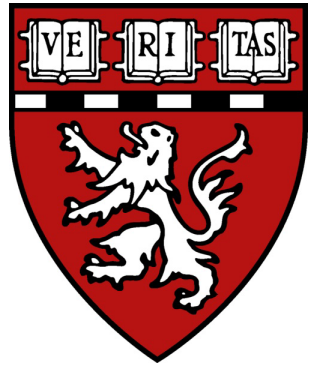
Ensembles in Public Health and Health Policy

Sherri Rose

Associate Professor
Department of Health Care Policy
Harvard Medical School

@sherrirose
drsherrirose.com

May 4, 2016



statistician

tell me something
interesting (fast)

insert data

big data system

Idreos/Rose

High Dimensional 'Big Data' Parametric Regression

- Often dozens, hundreds, or even thousands of potential variables

1513	4.163950	3.854244	3.827496
1555	4.277033	3.373982	489.825226
1597	4.390150	3.795142	221.608444
1639	4.503117	3.640379	26.986557
1681	4.616217	3.336954	104.501778
1723	4.729317	3.561723	8.354190
1765	4.842267	3.576960	146.476227
1807	4.955350	3.858309	58.118893
1849	5.068450	3.514176	3.682388
1891	5.181567	3.794615	32.864357
1933	5.294517	3.311670	1.653655
1975	5.407600	3.931615	72.284065
2017	5.520700	4.319901	15.170299
2059	5.633650	3.938955	2.626603
2101	5.746750	3.924497	16.581503
2143	5.859883	3.771340	33.761124
2185	5.972850	3.797512	9.262811
2227	6.085967	3.795501	126.762199
2269	6.199067	3.759673	108.416565
2311	6.312167	3.373145	10.712665
2353	6.425117	3.464702	56.385990
2395	6.538183	3.640879	30.747551
2437	6.651333	3.702649	5.748046
2479	6.764283	3.941036	58.997993
2521	6.877350	3.393778	24.935211
2563	6.990450	3.213435	6.881421
2605	7.103400	3.635089	12.697396
2647	7.216517	3.749416	4.405899
2689	7.329650	3.450428	6.340690
2731	7.442750	3.287580	231.588028

High Dimensional 'Big Data' Parametric Regression

- Often dozens, hundreds, or even thousands of potential variables
- Impossible challenge to correctly specify the parametric regression

1513	4.163950	3.854244	3.827496
1555	4.277033	3.373982	489.825226
1597	4.390150	3.795142	221.608444
1639	4.503117	3.640379	26.986557
1681	4.616217	3.336954	104.501778
1723	4.729317	3.561723	8.354190
1765	4.842267	3.576960	146.476227
1807	4.955350	3.858309	58.118893
1849	5.068450	3.514176	3.682388
1891	5.181567	3.794615	32.864357
1933	5.294517	3.311670	1.653655
1975	5.407600	3.931615	72.284065
2017	5.520700	4.319901	15.170299
2059	5.633650	3.938955	2.626603
2101	5.746750	3.924497	16.581503
2143	5.859883	3.771340	33.761124
2185	5.972850	3.797512	9.262811
2227	6.085967	3.795501	126.762199
2269	6.199067	3.759673	108.416565
2311	6.312167	3.373145	10.712665
2353	6.425117	3.464702	56.385990
2395	6.538183	3.640879	30.747551
2437	6.651333	3.702649	5.748046
2479	6.764283	3.941036	58.997993
2521	6.877350	3.393778	24.935211
2563	6.990450	3.213435	6.881421
2605	7.103400	3.635089	12.697396
2647	7.216517	3.749416	4.405899
2689	7.329650	3.450428	6.340690
2731	7.442750	3.287580	231.588028

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

1513	4.163950	3.854244	3.827496
1555	4.277033	3.373982	489.825226
1597	4.390150	3.795142	221.608444
1639	4.503117	3.640379	26.986557
1681	4.616217	3.336954	104.501778
1723	4.729317	3.561723	8.354190
1765	4.842267	3.576960	146.476227
1807	4.955350	3.858309	58.118893
1849	5.068450	3.514176	3.682388
1891	5.181567	3.794615	32.864357
1933	5.294517	3.311670	1.653655
1975	5.407600	3.931615	72.284065
2017	5.520700	4.319901	15.170299
2059	5.633650	3.938955	2.626603
2101	5.746750	3.924497	16.581503
2143	5.859883	3.771340	33.761124
2185	5.972850	3.797512	9.262811
2227	6.085967	3.795501	126.762199
2269	6.199067	3.759673	108.416565
2311	6.312167	3.373145	10.712665
2353	6.425117	3.464702	56.385990
2395	6.538183	3.640879	30.747551
2437	6.651333	3.702649	5.748046
2479	6.764283	3.941036	58.997993
2521	6.877350	3.393778	24.935211
2563	6.990450	3.213435	6.881421
2605	7.103400	3.635089	12.697396
2647	7.216517	3.749416	4.405899
2689	7.329650	3.450428	6.340690
2731	7.442750	3.287580	231.588028

