# Targeted Maximum Likelihood Super Learning

## Applications in Genomics, RCT, and Observational Studies

Mark van der Laan
works.bepress.com/mark_van_der_laan
*Division of Biostatistics,*
*University of California, Berkeley*

Biopharmaceutical Section of the American Statistical Association
Web-based Training Series
*August 24, 2009*

# Statistics are a tool of modern life

- Identifying associations, correlations and patterns
- Establishing causation based on randomized trials and observational studies.
- Making predictions
- Shaping strategies and future behavior
  - Of people and societies
  - Of machines and computing
  - Of complex systems

# Bias is a hazard of statistics

- Statistical *data samples* can be biased
  - The sample selected does not represent the population
  - Example: There are five red heads in a town of 100 people. Our sample of 20 people happens to include all five.
- Statistical *methods* for learning from data can be biased
  - The statistical model selected is not the one that best fits the data…
  - … for the question being asked!
- Statistical interpretations of findings can be biased.

# Statistics are often cited to denote a certainty that *does not, in fact, exist*

- "There is increasing concern that in modern research, false findings may be the majority, or even the vast majority, of published research claims."
  - "Simulations show that for most study designs and settings it is more likely for a research claim to be false than true."
  - "For many current scientific fields, claimed research findings often may be simply accurate measures of the prevailing ***bias***."


- J.P.A. Ioannidis, "Why Most Published Research Findings are False", *Chance*, Vol. 18, No. 4, 2005

## Why Most Published Research Findings Are False

John P. A. Ioannidis

**Corollary 2: The smaller the effect sizes in a scientific field, the less likely the research findings are to be true.** Power is also related to the effect size. Thus research findings are more

**Corollary 3: The greater the number and the lesser the selection of tested relationships in a scientific field, the less likely the research findings are to be true.** As shown above, the post-study probability that a finding is true (PPV)

**Corollary 4: The greater the flexibility in designs, definitions, outcomes, and analytical modes in a scientific field, the less likely the research findings are to be true.** Flexibility increases the potential for

**Corollary 6: The hotter a scientific field (with more scientific teams involved), the less likely the research findings are to be true.**

**Claimed Research Findings May Often Be Simply Accurate Measures of the Prevailing Bias**

As shown, the majority of modern biomedical research is operating in areas with very low pre- and post-study probability for true findings.

# False conclusions are expensive

- In medicine
  - False positives lead to expensive additional tests and anxiety
  - False negatives lead to delayed treatment with escalated costs and illness
- In drug discovery
  - False positives lead to failed trials
    - The average cost of a phase III clinical trial is $4m-$20m, some cost more than $100m
  - False negatives lead to failed trials
    - Missed contraindications, negative interactions and imprecise dosages
- In genomics, proteomics and chemoinformatics
  - False positives are abundant and lead to wasted time, effort and experimentation
  - False negatives lead to missed business opportunities
- In public policy
  - False positives and false negatives lead to action based on false premises and, frequently, public cynicism

## Government mandates and programs

- FDA Critical Path Initiative
  - Better evaluation tools
  - Streamlining clinical trials
  - Harnessing bioinformatics
- FDA Amendment Act of 2007
  - Post-market safety trials and studies
  - Requires FDA to develop systems and analyses
  - Risk evaluation and mitigation
  - Safety labeling
- FDA Chief Scientist's Challenge Grants
  - Biomarkers
  - Personalized medicine
  - Clinical trials design and analysis
  - Predicting safety and efficacy

## Solution

- Avoid reliance on human art and reliance on misspecified parametric models
- Adapt the model to fit the data
- Target the fit to the parameter of interest

TMLE/SL

Targeted Maximum Likelihood coupled with Super Learner methodology

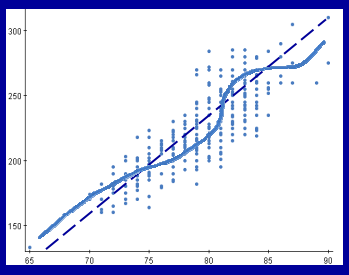
# Outline

- Two-stage Methodology: Targeted Maximum Likelihood and Super Learning
- Causal effect on survival, handling right-censoring and confounding
  - Application to Phase IV RCT
- Variable Importance Analysis in Genomic Studies
  - Biomarker Discovery
  - Effect of SNPs
  - Yeast Regulatory Network Collaborative double robust T-MLE
- Collaborative double robust T-MLE
- Causal effect in Case Control Studies

# Two-stage Methodology: TMLE/SL

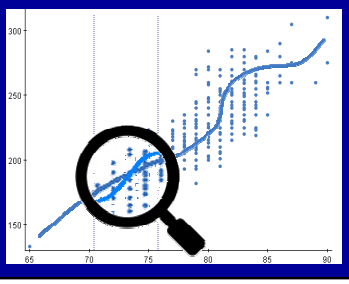
## 1. Super Learning

- Works on a library of model fits
- Builds data-adaptive composite model by assigning weights
- Weights are optimized based on loss-function specific cross-validation to guarantee best overall fit




## 2. Targeted Maximum Likelihood Estimation

- Zooms in on one aspect of the model fit—the target
- Removes bias for the target.



# TMLE/SL Toolbox

## Targeted effects

- Effect of static or dynamic treatments (e.g. on survival time)
- Direct and Indirect Effects
- Parameters of Marginal Structural Models
- Structural Nested Models

## Types of data

- Point treatment
- Longitudinal/Repeated Measures
- Censoring/Missingness/Time-dependent confounding.
- Case-Control
- Randomized clinical trials and observational data

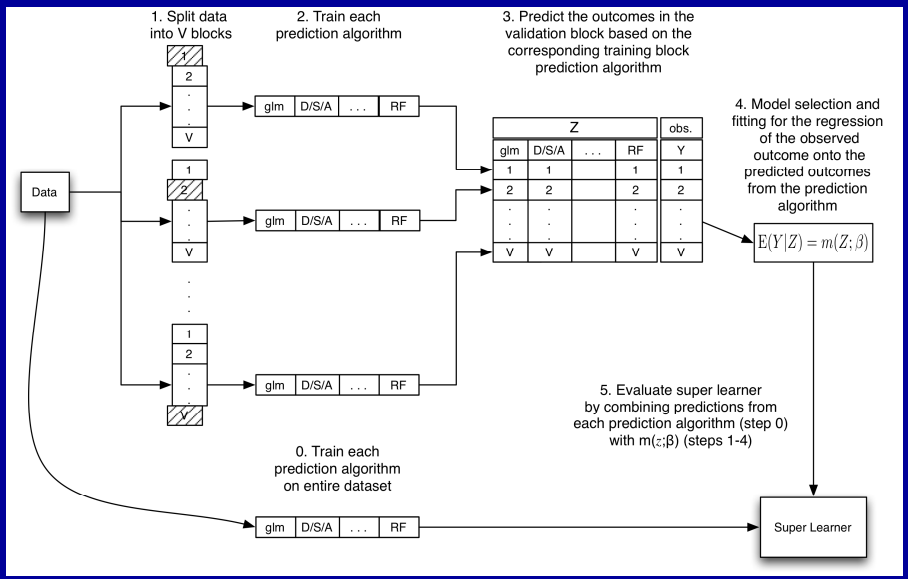


# Super Learner

# Super Learning in Semiparametric Models

- Allows one to combine many data adaptive MLEs into one improved MLE.
- Grounded by oracle results for loss-function based cross-validation (vdL&D, 2003). Loss function needs to be bounded.
- Performs asymptotically as well as best (oracle) weighted combination, or achieves parametric rate of convergence.

