 \mid W = w) > 0, $$

for each possible $w$. 
An **estimator** is an algorithm that can be applied to any empirical distribution to provide a mapping from the empirical distribution to the parameter space.

- **Maximum-Likelihood-Based Estimators**
- **Estimating-Equation-Based Methods**

The target parameters we discussed depend on $\mathcal{P}_0$ through the conditional mean $\bar{Q}_0(A, W) = E_0(Y \mid A, W)$, and the marginal distribution $Q_{W,0}$ of $W$. Thus we can also write $\Psi(Q_0)$, where $Q_0 = (\bar{Q}_0, Q_{W,0})$. 
Maximum-Likelihood-Based Estimators will be of the type

\[ \psi_n = \Psi(Q_n) = \frac{1}{n} \sum_{i=1}^{n} \{ \bar{Q}_n(1, W_i) - \bar{Q}_n(0, W_i) \}, \]

where this estimate is obtained by plugging in \( Q_n = (\bar{Q}_n, Q_{W,n}) \) into the mapping \( \Psi \). \( \bar{Q}_n(A = a, W_i) = E_n(Y \mid A = a, W_i) \).

Estimating-Equation-Based Methods An estimating function is a function of the data \( O \) and the parameter of interest. If \( D(\psi)(O) \) is an estimating function, then we can define a corresponding estimating equation: \( 0 = \sum_{i=1}^{n} D(\psi)(O_i) \), and solution \( \psi_n \) satisfying \( \sum_{i=1}^{n} D(\psi_n)(O_i) = 0 \).
Maximum-Likelihood-Based Methods

MLE using regression. Outcome regression estimated with parametric methods and plugged into

\[ \psi_n = \frac{1}{n} \sum_{i=1}^{n} \{ \bar{Q}_n(1, W_i) - \bar{Q}_n(0, W_i) \}. \]
**MLE using regression.** Outcome regression estimated with parametric methods and plugged into

$$\psi_n = \frac{1}{n} \sum_{i=1}^{n} \{ \bar{Q}_n(1, W_i) - \bar{Q}_n(0, W_i) \}.$$  

**STOP!** When does this differ from traditional regression?
Maximum-Likelihood-Based Methods

**MLE using regression: Continuous outcome example.**
True effect is -0.35

$W_1 = \text{gender}$

$W_2 = \text{medication use}$

$A = \text{high ozone exposure}$

$Y = \text{continuous measure of lung function}$

Model 1: $E(Y \mid A) = \alpha_0 + \alpha_1 A$

**Both Effects:** -0.23

Model 2: $E(Y \mid A, W) = \alpha_0 + \alpha_1 A + \alpha_2 W_1 + \alpha_3 W_2$

**Both Effects:** -0.36

Model 3: $E(Y \mid A, W) = \alpha_0 + \alpha_1 A + \alpha_2 W_1 + \alpha_3 A \cdot W_2$

**Regression Effect:** -0.49

**MLE Effect:** -0.34
MLE using regression: Binary outcomes.

\[ P(Y = 1 \mid A, W) = \frac{1}{1 + e^{-\beta_0 + \beta_1 A + \beta_2 W}} \]

\[ EY_a = P(Y_a = 1) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{1 + e^{-\beta_0 + \beta_1 A_i + \beta_2 W_i}} \]

\[ \frac{EY_1/(1 - EY_1)}{EY_0/(1 - EY_0)} \neq e^{\beta_1} \]
Medical Schools in Fragile States: Delivery of Care

We found that fragile states lack the infrastructure to train sufficient numbers of medical professionals to meet their population health needs. Fragile states were 1.76 (95%CI 1.07-2.45) to 2.37 (95%CI 1.44-3.30) times more likely to have < 2 medical schools than non-fragile states.

Mateen, McKenzie, Rose (2017)
MLE using machine learning. Outcome regression estimated with machine learning and plugged into

\[
\psi_n = \frac{1}{n} \sum_{i=1}^{n} \{ \bar{Q}_n(1, W_i) - \bar{Q}_n(0, W_i) \}.
\]
Machine Learning Estimation of $\bar{Q}(A, W) = E(Y \mid A, W)$
Machine learning aims to
▶ “smooth” over the data
▶ make fewer assumptions

Polley et al. (2011)
Purely nonparametric model with high dimensional data?

- $p > n!$
- data sparsity
Ensembling methods allow implementation of multiple algorithms. Do not need to decide beforehand which single technique to use; can use several by incorporating cross-validation.
Build a collection of algorithms consisting of all weighted averages of the algorithms.

One of these weighted averages might perform better than one of the algorithms alone.

Data

Algorithm_a
Algorithm_b
Algorithm_p
Algorithm_a
Algorithm_b
Algorithm_p
Algorithm_a
Algorithm_b
Algorithm_p
... of weighted combinations

\[ E_n[Y|Z] = \alpha_{a,n}Z_a + \alpha_{b,n}Z_b + \ldots + \alpha_{p,n}Z_p \]

Super learner function

Image credit: Polley et al. (2011)
Noncommunicable Disease and Poverty

Studied relative risk of death from noncommunicable disease on three poverty measures in Matlab, Bangladesh.

Parametric regression standardization
- Asset quintile: 1.19
- Self-rated condition: 1.16
- Landholding: 1.14

Machine-learning (super learner) estimation
- Asset quintile: 1.15
- Self-rated condition: 1.13
- Landholding: 1.11

Implemented parametric and machine learning substitution estimators.