High Dimensional 'Big Data' Parametric Regression

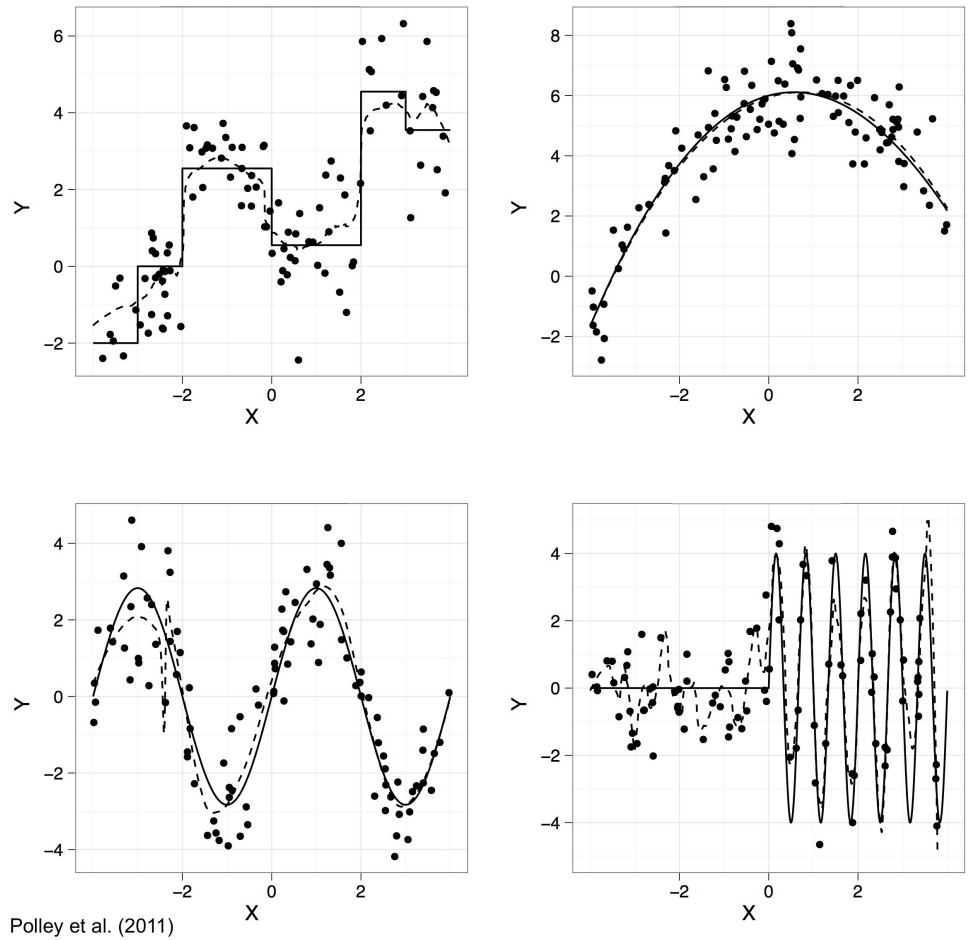
- Often dozens, hundreds, or even thousands of potential variables
- Impossible challenge to correctly specify the parametric regression
- May have more unknown parameters than observations
- True functional might be described by a complex function not easily approximated by main terms or interaction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

1513	4.163950	3.854244	3.827496
1555	4.277033	3.373982	489.825226
1597	4.390150	3.795142	221.608444
1639	4.503117	3.640379	26.986557
1681	4.616217	3.336954	104.501778
1723	4.729317	3.561723	8.354190
1765	4.842267	3.576960	146.476227
1807	4.955350	3.858309	58.118893
1849	5.068450	3.514176	3.682388
1891	5.181567	3.794615	32.864357
1933	5.294517	3.311670	1.653655
1975	5.407600	3.931615	72.284065
2017	5.520700	4.319901	15.170299
2059	5.633650	3.938955	2.626603
2101	5.746750	3.924497	16.581503
2143	5.859883	3.771340	33.761124
2185	5.972850	3.797512	9.262811
2227	6.085967	3.795501	126.762199
2269	6.199067	3.759673	108.416565
2311	6.312167	3.373145	10.712665
2353	6.425117	3.464702	56.385990
2395	6.538183	3.640879	30.747551
2437	6.651333	3.702649	5.748046
2479	6.764283	3.941036	58.997993
2521	6.877350	3.393778	24.935211
2563	6.990450	3.213435	6.881421
2605	7.103400	3.635089	12.697396
2647	7.216517	3.749416	4.405899
2689	7.329650	3.450428	6.340690
2731	7.442750	3.287580	231.588028

Big Picture

Machine learning aims to

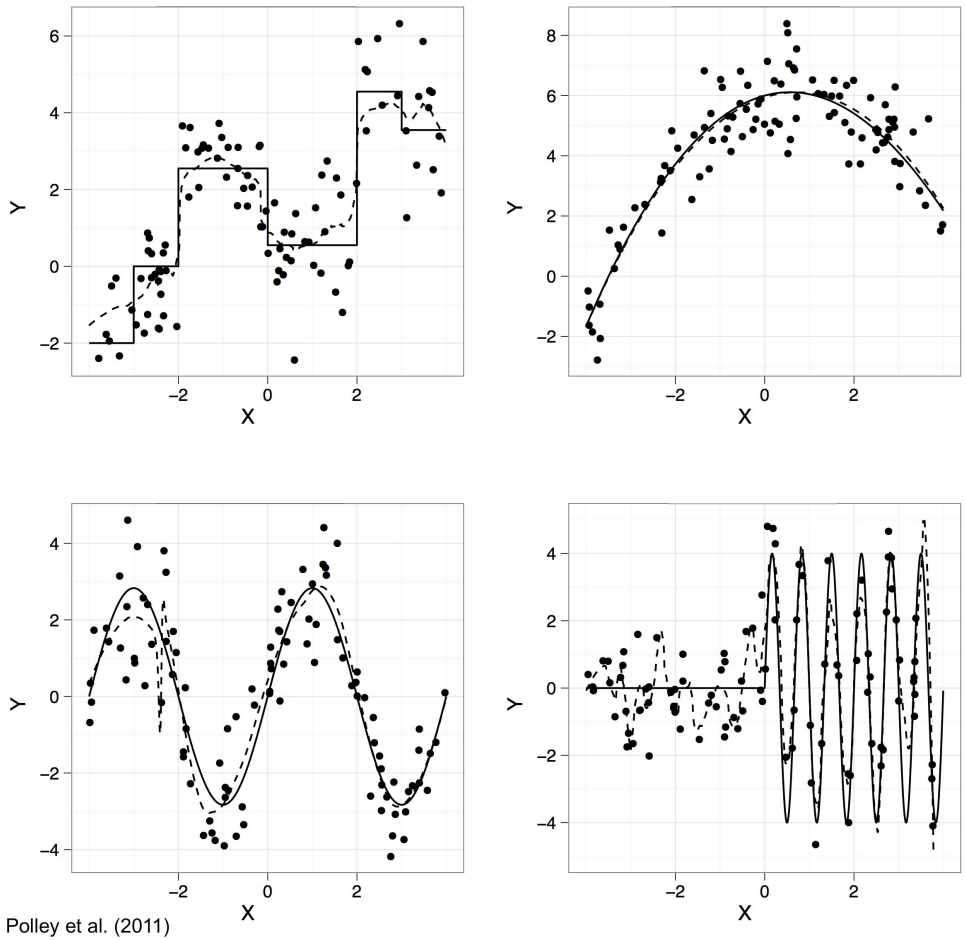
- "smooth" over the data
- make fewer assumptions



Big Picture

Purely nonparametric model with high dimensional data?