# Super Learner Flow Chart



# Super Learning in Prediction

| method | study 1 | study 2 | study 3 | study 4 | overall |
|---|---|---|---|---|---|
| Least Squares | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| LARS | 0.91 | 0.95 | 1.00 | 0.91 | 0.95 |
| D/S/A | 0.22 | 0.95 | 1.04 | 0.43 | 0.71 |
| Ridge | 0.96 | 0.99 | 1.02 | 0.98 | 1.00 |
| Random Forest | 0.39 | 0.72 | 1.18 | 0.71 | 0.91 |
| MARS | 0.02 | 0.82 | 0.17 | 0.61 | 0.38 |
| Super Learner | 0.02 | 0.67 | 0.16 | 0.22 | 0.19 |

Table 1: Simulation Example: Estimates of the relative mean squared prediction error (compared to least squares) based on the validation sample.

# Illustration of Super Learning in Hazard Estimation

Super Learner can be extended to prediction with right censored data.

Example: Interested in predicting survival curves for cancer patients based on baseline characteristics and treatment decisions

# Illustration of Super Learning in Hazard Estimation

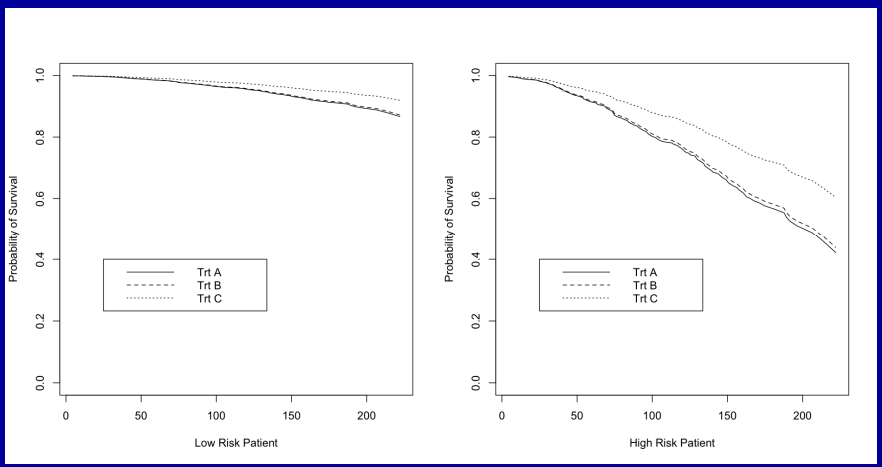
Observed data: (W,A,Y=min(T,C), $\Delta$) with

- Conditional hazard: $\lambda = \lambda(T|A,W)$
- Indicator of failure at t: dN(t) = I(Y=t, $\Delta$=1)
- Censoring survivor function: $\bar{G}(t) = \mathrm{P}(C > t)$

Loss functions for Super Learner:

- IPCW: $L(\lambda) = \Delta\{\bar{G}(T)\}^{-1}\left(\mathrm{I}(T > t_0) - S_\lambda(t_0|A,W)\right)^2$
- LogLik: $L(\lambda) = \sum_t \left(\mathrm{I}(\tilde{T} \geq t) \log \lambda^{dN(t)} \log(1-\lambda)^{(1-dN(T))}\right)$
- L2: $L(\lambda) = \sum_t \mathrm{I}(\tilde{T} \geq t)\left(dN(t) - \lambda\right)^2$

# Illustration of Super Learning in Hazard Estimation



# Targeted Maximum Likelihood

## Targeted Maximum Likelihood

- MLE/SL aims to do good job of estimating whole density
- Targeted MLE aims to do good job at parameter of interest
  - General decrease in bias for parameter of Interest
  - Fewer false positives
  - Honest p-values, inference, multiple testing

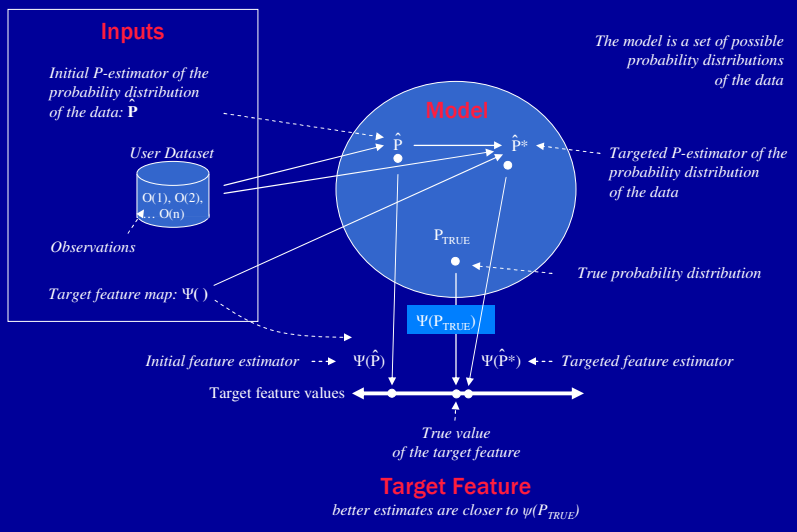
# Philosophy of Targeted Estimator

Given initial P-estimator, find updated $P^*$ in the model which gives:

- Large bias reduction for parameter of interest (target feature)
  - E.g. by requiring that it solves the efficient influence curve equation $\sum_{i=1} D^*(P)(O_i)=0$.
- Small increase of log-likelihood relative to the initial P estimator

Targeted log-likelihood loss -log p* can be used for selection.

# Targeted Maximum Likelihood Estimation Flow Chart



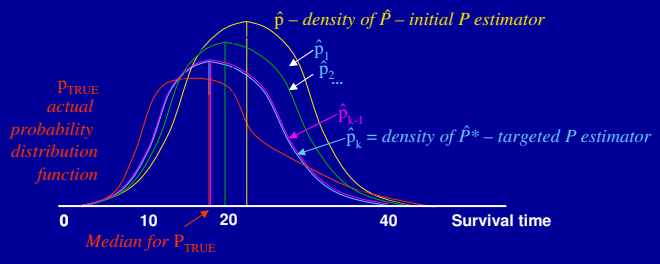
# (Iterative) Targeted MLE

1. Identify optimal strategy for "stretching" initial $\hat{P}$
   - Small "stretch" -> maximum change in target
2. Given strategy, identify optimum amount of stretch by MLE
3. Apply optimal stretch to $\hat{P}$ using optimal stretching function -> 1st-step targeted maximum likelihood estimator
4. Repeat until the incremental "stretch" is zero
   - Some important cases: 1 step to convergence
5. Final probability distribution solves efficient influence curve equation