Mirelman et al. (2016)
Estimating Equation Methods

**IPW.** Estimate causal risk difference with

\[ \psi_n = \frac{1}{n} \sum_{i=1}^{n} \{I(A_i = 1) - I(A_i = 0)\} \frac{Y_i}{g_n(A_i, W_i)}. \]

This estimator is a solution of an IPW estimating equation that relies on an estimate of the treatment mechanism, playing the role of a nuisance parameter of the IPW estimating function.

**A-IPW.** One estimates \( \Psi(P_0) \) with

\[ \psi_n = \frac{1}{n} \sum_{i=1}^{n} \{I(A_i = 1) - I(A_i = 0)\} \frac{Y_i}{g_n(A_i, W_i)} (Y_i - \bar{Q}_n(A_i, W_i)) \]

\[ + \frac{1}{n} \sum_{i=1}^{n} \{\bar{Q}_n(1, W_i) - \bar{Q}_n(0, W_i)\}. \]
Targeted Learning in Nonparametric Models

- Parametric MLE not targeted for effect parameters
- Need a subsequent targeted bias-reduction step

**Targeted Learning**

- Avoid reliance on human art and unrealistic parametric models
- Define interesting parameters
- Target the parameter of interest
- Incorporate machine learning
- Statistical inference
TMLE for Causal Effects

TMLE

Produces a well-defined, unbiased, efficient substitution estimator of target parameters of a data-generating distribution.

It is an iterative procedure that updates an initial (super learner) estimate of the relevant part $Q_0$ of the data generating distribution $P_0$, possibly using an estimate of a nuisance parameter $g_0$. 
Super Learner

Allows researchers to use multiple algorithms to outperform a single algorithm in nonparametric statistical models.

Builds weighted combination of estimators where weights are optimized based on loss-function specific cross-validation to guarantee best overall fit.

Targeted Maximum Likelihood Estimation

With an initial estimate of the outcome regression, the second stage of TMLE updates this initial fit in a step targeted toward making an optimal bias-variance tradeoff for the parameter of interest.
TMLE for Causal Effects

TMLE: Double Robust

- Removes asymptotic residual bias of initial estimator for the target parameter, if it uses a consistent estimator of censoring/treatment mechanism \( g_0 \).
- If initial estimator was consistent for the target parameter, the additional fitting of the data in the targeting step may remove finite sample bias, and preserves consistency property of the initial estimator.

TMLE: Efficiency

- If the initial estimator and the estimator of \( g_0 \) are both consistent, then it is also asymptotically efficient according to semi-parametric statistical model efficiency theory.
TMLE for Causal Effects

**TMLE: In Practice**

Allows the incorporation of machine learning methods for the estimation of both $Q_0$ and $g_0$ so that we do not make assumptions about the probability distribution $P_0$ we do not believe.

Thus, every effort is made to achieve minimal bias and the asymptotic semi-parametric efficiency bound for the variance.
Targeted Learning in Nonparametric Models

This chapter focuses on understanding targeted maximum likelihood estimation.

Recall that TMLE is a two-step procedure where one first obtains an estimate of the data-generating distribution \( \psi \), and then updates this initial fit in a step targeted towards making an optimal bias-variance trade-off.

The procedure is double robust and can incorporate data-adaptive likelihood based estimation procedures to estimate the parameter of interest \( \theta \).

The statistical model is set of possible probability distributions of the data. The targeted estimator of the probability distribution of the data is obtained by mapping the initial estimator to the truth, and the treatment mechanism. The targeted estimator of the probability distribution of the data is obtained by mapping the initial estimator of the probability distribution of the data to the truth.

The true value (estimand) of the target parameter is mapped to the real line with better estimates closer to the truth.
Example: TMLE for the Risk Difference

Note that $\epsilon_n$ is obtained by performing a regression of $Y$ on $H_n^*(A, W)$, where $\bar{Q}_n^0(A, W)$ is used as an offset, and extracting the coefficient for $H_n^*(A, W)$.

We then update $\bar{Q}_n^0$ with $\text{logit}\, \bar{Q}_n^1(A, W) = \text{logit}\, \bar{Q}_n^0(A, W) + \epsilon_n^1 H_n^*(A, W)$. This updating process converges in one step in this example, so that the TMLE is given by $Q_n^* = Q_n^1$. 
Example: Sonoma Cohort Study

Cohort study of \( n = 2,066 \) residents of Sonoma, CA aged 54 and over.

- Outcome was death.
- Covariates were gender, age, self-rated health, leisure-time physical activity, smoking status, cardiac event history, and chronic health condition status.
- The data structure is \( O = (W, A, Y) \), where \( Y = I(T \leq 5 \text{ years}) \), \( T \) is time to the event death.
- No right censoring in this cohort.
## Sonoma Study

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y$</td>
<td>Death occurring within 5 years of baseline</td>
</tr>
<tr>
<td>$A$</td>
<td>LTPA score $\geq 22.5$ METs at baseline$^\dagger$</td>
</tr>
<tr>
<td>$W_1$</td>
<td>Health self-rated as “excellent”</td>
</tr>
<tr>
<td>$W_2$</td>
<td>Health self-rated as “fair”</td>
</tr>
<tr>
<td>$W_3$</td>
<td>Health self-rated as “poor”</td>
</tr>
<tr>
<td>$W_4$</td>
<td>Current smoker</td>
</tr>
<tr>
<td>$W_5$</td>
<td>Former smoker</td>
</tr>
<tr>
<td>$W_6$</td>
<td>Cardiac event prior to baseline</td>
</tr>
<tr>
<td>$W_7$</td>
<td>Chronic health condition at baseline</td>
</tr>
<tr>
<td>$W_8$</td>
<td>$x \leq 60$ years old</td>
</tr>
<tr>
<td>$W_9$</td>
<td>$60 &lt; x \leq 70$ years old</td>
</tr>
<tr>
<td>$W_{10}$</td>
<td>$80 &lt; x \leq 90$ years old</td>
</tr>
<tr>
<td>$W_{11}$</td>
<td>$x &gt; 90$ years old</td>
</tr>
<tr>
<td>$W_{12}$</td>
<td>Female</td>
</tr>
</tbody>
</table>