- $p > n$!
- data sparsity



Options?

- Recent studies for prediction have employed newer **algorithms.**
  (any mapping from data to a predictor)

Options?

- Recent studies for prediction have employed newer **algorithms.**
- Researchers are then left with questions, e.g.,
  - *"When should I use random forest instead of standard regression techniques?"*

Options?

- Recent studies for prediction have employed newer **algorithms.**
- Researchers are then left with questions, e.g.,
  - *"When should I use random forest instead of standard regression techniques?"*

ELSEVIER

Journal of Clinical Epidemiology 63 (2010) 1145–1155

Journal of Clinical Epidemiology

Logistic regression had superior performance compared with regression trees for predicting in-hospital mortality in patients hospitalized with heart failure

Peter C. Austin[a,c,*], Jack V. Tu[a,b,c,d,e], Douglas S. Lee[a,e,f]

Options?

- Recent studies for prediction have employed newer **algorithms.**
- Researchers are then left with questions, e.g.,
  - *"When should I use random forest instead of standard regression techniques?"*

ELSEVIER

Journal of Clinical Epidemiology 63 (2010) 1145–1155

Journal of Clinical Epidemiology

Logistic regression had superior p
trees for predicting in-hospital n
hea

Peter C. Austin[a,c,*], Jack

European Journal of Neurology 2010, **17:** 945–950 doi:10.1111/j.1468-1331.2010.02955.x

Random forest can predict 30-day mortality of spontaneous intracerebral hemorrhage with remarkable discrimination

S. -Y. Peng[a,b,c], Y. -C. Chuang[b], T. -W. Kang[b] and K. -H. Tseng[d]

[a]*Institute of Biomedical Informatics, National Yang-Ming University, Taipei;* [b]*Department of Anesthesiology, Taichung Veterans General Hospital, Taichung;* [c]*School of Medicine, Chung Shan Medical University, Taichung; and* [d]*Department of Nephrology, Taoyuan Veterans Hospital, Taoyuan, Taiwan*

Ensembling: Cross-Validation

- 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.

Image credit: Rose (2016)

Ensembling: Cross-Validation

- 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**.

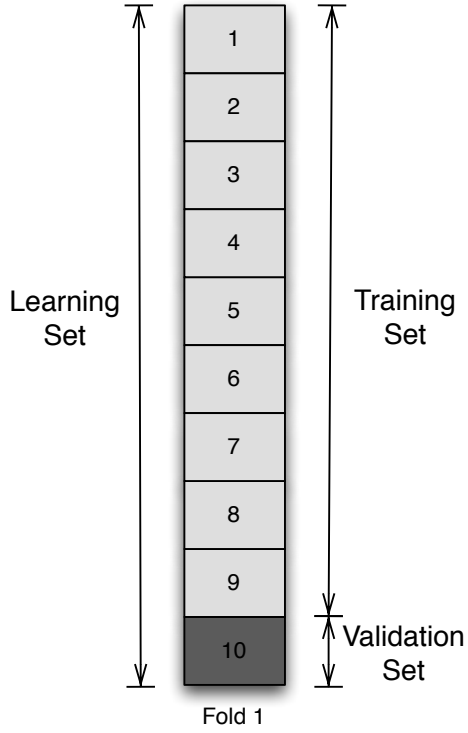


Image credit: Rose (2016)

Ensembling: Cross-Validation

- In V-fold cross-validation, our observed data $O_1, \ldots, O_n$ is referred to as the learning set and partition into V sets of size $\approx \frac{n}{V}$
- For any given fold, $V - 1$ sets comprise training set and remaining 1 set is validation set.

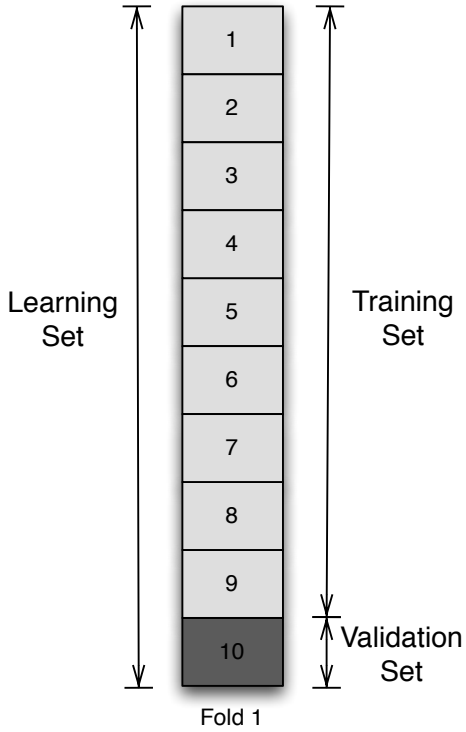


Image credit: Rose (2016)

Ensembling: Cross-Validation

- In V-fold cross-validation, our observed data $O_1, \ldots, O_n$ is referred to as the learning set and partition into V sets of size $\approx \frac{n}{V}$
- For any given fold, $V - 1$ sets comprise training set and remaining 1 set is validation set.