→ (Iterative) T-MLE is double robust & locally efficient

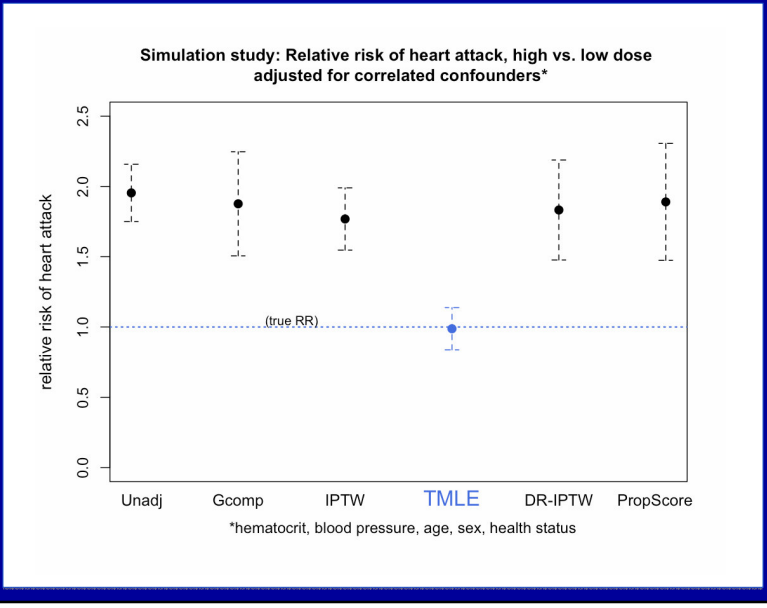
# (Iterative) Targeted MLE Estimating a Median

- Start with the initial $\hat{P}$-estimator $\hat{P}$
- Determine optimal "stretching function" and "amount of stretch", producing a new $\hat{P}$-estimator.
- Continue repeating until further stretching is essentially zero



## TMLE provides more accurate information with less data

Simulated Safety Analysis of Epogen (Amgen)




# Targeted MLE Applied to Clinical Trial or Observational Data

Impact of Treatment on Disease

# TMLE for Average Causal Effect

$$W \Longrightarrow A \Longrightarrow Y$$

- Observe predictors W, treatment A, missingness indicator Delta, and outcome Y: $(W, A, \Delta, \Delta Y)$
- Target is additive causal effect: EY(1)-Y(0)
- Regress Y on treatment A and W (e.g. Super Learning), and add clever covariate

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

where

$g(1 \mid W) = P(A = 1 \mid W)$ Treatment Mechanism
$\Pi(A,W) = P(\Delta = 1 \mid A, W)$ Missingness Mechanism

  - Then average regression over W for fixed treatment a: $E_n Y_a$
- Evaluate average effect: $E_n Y_1 - E_n Y_0$

# TMLE with Survival Outcome

- Suppose one observes baseline covariates, treatment, and one observes subject up till end of follow up or death:

$$(W, A, \Delta = I(T \leq C), \tilde{T} = \min(T, C))$$

- One wishes to estimate causal effect of treatment A on survival T
- Targeted MLE uses covariate information to adjust for confounding, informative drop out and to gain efficiency

# TMLE with Survival Outcome

- Target $\psi_1(t_0)=\Pr(T_1>t_0)$ and $\psi_0(t_0)=\Pr(T_0>t_0)$ – thereby target treatment effect, e.g.,

  1) Difference: $\Pr(T_1>t_0)$ - $\Pr(T_0>t_0)$, 2) Log RH: $\log \frac{\log \psi_1(t_0)}{\log \psi_0(t_0)}$

- Obtain initial conditional hazard fit (e.g. super learner for discrete survival) and add two time-dependent covariates

$$h_\delta(t,A,W) = \frac{I(A=\delta)}{g(A \mid W)\bar{G}(t \mid A,W)} \frac{S(t_0 \mid A,W)}{S(t \mid A,W)} I(t \le t_0)$$

  – Iterate until convergence, then use updated conditional hazard from final step, and average corresponding conditional survival over W for fixed treatments 0 and 1

# TMLE in RCT with Survival Outcome
## Difference at Fixed End Point

***Independent Censoring***

| | % Bias | Power | 95% Coverage | Relative Efficiency |
|---|---|---|---|---|
| KM | <1% | 0.79 | 0.95 | 1.00 |
| TMLE | <1% | 0.91 | 0.95 | 1.44 |

→ **TMLE: gain in power over KM**

***Informative Censoring***

| | % Bias | Power | 95% Coverage | Relative Efficiency |
|---|---|---|---|---|
| KM | 13% | 0.88 | 0.92 | 1.00 |
| TMLE | <1% | 0.92 | 0.95 | 1.50 |

→ **TMLE: unbiased**

# TMLE in Actual Phase IV RCT

- Study: RCT aims to evaluate safety based on mortality due to drug-to-drug interaction among patients with severe disease
- Data obtained with random sampling from original real RCT FDA dataset
- Goal: Estimate risk difference (RD) in survival at 28 days (0/1 outcome) between treated and placebo groups

# TMLE in Phase IV RCT

| | Unadjusted | TMLE |
|---|---|---|
| Estimate | 0.034 | 0.043 |
| p-value (RE) | 0.085 (1.000) | 0.009 (1.202) |

- TMLE adjusts for small amount of empirical confounding (imbalance in AGE covariate)
- TMLE exploits the covariate information to gain in efficiency and thus power over unadjusted
- TMLE Results significant at 0.05

## TMLE in RCT: Summary

- TMLE approach handles censoring and improves efficiency over standard approaches
  - Measure strong predictors of outcome
- Implications
  - Unbiased estimates with informative censoring
  - Improved power for clinical trials
  - Smaller sample sizes needed
  - Possible to employ earlier stopping rules
  - Less need for homogeneity in sample
    - More representative sampling
    - Expanded opportunities for subgroup analyses

## Targeted MLE Analysis of Genomic Data

Biomarker discovery, Impact of mutations on disease, or response to treatment

## The Need for Experimentation

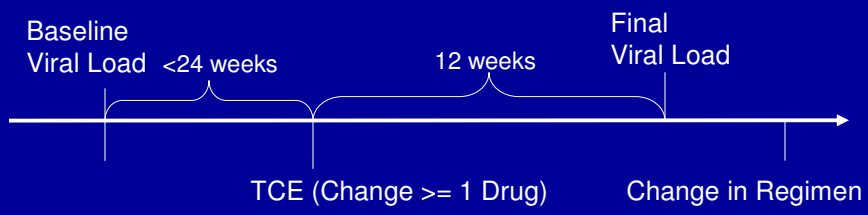
- Estimation of Variable Importance/Causal Effect requires assumption not needed for prediction
- "Experimental Treatment Assignment" (ETA)
  - Must be some variation in treatment variable A within every stratum of confounders W
    - W must not perfectly predict/determine A
    - g(a|W)>0 for all (a,W)

## Biomarker Discovery: HIV Resistance Mutations

- Goal: Rank a set of genetic mutations based on their importance for determining an outcome
  - Mutations (A) in the HIV protease enzyme
    - Measured by sequencing
  - Outcome (Y) = change in viral load 12 weeks after starting new regimen containing saquinavir
- How important is each mutation for viral resistance to this specific protease inhibitor drug?
  - Inform genotypic scoring systems

# Stanford Drug Resistance Database

- All Treatment Change Episodes (TCEs) in the Stanford Drug Resistance Database
  - Patients drawn from 16 clinics in Northern CA




