**eAppendix 1: Step-by-step implementation of TMLE**

We describe the four steps of implementing TMLE in detail, although similar procedures can be also found in the literature.1,2

Step 1: Estimate an initial probability function of *Y* given *A* and ***W***, denoted as . The standard logistic regression model is one possible approach:

Therefore, the initial probability can be estimated by:

(1)

Any terms that are functions of *A* and/or ***W*** can be included in the model, for example, polynomial terms of *A* and ***W*** as well as the interaction terms of *A* and ***W*** can be considered. Consequently, for each subject, the predicted probabilities for both counterfactual outcomes and can be attained as following by setting *A*=0 and *A*=1 for everyone respectively:

Step 2: Estimate a probability function of *A* given ***W***, denoted as . A logistic regression model can again be used:

(2)

The probability of *A* given ***W*** is then estimated by:

A logistic model is used in step 1 and step 2 for demonstration. In fact, any suitable modeling approach can be employed such as other parametric models, semi- or non-parametric approaches including advanced machine learning techniques.3

Step 3: Updating the density of *Y* given *A* and ***W***.

This step aims to find a better prediction model targeted at minimizing mean squared error for the estimation of and by using the so called efficient influence function of and .4 For and , it has been shown that the parametric working model satisfied specific required conditions for the so-called fluctuation parameters and .2

The model is defined as following:

Where , and are referred to as clever covariates, and and are estimated from step 2. The indicator function *I*(·) takes value one if its Boolean argument is true and value zero otherwise. The above parametric model is fitted with maximum likelihood estimation to obtain estimates for the fluctuation parameters (). This can be done with standard software by setting as an offset in an intercept-free logistic regression with covariates and .  
Explicitly: (3)

By substituting (), the estimated probability of *Y* given *A* and ***W*** and the probability of the counterfactual outcomes for each subject can be updated as follows:

so that,

This updating is performed by setting *A*=0 and *A*=1 for each subject in the probability functions , , as well as in the clever covariates and . In general, TMLE is an iterative procedure, where is replaced with , and is updated until convergence (until becomes sufficiently small), where the subscript denotes the th iteration. However, for the parameters and in this case, convergence is achieved in one step. Therefore is the only update needed. Model (3) targets and simultaneously by including both and in the model. Alternatively, and can be targeted separately by including the offset and or and in two different models. However, estimating and separately doubles this modeling step, and is thus less computationally favorable.

It is also possible to target RD, RR, and OR directly by applying the corresponding efficient influence functions. With this approach, iterative updates may be needed to achieve the convergence of when targeting RR and OR directly.

Step 4: Computing the substitution estimator of .

TMLE of and are given by the G-computation formula:5

and .

In this case, the G-computation formula is implemented simply by averaging the estimated individual predicted probabilities from the respective updated outcome model .

The RD, RR, and OR (in the log scale) are then estimated by:

**eAppendix 2: Standard errors, confidence intervals, and null-hypothesis testing in TMLE framework**

TMLE for the population parameter are asymptotically normally distributed with mean and variance , where is the variance of the influence function (denoted as for : . TMLE is asymptotically linear under the assumptions .14 The corresponding Wald-type 95% confidence interval is given by: . The causal null hypothesis : can be tested with the statistics: .6

Efficient influence functions are parameter specific and take the following values for the three effect measure statistics of interest, where denotes the last updated :

**eAppendix 3: SAS routine for a general implementation of TMLE in binary point exposure and outcome study**

/\* Specify the analysis dataset in option: Dataset=

Specify the name of the binary exposure variable in option: Var\_exposure=

Specify the name of the binary outcome variable in option: Var\_outcome=

List the covariates for the treatment model in option: Var\_ps=

List the treatment variable and the covariates for outcome model in the

option: Var\_out= \*/

**%macro** estimation\_TMLE(Dataset=,Var\_exposure=,Var\_outcome=,Var\_ps=,Var\_out=);

\* Propensity score estimation – Treatment model in TMLE;

proc logistic data=&Dataset descending noprint;

model &Var\_exposure=&Var\_ps;

output out=&Dataset(drop=\_level\_) pred=ps;

run;

\* Compute clever covariates;

data &Dataset;

set &Dataset end=eof;

H1AW=&Var\_exposure/ps;

H1W=**1**/ps;

H0AW=(**1**-&Var\_exposure)/(**1**-ps);

H0W=**1**/(**1**-ps);

if eof=**1** then call symput('n',\_n\_);

run;

\* Outcome model in TMLE;

proc logistic data=&Dataset descending;

model &Var\_outcome=&Var\_out;

store Q\_model;

output out=&Dataset xbeta=logit\_Y;

run;

\* Estimate fluctuation parameters;

ods output ParameterEstimates=epsilon;

proc logistic data=&Dataset descending;

model &Var\_outcome=H1AW H0AW/offset=logit\_Y noint;

quit;

\* Compute the probability of the counterfactual outcomes;

data A\_1;

set &Dataset;

&Var\_exposure=**1**;

run;

data A\_0;

set &Dataset;

&Var\_exposure=**0**;

run;

proc plm source=Q\_model noprint;

score data=&Dataset out=&Dataset predicted=logit\_Q\_A;

run;

proc plm source=Q\_model noprint;

score data=A\_0 out=A\_0 predicted=logit\_Q\_A0;

run;

proc plm source=Q\_model noprint;

score data=A\_1 out=A\_1 predicted=logit\_Q\_A1;

run;

data &Dataset;

merge &Dataset A\_1(keep=id logit\_Q\_A1) A\_0(keep=id logit\_Q\_A0);

by id;

run;

data epsilon1;

set epsilon;

where variable ='H1AW';

run;

data epsilon0;

set epsilon;

where variable ='H0AW';

run;