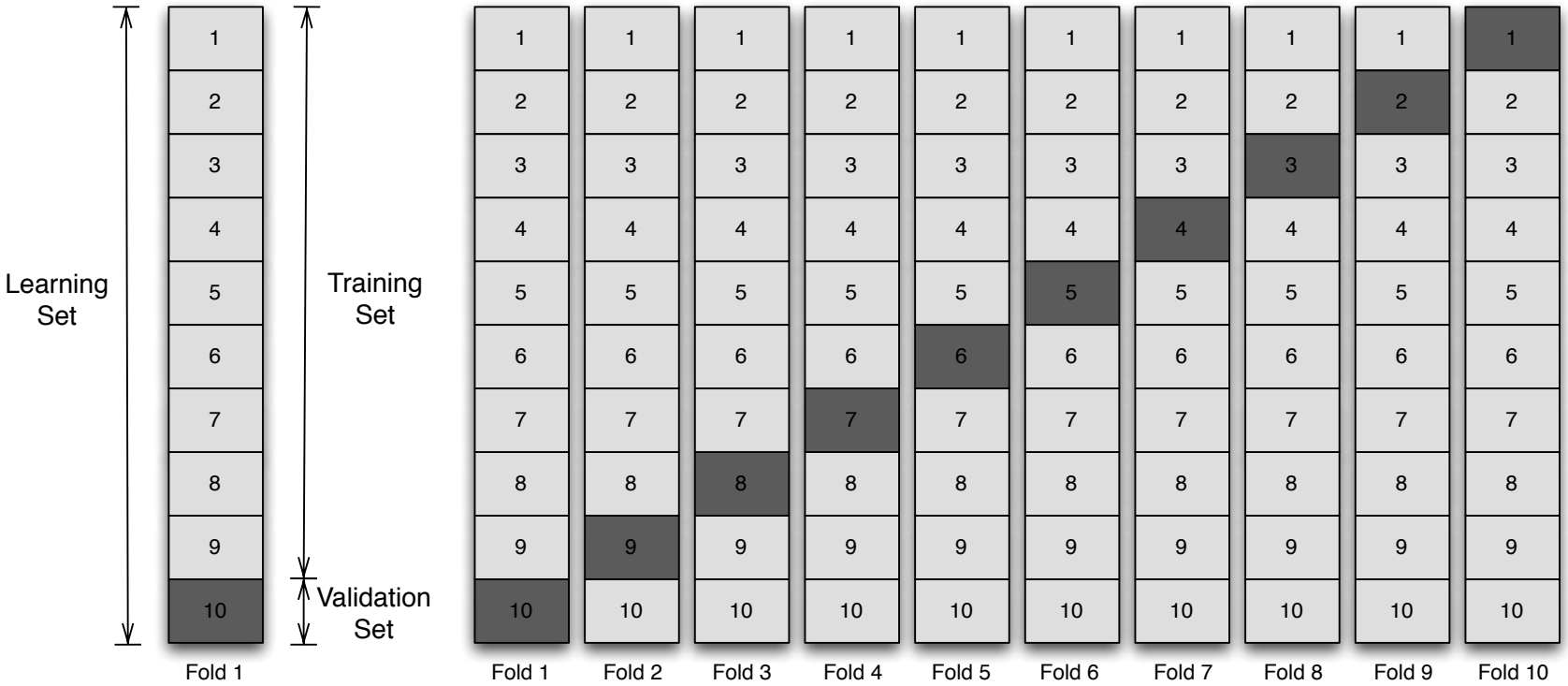


Image credit: Rose (2016)

Ensembling: Super Learner

Build collection of algorithms of all weighted averages of the algorithms.

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

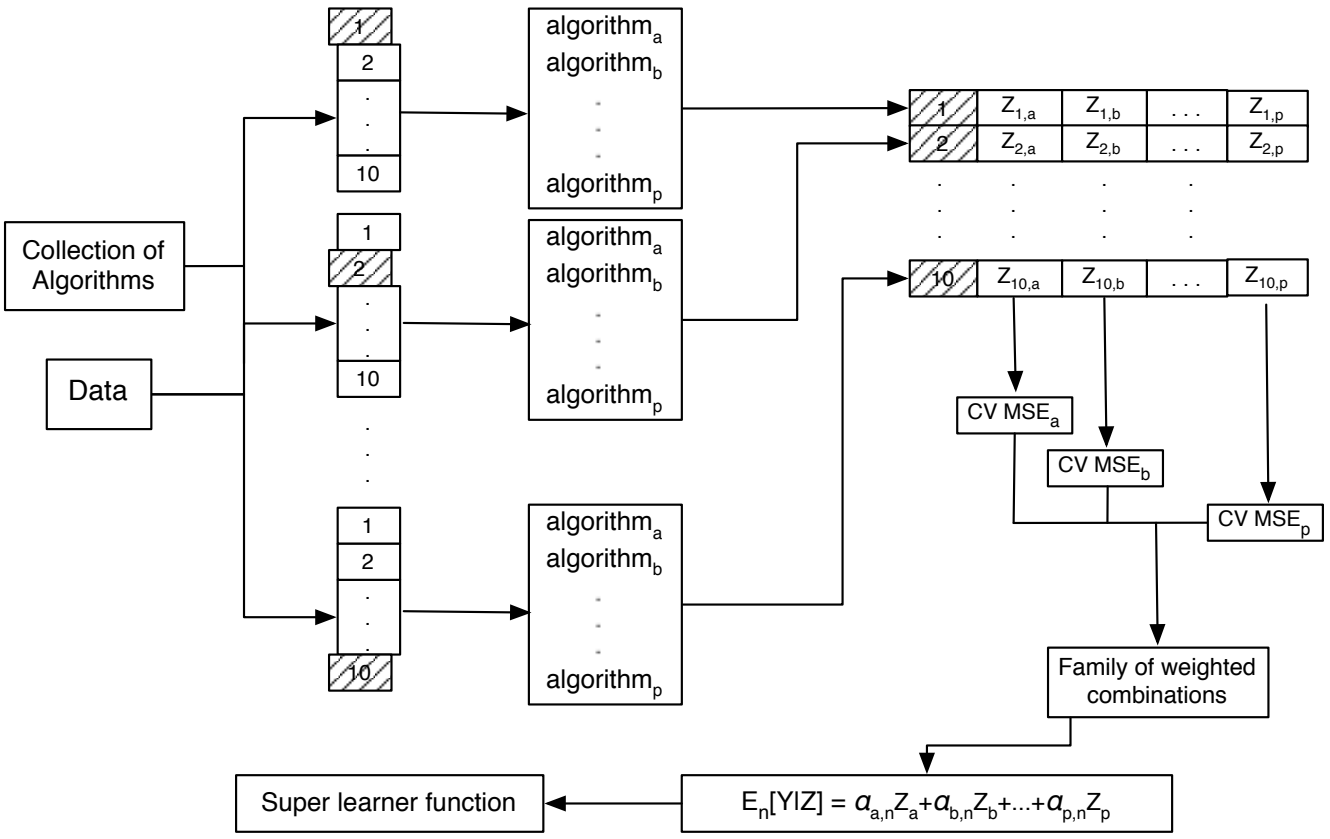


Image credit: Polley et al. (2011)