$^\dagger$ LTPA is calculated from a detailed questionnaire where prior performed vigorous physical activities are assigned standardized intensity values in metabolic equivalents (METs). The recommended level of energy expenditure for the elderly is 22.5 METs.
Sonoma Study

Super learner function

<table>
<thead>
<tr>
<th>ID</th>
<th>W1</th>
<th>...</th>
<th>W12</th>
<th>A</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>66</td>
<td>...</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2066</td>
<td>73</td>
<td>...</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ID</th>
<th>W1</th>
<th>...</th>
<th>Y</th>
<th>$Q_n^0(A_i, W_i)$</th>
<th>$Q_n^0(1, W_i)$</th>
<th>$Q_n^0(0, W_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>66</td>
<td>...</td>
<td>1</td>
<td>...</td>
<td>...</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2066</td>
<td>73</td>
<td>...</td>
<td>1</td>
<td>...</td>
<td>...</td>
<td>0.82</td>
</tr>
</tbody>
</table>
Sonoma Study: Estimating $\bar{Q}_0$

Super learner function

Super learner exposure mechanism function

<table>
<thead>
<tr>
<th>ID</th>
<th>W_1</th>
<th>...</th>
<th>W_{12}</th>
<th>A</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>66</td>
<td>...</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2066</td>
<td>73</td>
<td>...</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Step 1

<table>
<thead>
<tr>
<th>ID</th>
<th>W_1</th>
<th>...</th>
<th>Y</th>
<th>$\bar{Q}_n^0(A_i, W_i)$</th>
<th>$\bar{Q}_n^0(1, W_i)$</th>
<th>$\bar{Q}_n^0(0, W_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>66</td>
<td>...</td>
<td>1</td>
<td>...</td>
<td>...</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2066</td>
<td>73</td>
<td>...</td>
<td>1</td>
<td>...</td>
<td>...</td>
<td>0.82</td>
</tr>
</tbody>
</table>

Step 2

| ID | W_1 | ... | $\bar{Q}_n^0(0, W_i)$ | $g_n(1 | W_i)$ | $g_n(0 | W_i)$ |
|----|-----|-----|-----------------------|-------------|-------------|
| 1  | 66  | ... | 0.77                  | ...         | 0.32        |
|    |     |     |                       |             |             |
|    |     |     |                       |             |             |
| 2066 | 73 | ... | 0.82                  | ...         | 0.47        |

Step 3
At this stage we could plug our estimates $\bar{Q}_n^0(1, W_i)$ and $\bar{Q}_n^0(0, W_i)$ for each subject into our substitution estimator of the risk difference:

$$
\psi_{MLE,n} = \Psi(Q_n) = \frac{1}{n} \sum_{i=1}^{n} \{ \bar{Q}_n^0(1, W_i) - \bar{Q}_n^0(0, W_i) \}.
$$
Sonoma Study: **Estimating** $g_0$

Our targeting step required an estimate of the conditional distribution of LTPA given covariates $W$.

This estimate of $P_0(A \mid W) \equiv g_0$ is denoted $g_n$.

We estimated predicted values using a super learner prediction function, adding two more columns to our data matrix: $g_n(1 \mid W_i)$ and $g_n(0 \mid W_i)$.

(Step 3.)
### Step 2

<table>
<thead>
<tr>
<th>ID</th>
<th>W₁</th>
<th>...</th>
<th>Y</th>
<th>( \bar{Q}_{n}^0(A_i, W_i) )</th>
<th>( \bar{Q}_{n}^0(1, W_i) )</th>
<th>( \bar{Q}_{n}^0(0, W_i) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>66</td>
<td>...</td>
<td>1</td>
<td>...</td>
<td>...</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2066</td>
<td>73</td>
<td>...</td>
<td>1</td>
<td>...</td>
<td>...</td>
<td>0.82</td>
</tr>
</tbody>
</table>

### Step 3

<table>
<thead>
<tr>
<th>ID</th>
<th>W₁</th>
<th>...</th>
<th>( \bar{Q}_{n}^0(1, W_i) )</th>
<th>( g_n(1 \mid W_i) )</th>
<th>( g_n(0 \mid W_i) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>66</td>
<td>...</td>
<td>0.77</td>
<td>...</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2066</td>
<td>73</td>
<td>...</td>
<td>0.82</td>
<td>...</td>
<td>0.45</td>
</tr>
</tbody>
</table>

### Step 4

<table>
<thead>
<tr>
<th>ID</th>
<th>W₁</th>
<th>...</th>
<th>( g_n(0 \mid W_i) )</th>
<th>( H_n^*(A_i, W_i) )</th>
<th>( H_n^*(1, W_i) )</th>
<th>( H_n^*(0, W_i) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>66</td>
<td>...</td>
<td>0.32</td>
<td>...</td>
<td>...</td>
<td>-3.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2066</td>
<td>73</td>
<td>...</td>
<td>0.45</td>
<td>...</td>
<td>...</td>
<td>-2.22</td>
</tr>
</tbody>
</table>
The targeting step used the estimate $g_n$ in a clever covariate to define a parametric working model coding fluctuations of the initial estimator. This clever covariate $H^*_n(A, W)$ is given by

$$H^*_n(A, W) \equiv \left( \frac{I(A = 1)}{g_n(1 \mid W)} - \frac{I(A = 0)}{g_n(0 \mid W)} \right).$$
Sonoma Study: Determining a Submodel

Thus, for each subject with $A_i = 1$ in the observed data, we calculated the clever covariate as $H_n^*(1, W_i) = 1/g_n(1 \mid W_i)$.