- 333 patients on saquinavir regimen

# Parameter of Interest

- Need to control for a range of other covariates W
  - Include: past treatment history, baseline clinical characteristics, non-protease mutations, other drugs in regimen
- Parameter of Interest Variable Importance

  $\psi = E[E(Y|A_j=1,W)-E(Y|A_j=0,W)]$
  - For each protease mutation (indexed by j)

# Parameter of Interest

- Assuming no unmeasured confounders (W sufficient to control for confounding)
  - → Causal Effect is same as W-adjusted Variable Importance

$E(Y_1)-E(Y_0)=E[E(Y|A=1,W)-E(Y|A=0,W)]= \psi$

  - Same advantages to T-MLE

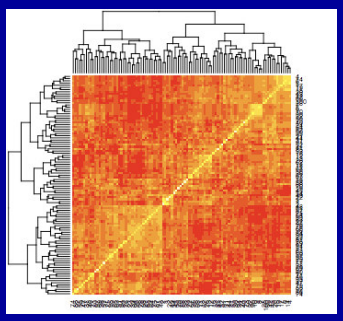
Targeted Maximum Likelihood Estimation of the Adjusted Effect of HIV Mutation on Resistance to Lopinavir

| mutation | score | estimate | 95% CI |
|---|---|---|---|
| p50V | 20 | 1.703 | ( 0.760, 2.645)* |
| p82AFST | 20 | 0.389 | ( 0.084, 0.695)* |
| p54VA | 11 | 0.505 | ( 0.241, 0.770)* |
| p54LMST | 11 | 0.369 | ( 0.002, 0.735)* |
| p84AV | 11 | 0.099 | (-0.130, 0.329) |
| p46ILV | 11 | 0.046 | (-0.222, 0.315) |
| p48VM | 10 | 0.306 | (-0.162, 0.774) |
| p47V | 10 | 0.805 | ( 0.282, 1.328)* |
| p32I | 10 | 0.544 | ( 0.312, 0.777)* |
| p90M | 10 | 0.209 | (-0.058, 0.476) |
| p82MLC | 10 | 1.610 | ( 1.330, 1.890)* |
| p84C | 10 | 0.602 | ( 0.471, 0.734)* |
| p33F | 5 | 0.300 | (-0.070, 0.669) |
| p53LY | 3 | 0.214 | (-0.266, 0.695) |
| p73CSTA | 2 | 0.635 | ( 0.278, 0.992)* |
| p24IF | 2 | 0.229 | (-0.215, 0.674) |
| p10FIRVY | 2 | -0.266 | (-0.522, -0.011)* |
| p71TVI | 2 | 0.019 | (-0.243, 0.281) |
| p30N | 0 | -0.440 | (-0.853, -0.028)* |
| p88S | 0 | -0.474 | (-0.840, -0.108)* |
| p88DTG | 0 | -0.426 | (-0.842, -0.010)* |
| p36ILVTA | 0 | 0.272 | (-0.001, 0.544) |
| p20IMRTVL | 0 | 0.178 | (-0.111, 0.467) |
| p23I | 0 | 0.822 | (-0.050, 1.694) |
| p16E | 0 | 0.239 | (-0.156, 0.633) |
| p63P | 0 | -0.131 | (-0.392, 0.131) |

Stanford mutation score, http://hivdb.stanford.edu, accessed September, 1997

# Evaluation of Biomarker Methods
## Sensitivity to Correlation

- Biological data is highly correlated
- Exploration of false positives is costly
- Compare under increasing correlation
  - Univariate Regression
  - LASSO Regression
  - randomForest (Breiman 1996, 1999)
  - tVIM using LASSO for estimation of E[Y|A,W] and E[A|W]



- Performance Assessment – distinguishing "true" variables from decoys
  - Minimum length of list necessary to find all "true" variables



# Targeted Variable Importance
## for continuous A

Parameter of Interest: $\mu(\mathrm{a}) = \mathrm{E}[E(Y \mid A=a,W) - E(Y \mid A=0,W)]$

define $m(A=a,W \mid \beta) = E(Y \mid A=a,W) - E(Y \mid A=0,W)$

$m(A=0,W \mid \beta) = 0 \quad \forall \beta, W$

- Update is obtained with least squares regression on a "clever covariate,'' h*(A,W), derived from the efficient influence curve
- For linear m(A,W|β), tMLE converges in one step

$$m(A,W \mid B) = A\beta^T W$$

$\mu(\mathrm{a}) = a\beta E[W]$ "Importance Curve"

$h^*(A,W) = W(A - E[A \mid W])$ Assuming σ(A,W)=σ(W)

- Simplest model used for simulations m(A,W|β)=Aβ

# Simulation Set-up

- Variables
  - p=100, n=300
  - Block Diagonal correlation structure: 10 independent sets of 10
  - Multivariate normal distribution
  - Constant ρ, variance=1,
  - ρ={0,0.1,0.2,0.3,…,0.9}
- Outcome
  - Main effect linear model
  - 10 "true" biomarkers, one variable from each set of 10
  - Equal coefficients
  - Noise term with mean=0 sigma=10
    - "realistic noise"

$$\Sigma_{nxp} = \begin{bmatrix} \Sigma_1 & & 0 \\ & \ddots & \\ 0 & & \Sigma_{10} \end{bmatrix}$$

$$W^*_{nxp} \sim MVN(\mu_{1xp}, \Sigma_{pxp})$$

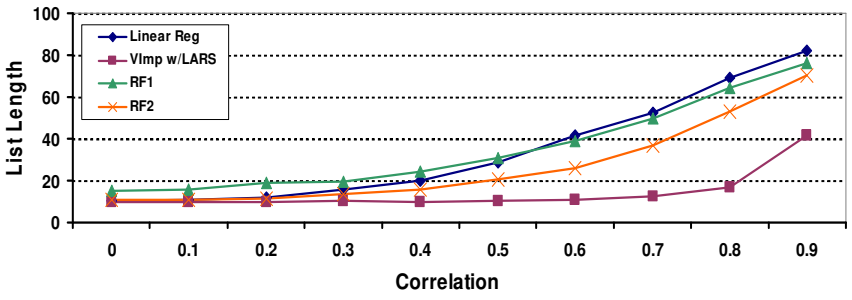
$$\mathrm{Y} = \beta X_{true} + \zeta$$

$$\mathrm{X}_{true} = \{W^*_i : i = 5, 15, \ldots, 85, 95\}$$

$$\zeta \sim N(0, 10)$$

# Evaluation of Biomarker Methods: Simulation

**Minimal List length to obtain all 10 "true" variables**



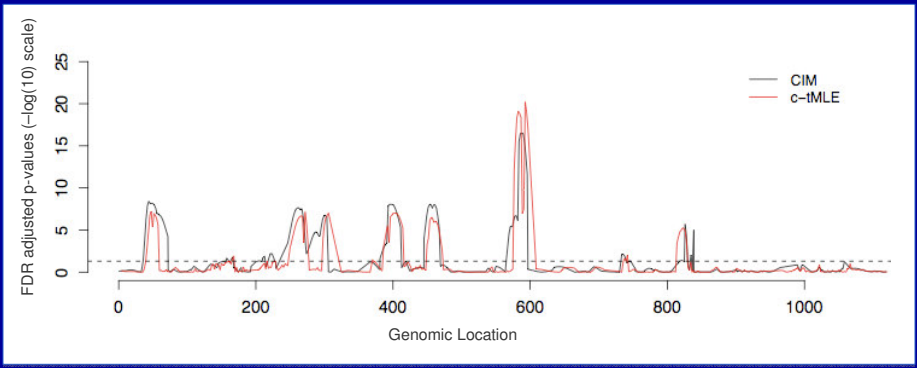

- **Variable importance coupled with LARS estimates true causal effect and outperforms both linear regression and randomForest**