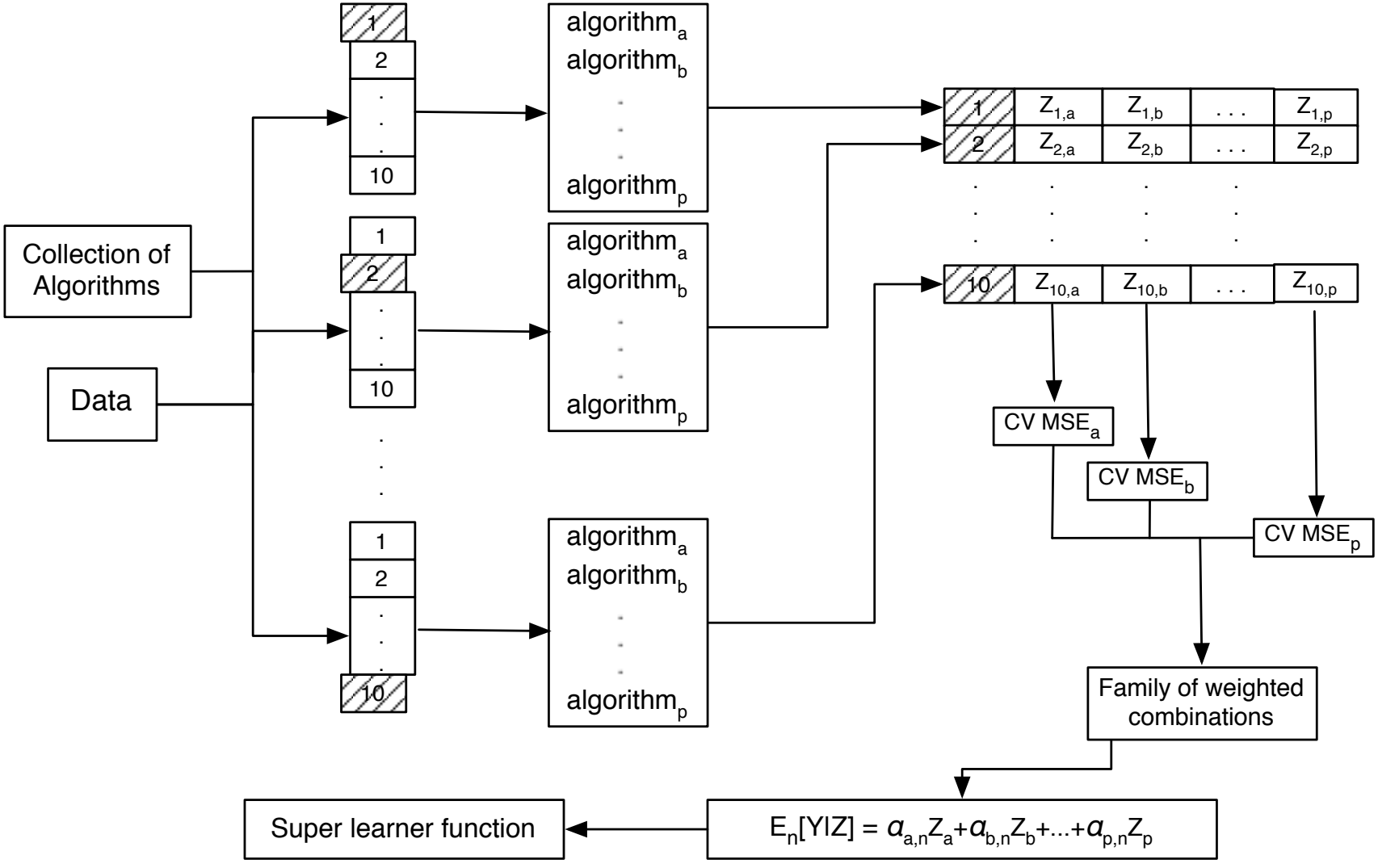


Image credit: Polley et al. (2011)

R Packages

- `SuperLearner` (Polley): Main super learner package
- `h2oEnsemble` (LeDell): Java-based, designed for big data, uses H2O `R` interface to run super learning

More: `targetedlearningbook.com/software`

Super Learner Sample Code

Code online: drsherrirose.com/slexample

```
library(SuperLearner)
set.seed(27)
n<-500

data <- data.frame(W1=runif(n, min = .5, max = 1),
W2=runif(n, min = 0, max = 1),
W3=runif(n, min = .25, max = .75),
W4=runif(n, min = 0, max = 1))
data <- transform(data,
W5=rbinom(n, 1, 1/(1+exp(1.5*W2-W3))))
data <- transform(data,
Y=rbinom(n, 1,1/(1+exp(-(-.2*W5-2*W1+4*W5*W1-1.5*W2+sin(W4))))))
```