Similarly, for each subject with $A_i = 0$ in the observed data, we calculated the clever covariate as $H_n^*(0, W_i) = -1/g_n(0 \mid W_i)$.

We combined these values to form a single column $H_n^*(A_i, W_i)$ in the data matrix. We also added two columns $H_n^*(1, W_i)$ and $H_n^*(0, W_i)$. The values for these columns were generated by setting $a = 0$ and $a = 1$.

(Step 4.)
### Step 3

<table>
<thead>
<tr>
<th>ID</th>
<th>W</th>
<th>$\bar{Q}_n^0(0, W_i)$</th>
<th>$g_n(1 \mid W_i)$</th>
<th>$g_n(0 \mid W_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>66</td>
<td>0.77</td>
<td></td>
<td>0.32</td>
</tr>
<tr>
<td>2066</td>
<td>73</td>
<td>0.82</td>
<td></td>
<td>0.45</td>
</tr>
</tbody>
</table>

### Step 4

<table>
<thead>
<tr>
<th>ID</th>
<th>W</th>
<th>$g_n(0 \mid W_i)$</th>
<th>$H_n^*(A_i, W_i)$</th>
<th>$H_n^*(1, W_i)$</th>
<th>$H_n^*(0, W_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>66</td>
<td>0.32</td>
<td></td>
<td></td>
<td>-3.13</td>
</tr>
<tr>
<td>2066</td>
<td>73</td>
<td>0.45</td>
<td></td>
<td></td>
<td>-2.22</td>
</tr>
</tbody>
</table>

### Step 5

<table>
<thead>
<tr>
<th>ID</th>
<th>W</th>
<th>$H_n^*(0, W_i)$</th>
<th>$\bar{Q}_n^1(1, W_i)$</th>
<th>$\bar{Q}_n^1(0, W_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>66</td>
<td>-3.13</td>
<td>0.74</td>
<td></td>
</tr>
<tr>
<td>2066</td>
<td>73</td>
<td>-2.12</td>
<td>0.81</td>
<td></td>
</tr>
</tbody>
</table>

### Step 6

$$\gamma_n = \frac{1}{n} \sum_{i=1}^{n} [\bar{Q}_n^1(1, W_i) - \bar{Q}_n^1(0, W_i)]$$
We then ran a logistic regression of our outcome $Y$ on the clever covariate using as intercept the offset logit $\bar{Q}^0_n(A, W)$ to obtain the estimate $\epsilon_n$, where $\epsilon_n$ is the resulting coefficient in front of the clever covariate $H^*_n(A, W)$.

We next wanted to update the estimate $\bar{Q}^0_n$ into a new estimate $\bar{Q}^1_n$ of the true regression function $\bar{Q}_0$:

$$\text{logit } \bar{Q}^1_n(A, W) = \text{logit } \bar{Q}^0_n(A, W) + \epsilon_n H^*_n(A, W).$$

This parametric working model incorporated information from $g_n$, through $H^*_n(A, W)$, into an updated regression.
Sonoma Study: **Updating \( \bar{Q}_n^0 \)**

The TMLE of \( Q_0 \) was given by \( Q_n^* = (\bar{Q}_n^1, Q_{W,n}^0) \). With \( \epsilon_n \), we were ready to update our prediction function at \( a = 1 \) and \( a = 0 \) according to the logistic regression working model. We calculated

\[
\text{logit } \bar{Q}_n^1(1, W) = \text{logit } \bar{Q}_n^0(1, W) + \epsilon_n H_n^*(1, W),
\]

for all subjects, and then

\[
\text{logit } \bar{Q}_n^1(0, W) = \text{logit } \bar{Q}_n^0(0, W) + \epsilon_n H_n^*(0, W)
\]

for all subjects and added a column for \( \bar{Q}_n^1(1, W_i) \) and \( \bar{Q}_n^1(0, W_i) \) to the data matrix.

Updating \( \bar{Q}_n^0 \) is also illustrated in Step 5.
\[
\psi_n = \frac{1}{n} \sum_{i=1}^{n} [\bar{Q}_n^1(1, W_i) - \bar{Q}_n^1(0, W_i)]
\]
Sonoma Study: Targeted Substitution Estimator

Our formula from the first step becomes

$$\psi_{TMLE,n} = \Psi(Q^*_n) = \frac{1}{n} \sum_{i=1}^{n} \{ \bar{Q}_n^1(1, W_i) - \bar{Q}_n^1(0, W_i) \}.$$  

This mapping was accomplished by evaluating \(\bar{Q}_n^1(1, W_i)\) and \(\bar{Q}_n^1(0, W_i)\) for each observation \(i\), and plugging these values into the above equation.