# Application of Targeted MLE to Assess Effects of Single Nucleotide Polymorphisms

## Wound Healing Genes in Mice

- Goal: map genes controlling for wound healing trait in mice.
- 119 microsatellite markers on 19 chromosomes were genotyped for 633 F2 mice, covering about 1100 cM (70%) of the entire mouse genome with an average marker interval about 15cM.
- 400 positions are tested
  - Reference: MASINDE, G. L., X. LI, W. GU, H. DAVIDSON, S. MOHAN and D. J. BAYLINK, 2001. Identification of wound healing/regeneration quantitative trait loci (QTL) at multiple time points that explain seventy percent of variance in (MRL/MpJ and SJL/J) mice F2 population. *Genome Research* **11:** 2027-2033.

# Analyses of the Mouse Data




- Compare TMLE to Composite interval mapping (CIM)
  - A traditional MLE based approach, obtained from QTL mapping software QTL Cartographer.
- Dashed line is 0.05 significant line.

## Significant Genes at 0.05 in Mouse Data

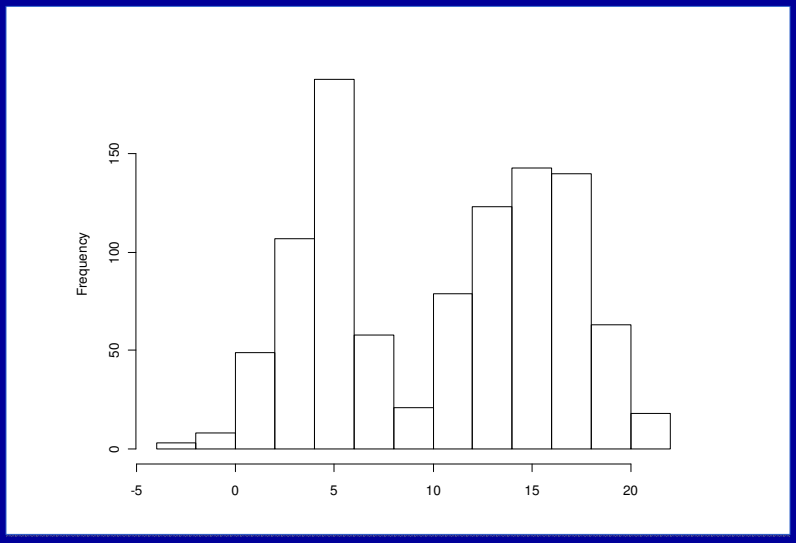
| | CIM | | | c-tMLE | | |
|---|---|---|---|---|---|---|
| QTL ID | Chr | cM | Prop(%) | Chr | cM | Prop(%) |
| 1 | 1 | 43.91 | 6.72 | 1 | 47.40 | 5.47 |
| 2 | 2 | 56.81 | 0.98 | 2 | 58.30 | 1.02 |
| 3 | 3 | 35.31 | 3.54 | - | - | - |
| 4 | 4 | 14.01 | 5.02 | 4 | 25.40 | 5.90 |
| 5 | 4 | 55.01 | 4.28 | 4 | 59.40 | 4.63 |
| 7 | 6 | 28.21 | 4.90 | 6 | 35.40 | 4.79 |
| 8 | 7 | 30.91 | 5.87 | 7 | 33.40 | 5.29 |
| 9 | 9 | 41.31 | 10.39 | 9 | 38.30 | 11.82 |
| 10 | - | - | - | 9 | 48.30 | 13.11 |
| 11 | 12 | 2.01 | 1.58 | 12 | 8.30 | 1.35 |
| 13 | 13 | 45.91 | 2.93 | 13 | 42.10 | 3.70 |
| 14 | 13 | 58.61 | 2.92 | - | - | - |
| Σ | | | 49.14 | | | 58.06 |

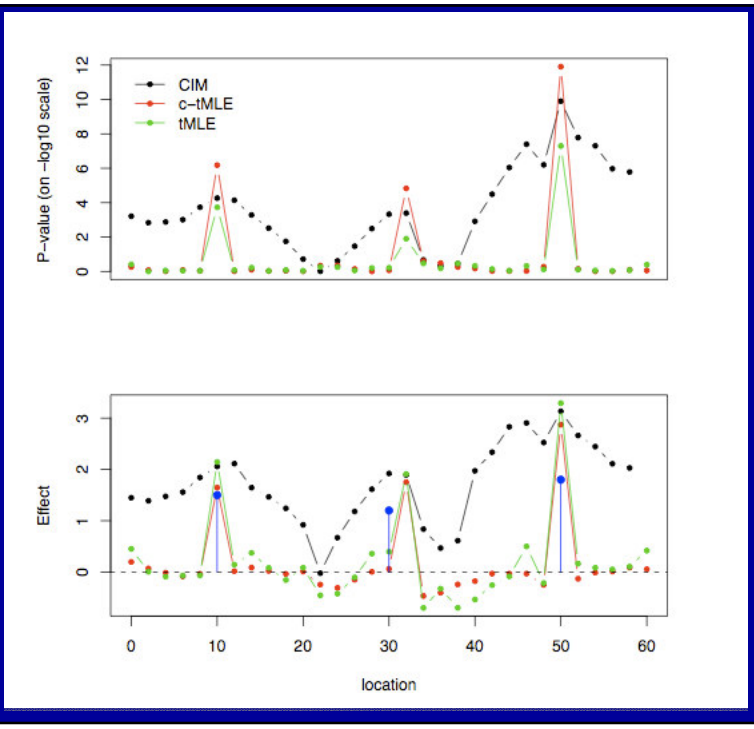
--- *Chr* means chromosome; *cM* means position of the gene on chromosome, in centi-morgans; *Prop* is the percentage of variance explained by the gene.
--- Genes in red are supported by various previous studies.

# Simulation

- Objective: Composite Interval Mapping Method (standard), and CTMLE
- Data:
  - 1000 Backcross mice
  - 30 markers on one chromosome, spaced at 2 centi-morgans. Due to this small space, nearby markers are highly correlated.
  - 3 main effects at marker position 10, 30 and 50 centi-morgans.
  - Outcome y is simulated with the linear model: $y = 5 + 1.5\ x_{10} + 1.2\ x_{30} + 1.8\ x_{50} + N(0, 8)$ and then truncated at 6.5, with y = 15 + rnorm(0, 8) This creates a bi-mode distribution of the outcome y.