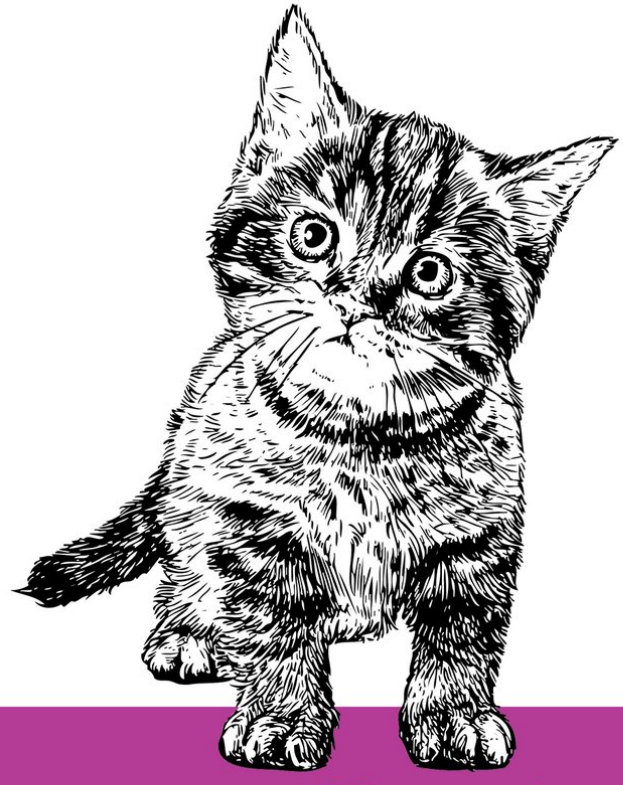
Super Learner Sample Code

```
SL.library <- c("SL.nnet", "SL.glm", "SL.mean", "SL.randomForest")

fit.data.SL<-SuperLearner(Y=data[,6],X=data[,1:5],
SL.library=SL.library, family=binomial(),method="method.NNLS",
verbose=TRUE)

fit.data.SL     #CV risks for algorithms in the library
```

Super Learner Sample Code

```
fitSL.data.CV <- CV.SuperLearner(Y=data[,6],X=data[,1:5],
V=10, SL.library=SL.library,verbose = TRUE,
method = "method.NNLS", family = binomial())

mean((data[,6]-fitSL.data.CV$SL.predict)^2)
```

How to actually learn any new programming concept



Essential

Changing Stuff and Seeing What Happens

O RLY?

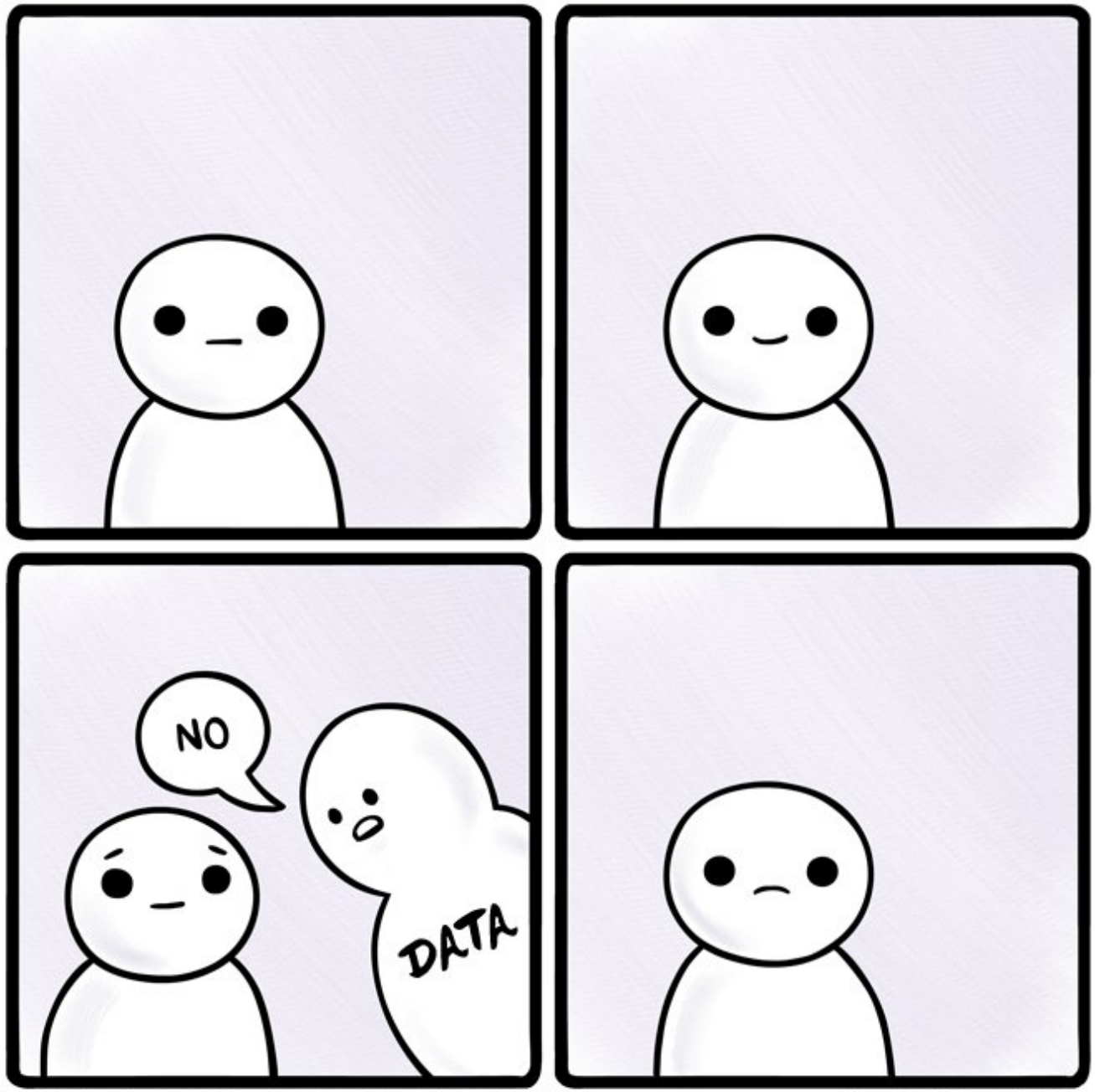
@ThePracticalDev

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.



THIS COMIC MADE POSSIBLE THANKS TO ADAM LINGELBACH

MRLOVENSTEIN.COM

Application: Kaiser Permanente Database

Nested case-control sample (n=27,012) from database of 345,191 persons over the age of 65 in 2003.