Our estimate of the causal risk difference for the mortality study was \(\psi_{TMLE,n} = -0.055\).
\[ \psi_n = \frac{1}{n} \sum_{i=1}^{n} [\tilde{Q}_{n}^1(1, W_i) - \tilde{Q}_{n}^1(0, W_i)] \]
Sonoma Study: Inference (Standard errors)

We then needed to calculate the influence curve for our estimator in order to obtain standard errors:

$$IC_n(O_i) = \left( \frac{I(A_i = 1)}{g_n(1 \mid W_i)} - \frac{I(A_i = 0)}{g_n(0 \mid W_i)} \right) (Y - \bar{Q}_n^1(A_i, W_i))$$

$$+ \bar{Q}_n^1(1, W_i) - \bar{Q}_n^1(0, W_i) - \psi_{TMLE,n},$$

where $I$ is an indicator function: it equals 1 when the logical statement it evaluates, e.g., $A_i = 1$, is true.
Sonoma Study: Inference (Standard errors)

Note that this influence curve is evaluated for each of the $n$ observations $O_i$.

With the influence curve of an estimator one can now proceed with statistical inference as if the estimator minus its estimand equals the empirical mean of the influence curve.
Next, we calculated the sample mean of these estimated influence curve values:

$$\bar{IC}_n = \frac{1}{n} \sum_{i=1}^{n} IC_n(o_i).$$

For the TMLE we have $$\bar{IC}_n = 0$$. Using this mean, we calculated the sample variance of the estimated influence curve values:

$$S^2(IC_n) = \frac{1}{n} \sum_{i=1}^{n} (IC_n(o_i) - \bar{IC}_n)^2.$$ 

Lastly, we used our sample variance to estimate the standard error of our estimator:

$$\sigma_n = \sqrt{\frac{S^2(IC_n)}{n}}.$$ 

This estimate of the standard error in the mortality study was $$\sigma_n = 0.012$$. 
Sonoma Study: Inference (CIs)

\[ \psi_{TMLE,n} \pm z_{0.975} \frac{\sigma_n}{\sqrt{n}}, \]

where \( z_{\alpha} \) denotes the \( \alpha \)-quantile of the standard normal density \( N(0,1) \).
A $p$-value for $\psi_{TMLE,n}$ can be calculated as:

$$2 \left[ 1 - \Phi \left( \left| \frac{\psi_{TMLE,n} \sigma_n}{\sqrt{n}} \right| \right) \right],$$

where $\Phi$ denotes the standard normal cumulative distribution function.

The $p$-value was $< 0.001$ and the confidence interval was $[-0.078, -0.033]$. 
The interpretation of our estimate $\psi_{TMLE,n} = -0.055$, under causal assumptions, is that meeting or exceeding recommended levels of LTPA decreases 5-year mortality in an elderly population by 5.5 percentage points.

This result was significant, with a $p$-value of $< 0.001$ and a confidence interval of $[-0.078, -0.033]$. 
Example: TMLE with Missingness

SCM for a point treatment data structure with missing outcome

\[ W = f_W(U_W), \]
\[ A = f_A(W, U_A), \]
\[ \Delta = f_A(W, A, U_\Delta), \]
\[ Y = f_Y(W, A, \Delta, U_Y). \]

We can now define counterfactuals \( Y_{1,1} \) and \( Y_{0,1} \) corresponding with interventions setting \( A \) and \( \Delta \).

The additive causal effect \( EY_{1} - EY_{0} \) equals:
\[
\psi(P) = E[E(Y | A = 1, \Delta = 1, W) - E(Y | A = 0, \Delta = 1, W)]
\]
Example: TMLE with Missingness

Our first step is to generate an initial estimator of $P^0_n$ of $P$; we estimate $E(Y \mid A, \Delta = 1, W)$, possible with super learning.

We fluctuate this initial estimator with a logistic regression:

$$\text{logit} P^0_n(\epsilon)(Y = 1 \mid A, \Delta = 1, W) = \text{logit} P^0_n(Y = 1 \mid A, \Delta = 1, W) + \epsilon h$$

where

$$h(A, W) = \frac{1}{\Pi(A, W)} \left( \frac{A}{g(1 \mid W)} - \frac{1 - A}{g(0 \mid W)} \right)$$

and

$$g(1 \mid W) = P(A = 1 \mid W) \text{ Treatment Mechanism}$$

$$\Pi(A, W) = P(\Delta = 1 \mid A, W) \text{ Missingness Mechanism}$$

Let $\epsilon_n$ be the maximum likelihood estimator and

$$P^*_n = P^0_n(\epsilon_n).$$

The TMLE is given by $\Psi(P^*_n)$. 
Plan Payment Risk Adjustment

Over 50 million people in the United States currently enrolled in an insurance program that uses risk adjustment.

- Redistributes funds based on health
- Encourages competition based on efficiency/quality

**Results**

- Machine learning finds novel insights
- Potential to impact policy, including diagnostic upcoding and fraud

Rose (2016)
Plan Payment Risk Adjustment: Key Results

1. **Super Learner had best performance.**

2. **Top 5 algorithms with reduced set of variables retained 92% of the relative efficiency of their full versions (86 variables).**
   - age category 21-34
   - all five inpatient diagnoses categories
     - heart disease
     - cancer
     - diabetes
     - mental health
     - other inpatient diagnoses
   - metastatic cancer
   - stem cell transplantation/complication
   - multiple sclerosis
   - end stage renal disease

But what if we care about the individual impact of medical condition categories on health spending?
Evaluate how much more enrollees with each medical condition cost after controlling for demographic information and other medical conditions.
Evaluate how much more enrollees with each medical condition cost after controlling for demographic information and other medical conditions.