\* Update the probability of the counterfactual outcomes;

\* Estimate the parameter of interest, e.g. odds ratio;

\* Estimate the standard error and 95% confidence interval;

proc iml;

use epsilon1; read all var{estimate} into epsilon1;

use epsilon0; read all var{estimate} into epsilon0;

use &Dataset;

read all var{H1AW} into H1AW;

read all var{H0AW} into H0AW;

read all var{H1W} into H1W;

read all var{H0W} into H0W;

read all var{&Var\_exposure} into E;

read all var{&Var\_outcome} into Y;

read all var{ps} into ps;

read all var{logit\_Q\_A} into logit\_Q\_A;

read all var{logit\_Q\_A0} into logit\_Q\_A0;

read all var{logit\_Q\_A1} into logit\_Q\_A1;

Q1\_1W=epsilon1\*H1W+logit\_Q\_A1;

Q1\_0W=epsilon0\*H0W+logit\_Q\_A0;

Q1\_AW=epsilon1\*H1AW+logit\_Q\_A;

Y1=exp(Q1\_1W)/(**1**+exp(Q1\_1W));

Y0=exp(Q1\_0W)/(**1**+exp(Q1\_0W));

YA=exp(Q1\_AW)/(**1**+exp(Q1\_AW));

u1=mean(Y1);

u0=mean(Y0);

\*log RD estimation;

effect="log RD";

est=u1-u0;

IC=(E/ps-(**1**-E)/(**1**-ps))#(Y-YA)+Y1-Y0-est;

Variance=var(IC)/&n;

create TMLE\_EST var{effect est Variance};

append;

\*log RR estimation;

edit TMLE\_EST;

effect="log RR";

est=log(u1)-log(u0);

IC=(**1**/u1)#(E/ps#(Y-YA)+Y1-u1)-(**1**/u0)#((**1**-E)/(**1**-ps)#(Y-YA)+Y0-u0);

Variance=var(IC)/&n;

append;

\*log OR estimation;

edit TMLE\_EST;

effect="log OR";

est=log(u1/(**1**-u1))-log(u0/(**1**-u0));

IC=(**1**/u1+**1**/(**1**-u1))#(E/ps#(Y-YA)+Y1)-(**1**/u0+**1**/(**1**-u0))#((**1**-E)/(**1**-ps)#(Y-YA)+Y0);

Variance=var(IC)/&n;

append;

quit;

data TMLE\_EST;

length var\_ps $200.;

length var\_out $200.;

var\_ps="&var\_ps";

var\_out="&var\_out";

set TMLE\_EST;

Stderr=sqrt(Variance);

LowerCL=est-probit(**0.975**)\*sqrt(Variance);

UpperCL=est+probit(**0.975**)\*sqrt(Variance);

Method='TMLE';

rename est=Estimate;

run;

**%mend**;

**eAppendix 4:** **Results of the sensitivity analysis with propensity score truncated at the 1st and 99th percentile**

eTable 1 Effect estimate of statins use post-MI on 1-year all-cause mortality by different TMLE and IPW estimator modeling approaches with truncated propensity score at the 1st and 99th percentile

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Outcome Model | *A* | | | | | | *A*, ***W***1 | | | *A*, ***W***2 | | | *A*, ***W***3 | | |
| Method | IPW1-4 | | | TMLE1-4 | | | TMLE5-8 | | | TMLE9-12 | | | TMLE13-16 | | |
| Treatment Model | log(OR) | SE(log(OR)) | OR | log(OR) | SE(log(OR)) | OR | log(OR) | SE(log(OR)) | OR | log(OR) | SE(log(OR)) | OR | log(OR) | SE(log(OR)) | OR |
| NULL | -1.12 | 0.04 | 0.30 | -1.2 | 0.01 | 0.31 | -0.85 | 0.04 | 0.43 | -0.57 | 0.04 | 0.57 | -0.32 | 0.04 | 0.73 |
| ***W***1a | -0.84 | 0.04 | 0.43 | -0.84 | 0.04 | 0.44 | -0.84 | 0.04 | 0.43 | -0.56 | 0.04 | 0.57 | -0.32 | 0.04 | 0.73 |
| ***W***2b | -0.47 | 0.08 | 0.62 | -0.49 | 0.09 | 0.61 | -0.51 | 0.08 | 0.60 | -0.58 | 0.07 | 0.56 | -0.36 | 0.07 | 0.70 |
| ***W***3c | -0.14 | 0.09 | 0.87 | -0.23 | 0.11 | 0.80 | -0.22 | 0.01 | 0.80 | -0.34 | 0.09 | 0.71 | -0.37 | 0.08 | 0.69 |

a***W***1: Predefined important confounders (age, sex, obesity, smoking, history of diabetes);

b***W***2: All pre-specified confounders;

c***W***3: All potential confounders (*W*2 + the 400 empirical selected variables).

REFERENCES

1. Laan MJ van der, Rose S. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer Science & Business Media; 2011.

2. Gruber S, van der Laan MJ. tmle: An R Package for Targeted Maximum Likelihood Estimation. 2011. http://biostats.bepress.com/ucbbiostat/paper275/. Accessed May 22, 2014.

3. Van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Statistical Applications in Genetics and Molecular Biology.* 2007;6(1): Article 25.

4. Moore KL, van der Laan MJ. Covariate adjustment in randomized trials with binary outcomes: Targeted maximum likelihood estimation. *Stat Med*. 2009;28(1):39-64.

5. Robins J. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Math Model*. 1986;7(9):1393-1512.

6. Rosenblum M, van der Laan MJ. Targeted maximum likelihood estimation of the parameter of a marginal structural model. *Int J Biostat*. 2010;6(2): Article 19.