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

Ensembling method outperformed all other algorithms.

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

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

van der Laan & Rose (2011)

Application: Sonoma Cohort Study

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

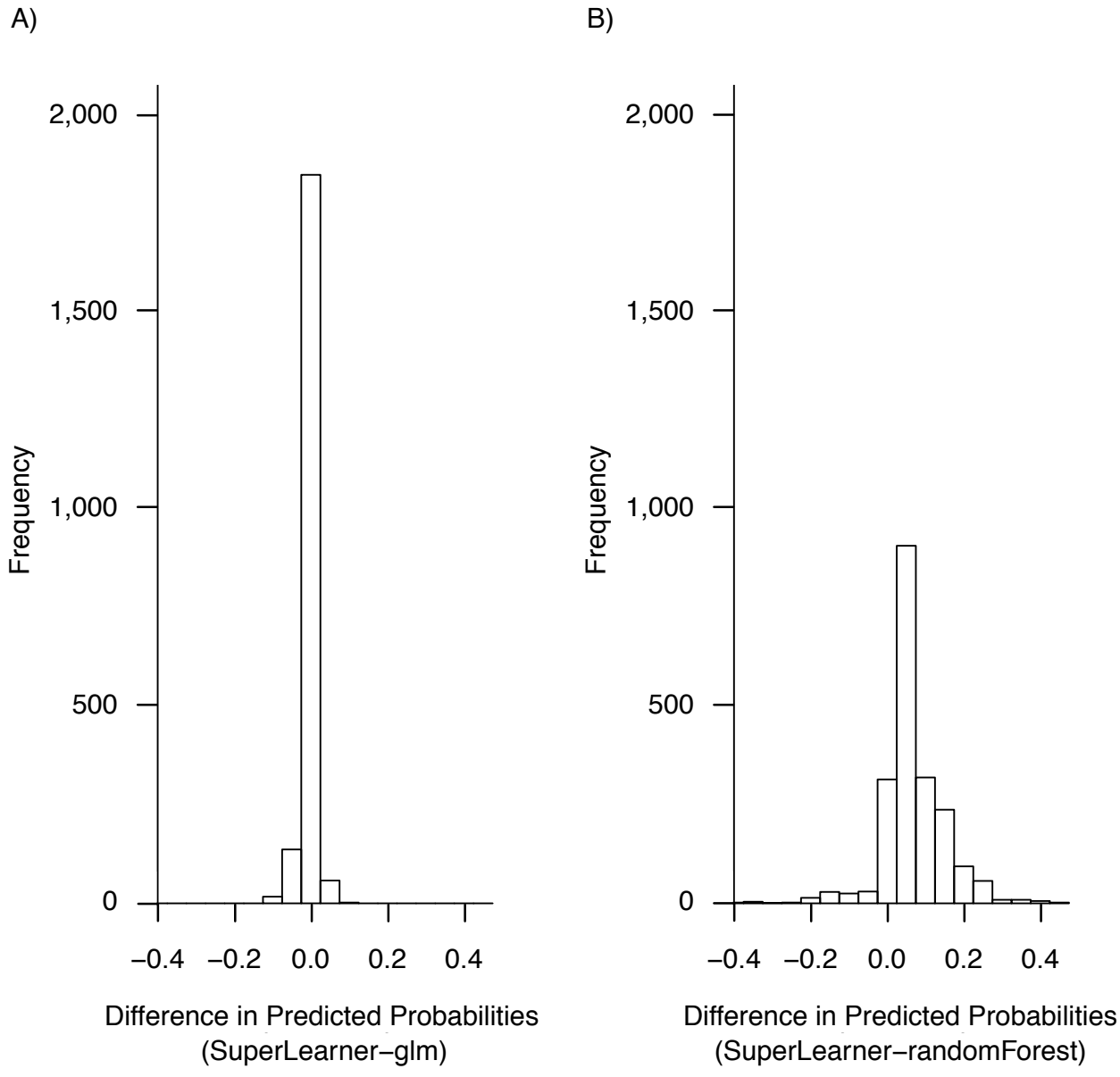
- Outcome: death.
- Covariates: gender, age, **self-rated health**, **leisure-time physical activity**, smoking status, cardiac event history, and chronic health condition status.
- $R^2 = 0.201$

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)