**Health Tracking**

**Trends**

National Health Spending By Medical Condition, 1996–2005

Mental disorders and heart conditions were found to be the most costly.

by Charles Roehrig, George Miller, Craig Lake, and Jenny Bryant

**ABSTRACT:** This study responds to recent calls for information about how personal health expenditures from the National Health Expenditure Accounts are distributed across medical conditions. It provides annual estimates from 1996 through 2005 for thirty-two conditions mapped into thirteen all-inclusive diagnostic categories. Circulatory system spending was highest among the diagnostic categories, accounting for 17 percent of spending in 2005. The most costly conditions were mental disorders and heart conditions. Spending growth rates were lowest for lung cancer, chronic obstructive pulmonary disease, pneumonia, coronary heart disease, and stroke, perhaps reflecting benefits of preventive care. ([Health Affairs 28, no. 2 (2009): w358–w367 (published online 24 February 2009; 10.1377/hlthaff.28.2.358])
Evaluate how much more enrollees with each medical condition cost after controlling for demographic information and other medical conditions.

**TMLE Example: Impact of Medical Conditions**

Evaluate how much more enrollees with each medical condition cost after controlling for demographic information and other medical conditions.

**Trends**

**National Health Spending By Medical Condition, 1996–2005**

Mental disorders and heart conditions were found to be the most costly.

_by Charles Roehrig, George Miller, Craig Lake, and Jenny Bryant_

**ABSTRACT:** This study responds to recent calls for information about how personal health expenditures from the National Health Expenditure Accounts are distributed across medical conditions. It provides annual estimates from 1996 through 2005 for thirty-two conditions mapped into thirteen all-inclusive diagnostic categories. Circulatory system spending was highest among the diagnostic categories, accounting for 17 percent of spending in 2005. The most costly conditions were mental disorders and heart conditions. Spending growth rates were lowest for lung cancer, chronic obstructive pulmonary disease, pneumonia, coronary heart disease, and stroke, perhaps reflecting benefits of preventive care. [Health Affairs 28, no. 2 (2009): w358–w367 (published online 24 February 2009; 10.1377/hlthaff.28.2.358)]

**Which Medical Conditions Account For The Rise In Health Care Spending?**

The fifteen most costly medical conditions accounted for half of the overall growth in health care spending between 1987 and 2000.

_by Kenneth E. Thorpe, Curtis S. Florence, and Peter Joski_

**ABSTRACT:** We calculate the level and growth in health care spending attributable to the fifteen most expensive medical conditions in 1987 and 2000. Growth in spending by medical condition is decomposed into changes attributable to rising cost per treated case, treated prevalence, and population growth. We find that a small number of conditions account for most of the growth in health care spending—the top five medical conditions accounted for 31 percent. For four of the conditions, a rise in treated prevalence, rather than rising treatment costs per case or population growth, accounted for most of the spending growth.
TMLE Example: Impact of Medical Conditions

- **Truven MarketScan** database, those with continuous coverage in 2011-2012; 10.9 million people. Variables: age, sex, region, procedures, expenditures, etc.
- Enrollment and claims from private health plans and employers.
- Extracted random sample of 1,000,000 people.
- Enrollees were eligible for insurance throughout this entire 24 month period and thus there is no drop-out due to death.
TMLE Example: Impact of Medical Conditions

Sex and Location

- Female: 40%
- Metropolitan: 80%

Age

- 21 to 34: 20%
- 35 to 54: 40%
- 55+: 40%

Region

- Northeast: 15%
- Midwest: 30%
- South: 40%
- West: 10%

Inpatient Diagnoses

- Heart Disease: 10%
- Cancer: 8%
- Diabetes: 4%
- Other: 8%

n=1,000,000
TMLE Example: Impact of Medical Conditions

Medical Condition Categories

- Stroke
- Acute Myocardial Infarction
- Lung, Brain & Severe Cancers
- Metastatic Cancer
- Chronic Skin Ulcer
- Lung Fibrosis
- Acute Ischemic Heart Disease
- Intestinal Obstruction
- Chronic Hepatitis
- Non-Hodgkin's Lymphomas
- Sepsis
- HIV/AIDS
- Pulmonary Embolism
- Multiple Sclerosis
- Hematological Disorders
- Pancreatic Disorders & Intestinal Malabsorption
- Thyroid Cancer & Melanoma
- Lupus
- Colorectal, Breast (Age <50) & Kidney Cancer
- Seizure Disorders
- Inflammatory Bowel Disease
- Congestive Heart Failure
- Rheumatoid Arthritis
- Heart Arrhythmias
- Breast (Age 50+) & Prostate Cancer
- Major Depression & Bipolar

Percent

n=1,000,000
$\psi = E_{W,M^-}[E(Y \mid A = 1, W, M^-) - E(Y \mid A = 0, W, M^-)]$,

represents the effect of $A = 1$ versus $A = 0$ after adjusting for all other medical conditions $M^-$ and baseline variables $W$.

### Interpretation

The difference in total annual expenditures when enrollees have the medical condition under consideration (i.e., $A = 1$).

$Y=$total annual expenditures, $A=$medical condition category of interest
Leverage

- available big data
- novel machine learning tools
to improve conclusions and policy insights

Rose (2017)
First investigation of the impact of medical conditions on health spending as a variable importance question using double robust estimators.

Five most expensive medical conditions were

1. multiple sclerosis
2. congestive heart failure
3. lung, brain, and other severe cancers
4. major depression and bipolar disorders
5. chronic hepatitis.