# Distribution of Outcome Y






# Application of Repeated Measures TMLE to Yeast Regulatory Network Cell Cycle Data

# TVIM for Repeated Measures
## Generalization to Multivariate Outcome

Parameter of Interest: $\mu(a,t) = E[E(Y(t) \mid A = a, W) - E(Y(t) \mid A = 0, W)]$

define $m_t(a, W \mid \beta) = E(Y(t) \mid A = a, W) - E(Y(t) \mid A = 0, W)$

$m_t(A = 0, t, W \mid \beta) = 0 \quad \forall \beta, W$

- Update achieved using Gaussian generalized estimating equation with "clever covariate"

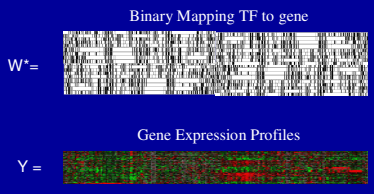
$$h^*(A, W) = \left( \frac{d}{d\beta} m(A, W \mid \beta) - E\left[\Sigma(A, W)^{-1} \mid W\right]^{-1} E\left[\Sigma(A, W)^{-1} \frac{d}{d\beta_0} m(A, W \mid \beta) \mid W\right] \right)$$

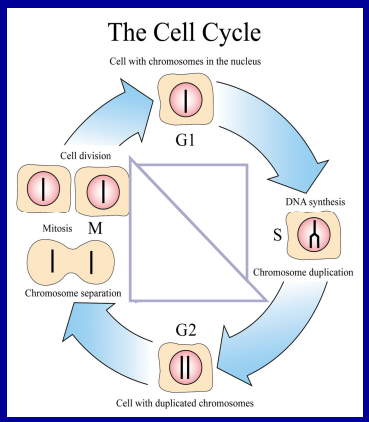
- Applied using geeglm() implementation of generalized estimating equations (R library geepack)



# Yeast Regulatory Network: Cell Cycle

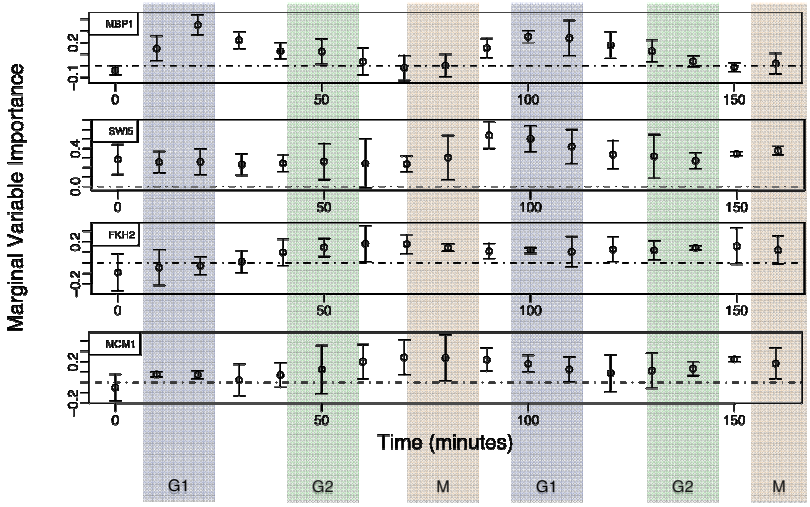
- Importance of transcription factor over cell cycle




http://nobelprize.org/nobel_prizes/medicine/laureates/2001/press.html

- Cho et al 1998 data
  - t=0,...,160 min (2 cycles)
  - $m(A,W|\beta)=At_0\beta_0 + At_{10}\beta_{10}+ \ldots + At_{160}\beta_{160}$
- Examples:
  - MBP1 (G1 phase)
  - SWI5 (M and M to G1)
  - FKH2, MCM1 (G2, G2 to M, M, M to G1 phases)




55

# Collaborative Targeted MLE in Observational Studies:

Causal Effect of Treatment

## Why Collaborative TMLE?

- TMLE requires estimation of a treatment mechanism (i.e. p(A|W), E[A|W])
  - Getting this right results in targeted bias reduction
  - Often many covariates, requires using non-targeted machine learning algorithm based on irrelevant likelihood of treatment mechanism
- Collaborative TMLE eliminates reliance on external estimate of treatment mechanism
- Results in a fully targeted estimating procedure

## Collaborative Robustness of T-MLE

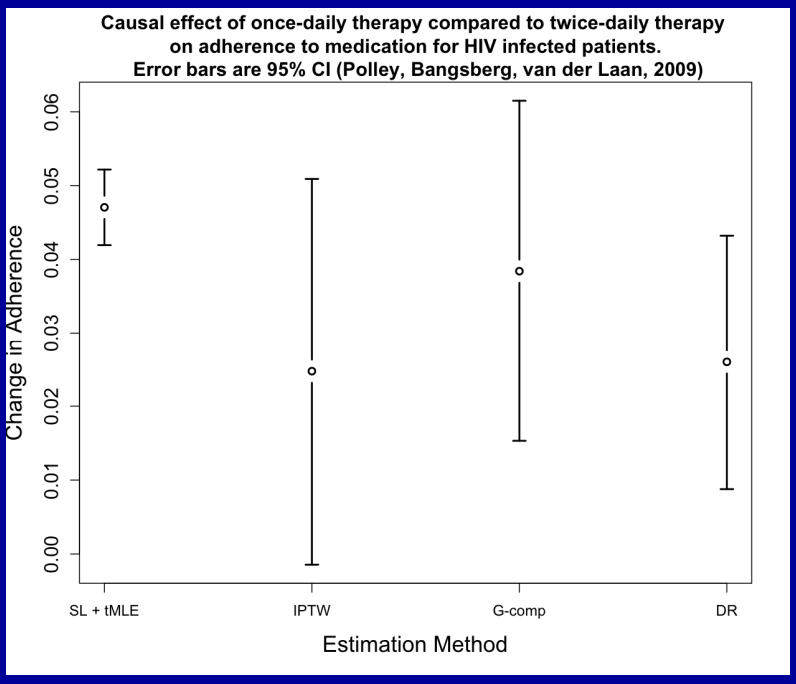
- Suppose the initial outcome regression of Y on A,W, minus true regression only depends on a subset S of all confounders W
- Suppose the treatment mechanism in clever covariate only adjusts correctly for a set of confounders that includes S
- Then, the Targeted MLE is consistent!
- Thus the treatment mechanism in clever covariate only needs to adjust for covariates whose effect has not been captured by the initial regression yet.

# Collaborative TMLE

## Building the Propensity Score Based on Outcome Data

- Initial regression based on super learning
- Construct rich set of one dimensional dimension reductions of W, that will be used as main terms below
- Select main terms in propensity score in clever covariate using forward selection based on fit (e.g, loglik) of

  TMLE using this clever covariate
- If no main term increases emp. fit of TMLE, then carry out TMLE update with the current clever covariate
- Proceed to generate a sequence of T-MLE's using increasingly nonparametric treatment mechanisms in the subsequent clever covariates
- Select the wished T-MLE with cross-validation



# Causal Effect on Survival

## Simulation Study Investigating C-TMLE

- Generated 500 data sets of 500 observations each per scenario.
- Varied How Well Hazard Was Specified (Well/Badly)
- Varied Level of ETA Violation (Low/Med/High)
- Observed performance of: IPCW, G-comp, Double Robust, TMLE, C-TMLE and TMLE excluding the variable that causes the eta violation in the treatment mechanism.