- Differing results compared to parametric regression.
- What does this mean for incentives for prevention and care?
Effect of Drug-Eluting Stents

Expected Outcome by Stent

1-Year MACE %

A1
n = 709
C1
1840
B1
1273
A4
4518
A3
622
B2
70
C4
31
C2
72
A2
640
C3
227

TMLE
MLE
Ridge
RF
Truth

Rose and Normand (2017)
Hospital Profiling

Spertus et al. (2016)
Effect Estimation Literature

- Maximum-Likelihood-Based Estimators: g-formula, Robins 1986
- Additional bibliographic history found in Chapter 1 of van der Laan and Robins 2003.
- For even more references, see Chapter 4 of Targeted Learning.
[TMLE Example Code]
TMLE Packages

- `tmle` (Gruber): Main point-treatment TMLE package
- `ltmle` (Schwab): Main longitudinal TMLE package
- SAS code (Brooks): Github
- Julia code (Lendle): Github

More: targetedlearningbook.com/software
[TMLE Example Code]
TMLE Sample Code

```r
##Code lightly adapted from Schuler & Rose, 2017, AJE##
library(tmle)
set.seed(1)
N <- 1000
```
##Generate simulated data##

#X1=Gender; X2=Therapy; X3=Antidepressant use

X1 <- rbinom(N, 1, prob=.55)
X2 <- rbinom(N, 1, prob=.30)
X3 <- rbinom(N, 1, prob=.25)
W <- cbind(X1,X2,X3)

#Exposure=regular physical exercise
A <- rbinom(N, 1, plogis(-0.5 + 0.75*X1 + 1*X2 + 1.5*X3))

#Outcome=CES-D score
Y <- 24-3*A+3*X1-4*X2-6*X3-1.5*A*X3+rnorm(N,mean=0,sd=4.5)
## Examine simulated data ##

```r
data <- data.frame(cbind(A,X1,X2,X3,Y))
summary(data)
barplot(colMeans(data[,1:4]))
```

# Examine simulated data #

```r
data <- data.frame(cbind(A,X1,X2,X3,Y))
summary(data)
barplot(colMeans(data[,1:4]))
```
```r
> summary(data)

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>X1</td>
<td>X2</td>
</tr>
<tr>
<td>Min.</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Median</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Mean</td>
<td>0.632</td>
<td>0.568</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Max.</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>X3</td>
<td>Y</td>
</tr>
<tr>
<td>Min.</td>
<td>0.000</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>0.000</td>
</tr>
<tr>
<td>Median</td>
<td>0.000</td>
</tr>
<tr>
<td>Mean</td>
<td>0.254</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>1.000</td>
</tr>
<tr>
<td>Max.</td>
<td>1.000</td>
</tr>
<tr>
<td>Min.</td>
<td>0.9629</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>16.4549</td>
</tr>
<tr>
<td>Median</td>
<td>21.2744</td>
</tr>
<tr>
<td>Mean</td>
<td>20.8369</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>25.2316</td>
</tr>
<tr>
<td>Max.</td>
<td>39.8796</td>
</tr>
</tbody>
</table>
```
TMLE Sample Code
##Specify a library of algorithms##

SL.library <- c("SL.glm","SL.step.interaction","SL.glmnet", "SL.randomForest","SL.gam","SL.rpart" )
Could use various forms of "screening" to consider differing variable sets

```r
SL.library <- list(c("SL.glm","screen.randomForest", "All"),
                   c("SL.mean", "screen.randomForest", "All"),
                   c("SL.randomForest", "screen.randomForest", "All"),
                   c("SL.glmnet", "screen.randomForest","All"))
```

Or the same algorithm with different tuning parameters

```r
SL.glmnet.alpha0 <- function(..., alpha=0){
  SL.glmnet(..., glmnet.alpha=alpha)}
SL.glmnet.alpha50 <- function(..., alpha=.50){
  SL.glmnet(..., glmnet.alpha=alpha)}
```

```r
SL.library <- c("SL.glm","SL.glmnet", "SL.glmnet.alpha50",
                "SL.glmnet.alpha0","SL.randomForest")
```
## Specify a library of algorithms ##

SL.library <- c("SL.glm","SL.step.interaction","SL.glmnet", 
                "SL.randomForest","SL.gam","SL.rpart")
++TMLE approach: Super Learning++

tmleSL1 <- tmle(Y, A, W,
              Q.SL.library = SL.library, g.SL.library = SL.library)
tmleSL1
> tmleSL1 <- tmle(Y, A, W, Q.SL.library = SL.library, g.SL.library = SL.library)
Loading required package: gam
Loading required package: splines
Loading required package: foreach
foreach: simple, scalable parallel programming from Revolution Analytics
Use Revolution R for scalability, fault tolerance and more.
http://www.revolutionanalytics.com
Loaded gam 1.14

Loading required package: glmnet
Loading required package: Matrix
Loaded glmnet 2.0-2

Loading required package: randomForest
randomForest 4.6-12
Type rfNews() to see new features/changes/bug fixes.
Loading required package: rpart
> tmleSL1
Additive Effect
Parameter Estimate: -3.4074
Estimated Variance: 0.11084
p-value: <2e-16
95% Conf Interval: (-4.0599, -2.7549)