# MSE and Relative Efficiency of Estimating $P(T_{A=1}>t_0)$

| | Low ETA | | Medum ETA | | High ETA | |
|---|---|---|---|---|---|---|
| Method | Well Specified λ | Badly Specified λ | Well Specified λ | Badly Specified λ | Well Specified λ | Badly Specified λ |
| IPCW | 0.00086 (1.1) | 0.00086 (1.1) | 0.00117 (1.2) | 0.00117 (1.2) | 0.00698 (1.9) | 0.00698 (1.9) |
| G-comp | 0.00071 (0.9) | 0.00086 (1.1) | 0.00073 (0.7) | 0.00161 (1.6) | 0.00071 (0.2) | 0.00231 (0.6) |
| Double Robust | 0.00085 (1.1) | 0.00088 (1.1) | 0.00106 (1.1) | 0.00116 (1.2) | 0.00455 (1.3) | 0.00819 (2.3) |
| TMLE | 0.00085 (1.1) | 0.00088 (1.1) | 0.00105 (1.1) | 0.00111 (1.1) | 0.00347 (1.0) | 0.00375 (1.0) |
| TMLE w/o ETA | 0.00084 (1.0) | 0.00088 (1.1) | 0.00083 (0.8) | 0.00087 (1.2) | 0.00079 (0.2) | 0.00085 (0.2) |
| c-TMLE | 0.00085 (1.1) | 0.00104 (1.3) | 0.00085 (0.9) | 0.00122 (1.2) | 0.00080 (0.2) | 0.00101 (0.3) |

# Percent of Time Influence Curve Based 95% CI Includes Truth

| | Low ETA | | Medum ETA | | High ETA | |
|---|---|---|---|---|---|---|
| Method | Well Specified λ | Badly Specified λ | Well Specified λ | Badly Specified λ | Well Specified λ | Badly Specified λ |
| IPCW | 98.4% | 98.4% | 97.2% | 97.2% | 89.8% | 89.8% |
| Double Robust | 94.2% | 95.4% | 95.0% | 96.4% | 94.8% | 95.8% |
| TMLE | 94.0% | 95.2% | 95.4% | 95.6% | 84.8% | 87.0% |
| TMLE w/o ETA | 94.2% | 94.8% | 94.4% | 95.6% | 94.6% | 95.0% |
| c-TMLE | 94.6% | 93.6% | 96.0% | 95.0% | 94.8% | 93.2% |

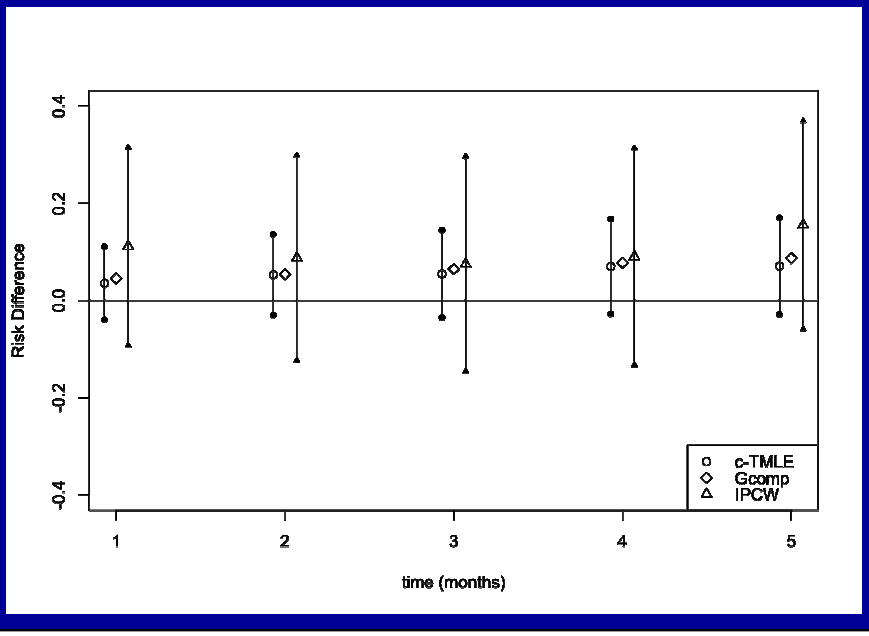
# Properties of c-TMLE Algorithm

| | Low ETA | | Medum ETA | | High ETA | |
|---|---|---|---|---|---|---|
| Method | Well Specified λ | Badly Specified λ | Well Specified λ | Badly Specified λ | Well Specified λ | Badly Specified λ |
| Mean Moves | 2.7 | 9.1 | 2.1 | 7.3 | 2.9 | 7.1 |
| % No Moves | 68.8% | 0.2% | 74.2% | 6.4% | 62.6% | 1.6% |
| % ETA Variable | 27.8% | 60.4% | 1.2% | 6.0% | 0.0% | 0.4% |

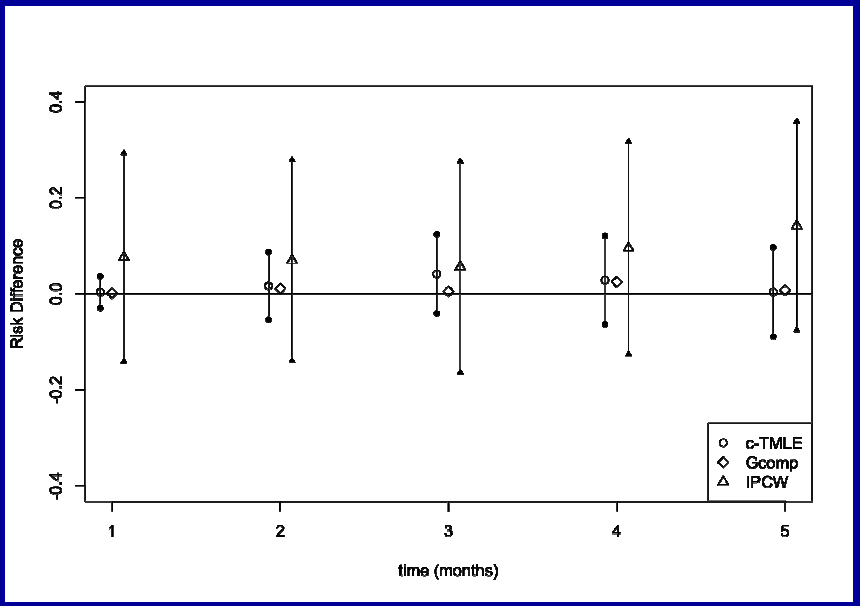
# Protease Inhibitors (PI) Effect on Viral Changes After Viral Failure

- 608 Subjects That Experience Viral Failure
- 478 Treated With Protease Inhibitor
- Baseline covariates adjusted for included: age, race, sexual orientation, IV drug use, baseline CD4, Viral Load, and prior therapeutic history.
- Two Outcomes of Interest Explored
  - T1 = Time until CD4 below 75 percent of baseline
  - T2 = Time until Viral Load below 500.