True value is -3.38
TMLE Sample Code

```r
##TMLE approach: GLM, MT misspecification of outcome##
#Misspecified outcome regression:  Y ~ A + X1 + X2 + X3#

tmleGLM1 <- tmle(Y, A, W, Qform=Y~A+X1+X2+X3, gform=A~X1+X2+X3)
tmleGLM1
```
TMLE Sample Code

```r
> tmleGLM1 <- tmle(Y, A, W, Qform=Y~A+X1+X2+X3,
gform=A~X1+X2+X3)
> tmleGLM1
Additive Effect
  Parameter Estimate:  -3.416
  Estimated Variance:  0.1132
  p-value:            <2e-16
  95% Conf Interval:  (-4.0754, -2.7565)
```

True value is -3.38
### TMLE approach: GLM, OV misspecification of outcome###

Misspecified outcome regression: $Y \sim A + X_1 + X_2$

```r
tmleGLM2 <- tmle(Y, A, W, Qform=Y~A+X1+X2, gform=A~X1+X2+X3)
tmleGLM2
```
TMLE Sample Code

> tmleGLM2 <- tmle(Y, A, W, Qform=Y~A+X1+X2, gform=A~X1+X2+X3)
> tmleGLM2

Additive Effect
Parameter Estimate:  -3.3976
Estimated Variance:  0.1416
p-value:  <2e-16
95% Conf Interval:  (-4.1351,  -2.6601)

True value is -3.38
TMLE Sample Code

##TMLE approach: GLM, OV misspecification of exposure##
#Misspecified exposure regression:  A ~ X1 + X2#

tmleGLM3 <- tmle(Y, A, W, Qform=Y~A+X1+X2+X3+A:X3, gform=A~X1+X2)
tmleGLM3
> tmleGLM3 <- tmle(Y, A, W, Qform=Y~A+X1+X2+X3+A:X3, gform=A~X1+X2)
> tmleGLM3

Additive Effect
Parameter Estimate:  -3.4277
Estimated Variance:  0.10156
p-value:  <2e-16
95% Conf Interval:  (-4.0524,  -2.8031)

True value is -3.38
Researchers have recently implemented super learning for propensity score estimation (40, 41). Fundamentally, TMLE and the other estimators discussed can be implemented using machine learning algorithms, which can prove advantageous in complex observational data. As our simulation results demonstrate, super learning performed better than or equal to parametric regression for all 3 estimators. The ability of super learning to protect against certain types of functional form misspecification is demonstrated by the G-computation results, as super learning yielded smaller bias than parametric regression with only main terms. Given the complexity of data in typical observational studies, correct specification of all parametric regressions is unlikely, yet bias can arise from even minor functional form misspecification. Machine learning algorithms, particularly ensemble methods such as super learning, can empirically identify interaction, nonlinear, and higher-order relationships among variables; therefore, the corresponding ATE estimate is less likely to be biased due to a misspecified functional form in comparison with main-terms parametric regression. Additionally, while TMLE with super learning and parametric regression performed equivalently in our simulation study, TMLE with super learning may outperform parametric regression in cases of more complex data (14, 39, 41).

Implementation of TMLE is conceptually similar to and shares estimation steps with other statistical methods used for causal estimation in epidemiologic research. TMLE has attractive statistical properties, particularly double robustness. Additionally, it is straightforward to implement, as open-source software is available (32, 45). As a flexible estimation method that can easily incorporate nonparametric machine learning methods, TMLE is another advanced tool that can be added to the applied practitioner’s statistical toolbox.

ACKNOWLEDGMENTS

Author affiliations: Department of Health Care Policy, Harvard Medical School, Boston, Massachusetts (Megan S. Schuler, Sherri Rose).

This work was generously supported by the Marshall J. Seidman Center for Studies in Health Economics and Health Care Policy at Harvard Medical School (M.S.) and the University of Utah (grant P0163947 to S.R.).

Conflict of interest: none declared.

Schuler and Rose (2017)
## TMLE Sample Code

### G-Computation

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Outcome Variables</th>
<th>Estimate (SE)</th>
<th>p-value</th>
<th>Lower, Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Super learner</td>
<td>$A, X_1, X_2, X_3$</td>
<td>$-3.27 (0.35)$</td>
<td>0.11</td>
<td>$-3.98, -2.56$</td>
</tr>
<tr>
<td>Misspecified parametric regression</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Main-terms misspecification</td>
<td>$A, X_1, X_2, X_3$</td>
<td>$-3.25 (0.33)$</td>
<td>0.13</td>
<td>$-3.91, -2.59$</td>
</tr>
<tr>
<td>Omitted-variable misspecification</td>
<td>$A, X_1, X_2$</td>
<td>$-4.98 (0.37)$</td>
<td>-1.60</td>
<td>$-5.69, -4.24^b$</td>
</tr>
</tbody>
</table>

### Inverse Probability Weighting

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Exposure Variables</th>
<th>Estimate (SE)</th>
<th>p-value</th>
<th>Lower, Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Super learner</td>
<td>$X_1, X_2, X_3$</td>
<td>$-3.43 (0.37)$</td>
<td>-0.05</td>
<td>$-4.17, -2.63$</td>
</tr>
<tr>
<td>Misspecified parametric regression</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Omitted-variable misspecification</td>
<td>$X_1, X_2$</td>
<td>$-4.96 (0.37)$</td>
<td>-1.58</td>
<td>$-5.67, -4.21^b$</td>
</tr>
</tbody>
</table>

Schuler and Rose (2017)
TMLE Packages

- `tmle` (Gruber): Main point-treatment TMLE package
- `ltmle` (Schwab): Main longitudinal TMLE package
- SAS code (Brooks): Github
- Julia code (Lendle): Github

More: targetedlearningbook.com/software
[Q & A]