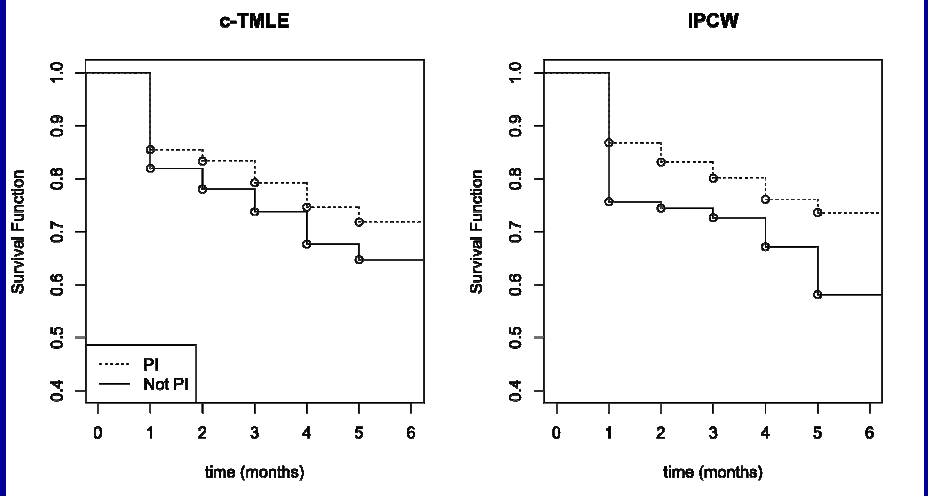
# Causal effect (RD) of PI on T1 (CD4)



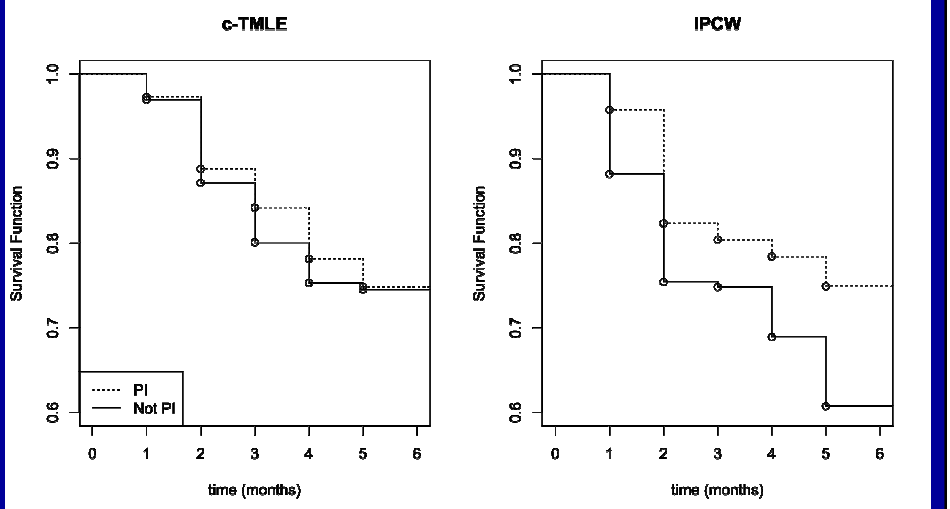
# Causal effect (RD) of PI on T2 (VL)




# Survival Curve, T1 (CD4)



## Survival Curve, T2 (VL)




# Targeted MLE of Effects in Case Control Studies

# Case-Control Weighted Targeted MLE

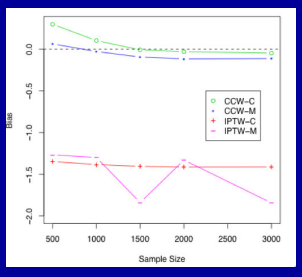
- Case-control weighting of the targeted MLE designed for prospective sampling results in wished targeted MLE for case-control sampling.
- This technique relies on knowledge of the true prevalence probability $P(Y=1)=q_0$ to eliminate the bias due to the case-control sampling design.
- The CC-W-TMLE is double robust and locally efficient. It produces efficient estimators when its prospective sample counterpart is efficient.

# Case-Control Weighted Targeted MLE

Table: **Odds Ratio** – MSEs and REs

| | | |
|---|---|---|
| IPTW MSE | M | 3.39 |
| IPTW RE | C | 1.69 |
| CCW T-MLE RE | C/C | 14.58 |
| | C/M | 14.57 |
| | M/C | 16.68 |

Table results for a sample of 500 cases and 1000 controls taken from a population of 120,000 where q0 = 0.035




## Simulation Results

- We showed striking improvements in efficiency and bias of case-control weighted method versus the IPTW estimator of causal odds ratio (Mansson 2007, Robins 1999), which does not utilize $q_0$.
- Our complete simulation results bolster our theoretical arguments that gains in efficiency and reductions in bias can be obtained by having known $q_0$ and using a targeted estimator.

# Closing Remarks

- True knowledge is embodied by semi or non-parametric models
- Semi-parametric models require fully automated state of the art machine learning (super learning)
- Targeted bias removal is essential and is achieved by collaborative targeted MLE
- Statistical inference is now sensible
- The machine learning algorithms are (super) efficient for the target parameters.


# Acknowledgements:

## Special Thanks to Cathy Tuglus

- UC Berkeley
  - Susan Gruber
  - Ori Stittelman
  - Hui Wang
  - Kelly Moore
  - Maya Petersen
  - Dan Rubin
  - Cathy Tuglus
  - Sherri Rose
  - Michael Rosenblum
  - Eric Polley
- Stanford: Dr. Robert Shafer
- Kaiser: Dr. Jeffrey Fessels
- Harvard: Dr. David Bangsberg
- UCSF: Dr. Steve Deeks
- FDA: Thamban Vallappil, Greg Soon

-P.I. Ira Tager (Epi).

# References


- Oliver Bembom, Maya L. Petersen , Soo-Yon Rhee , W. Jeffrey Fessel , Sandra E. Sinisi, Robert W. Shafer, and Mark J. van der Laan, "Biomarker Discovery Using Targeted Maximum Likelihood Estimation: Application to the Treatment of Antiretroviral Resistant HIV Infection" (August 2007). *U.C. Berkeley Division of Biostatistics Working Paper Series*. Working Paper 221. http://www.bepress.com/ucbbiostat/paper221
- Mark J. van der Laan and Susan Gruber, "Collaborative Double Robust Targeted Penalized Maximum Likelihood Estimation" (April 2009). U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 246, http://www.bepress.com/ucbbiostat/paper246
- Mark J. van der Laan, Eric C. Polley, and Alan E. Hubbard, "Super Learner" (July 2007). *U.C. Berkeley Division of Biostatistics Working Paper Series*. Working Paper 222. http://www.bepress.com/ucbbiostat/paper222
- Mark J. van der Laan and Daniel Rubin, "Targeted Maximum Likelihood Learning" (October 2006). *U.C. Berkeley Division of Biostatistics Working Paper Series.* Working Paper 213. http://www.bepress.com/ucbbiostat/paper213
- Oliver Bembom, Mark van der Laan (2008), A practical illustration of the importance of realistic individualized treatment rules in causal inference, Electronic Journal of Statistics.
- Mark J. van der Laan, "Statistical Inference for Variable Importance" (August 2005). *U.C. Berkeley Division of Biostatistics Working Paper Series*. Working Paper 188. http://www.bepress.com/ucbbiostat/paper188