Application: Sonoma Cohort Study

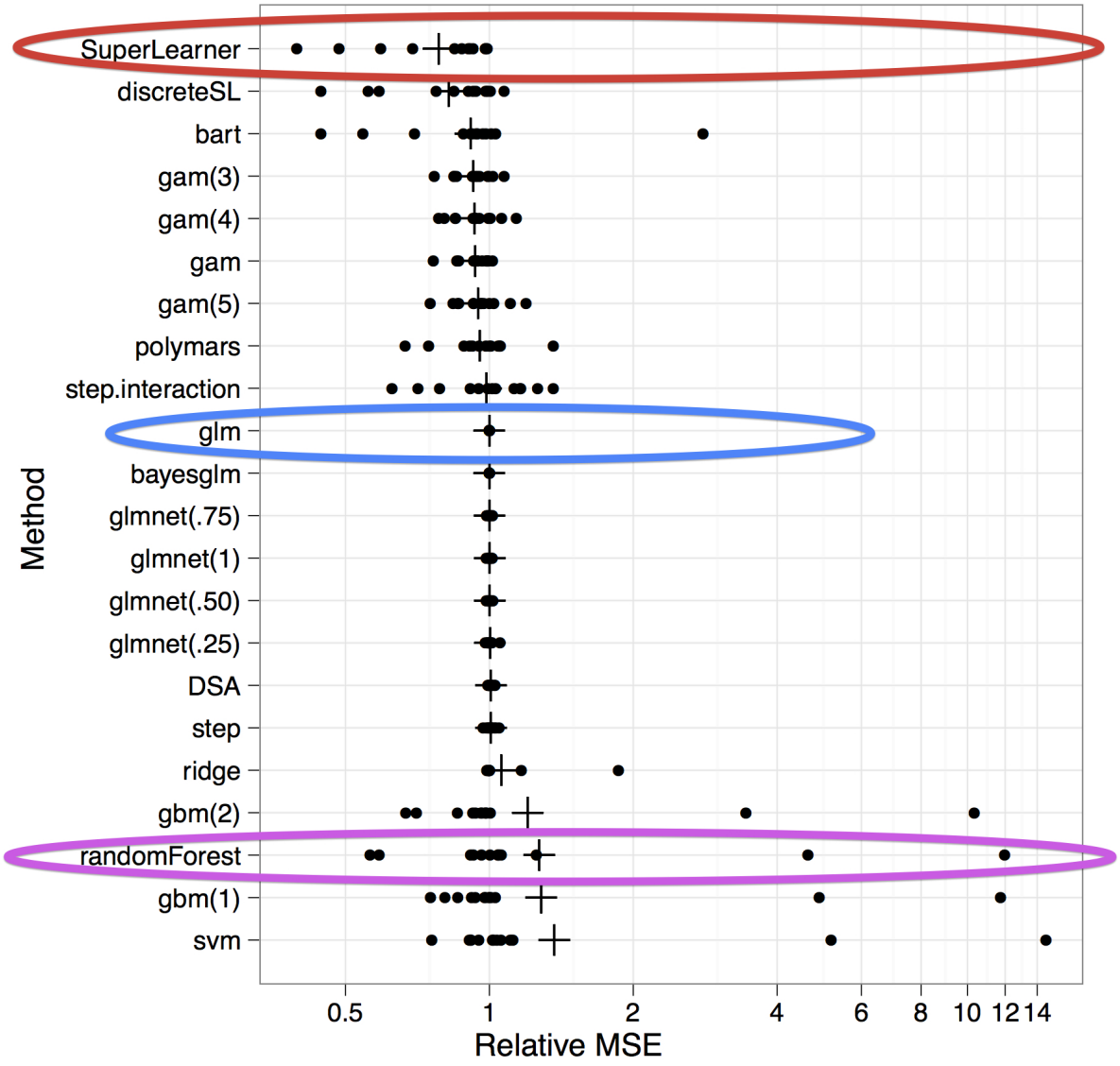


Application: Public Datasets

Name	n	p	Source
ais	202	10	Cook and Weisberg (1994)
diamond	308	17	Chu (2001)
cps78	550	18	Berndt (1991)
cps85	534	17	Berndt (1991)
cpu	209	6	Kibler et al. (1989)
FEV	654	4	Rosner (1999)
Pima	392	7	Newman et al. (1998)
laheart	200	10	Afifi and Azen (1979)
mussels	201	3	Cook (1998)
enroll	258	6	Liu and Stengos (1999)
fat	252	14	Penrose et al. (1985)
diabetes	366	15	Harrell (2001)
house	506	13	Newman et al. (1998)

Polley et al. (2011)

Application: Public Datasets



Polley et al. (2011)

Screening: Will Be Useful for Parsimony

- Often beneficial to screen variables before running algorithms.
- Can be coupled with prediction algorithms to create new algorithms in the library.

Screening: Will Be Useful for Parsimony

- Often beneficial to screen variables before running algorithms.
- Can be coupled with prediction algorithms to create new algorithms in the library.
- Clinical subsets

Screening: Will Be Useful for Parsimony

- Often beneficial to screen variables before running algorithms.
- Can be coupled with prediction algorithms to create new algorithms in the library.
- Clinical subsets
- Test each variable with the outcome, rank by p-value

Screening: Will Be Useful for Parsimony

- Often beneficial to screen variables before running algorithms.
- Can be coupled with prediction algorithms to create new algorithms in the library.

- Clinical subsets
- Test each variable with the outcome, rank by p-value
- Lasso

Risk Adjustment in Plan Payment

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

- Redistributes funds based on enrollee health
- Encourages competition based on efficiency & quality
- Huge financial implications in health care

Risk Adjustment in Plan Payment

HSR Health Services Research

DOI: 10.1111/1475-6773.12464
METHODS ARTICLE

A Machine Learning Framework for Plan Payment Risk Adjustment

Sherri Rose

DOI: 10.1377/hlthaff.2015.1026
HEALTH AFFAIRS 35,
NO. 3 (2016): 440–448

By Sherri Rose, Alan M. Zaslavsky, and J. Michael McWilliams

Variation In Accountable Care Organization Spending And Sensitivity To Risk Adjustment: Implications For Benchmarking

Sherri Rose is an assistant professor of health care policy (biostatistics) in the Department of Health Care Policy at Harvard Medical School, in Boston, Massachusetts.

Alan M. Zaslavsky is a professor of health care policy (statistics) in the Department of Health Care Policy at Harvard Medical School.

J. Michael McWilliams (mcwilliams@hcp.med.harvard.edu) is an associate professor of health care policy and medicine in the Department of Health Care Policy at Harvard Medical School.

ABSTRACT **Spending targets (or benchmarks) for accountable care organizations (ACOs) participating in the Medicare Shared Savings Program must be set carefully to encourage program participation while achieving fiscal goals and minimizing unintended consequences, such as penalizing ACOs for serving sicker patients. Recently proposed regulatory changes include measures to make benchmarks more similar for ACOs in the same area with different historical spending levels. We found that ACOs vary widely in how their spending levels compare with those of other local providers after standard case-mix adjustments. Additionally adjusting for survey measures of patient health meaningfully reduced the variation in differences between ACO spending and local average fee-for-service spending, but substantial variation remained, which suggests that differences in care efficiency between ACOs and local non-ACO providers vary widely. Accordingly, measures to equilibrate benchmarks between high- and low-spending ACOs—such as setting benchmarks to risk-adjusted average fee-for-service spending in an area—should be implemented gradually to maintain participation by ACOs with high spending. Use of survey information also could help mitigate perverse incentives for risk selection and upcoding and limit unintended consequences of new benchmarking methodologies for ACOs serving sicker patients.**

Risk Adjustment in Plan Payment = Prediction

Prediction: Generate function to input covariates and predict outcome value.

Adjust for patient characteristics to predict cost.

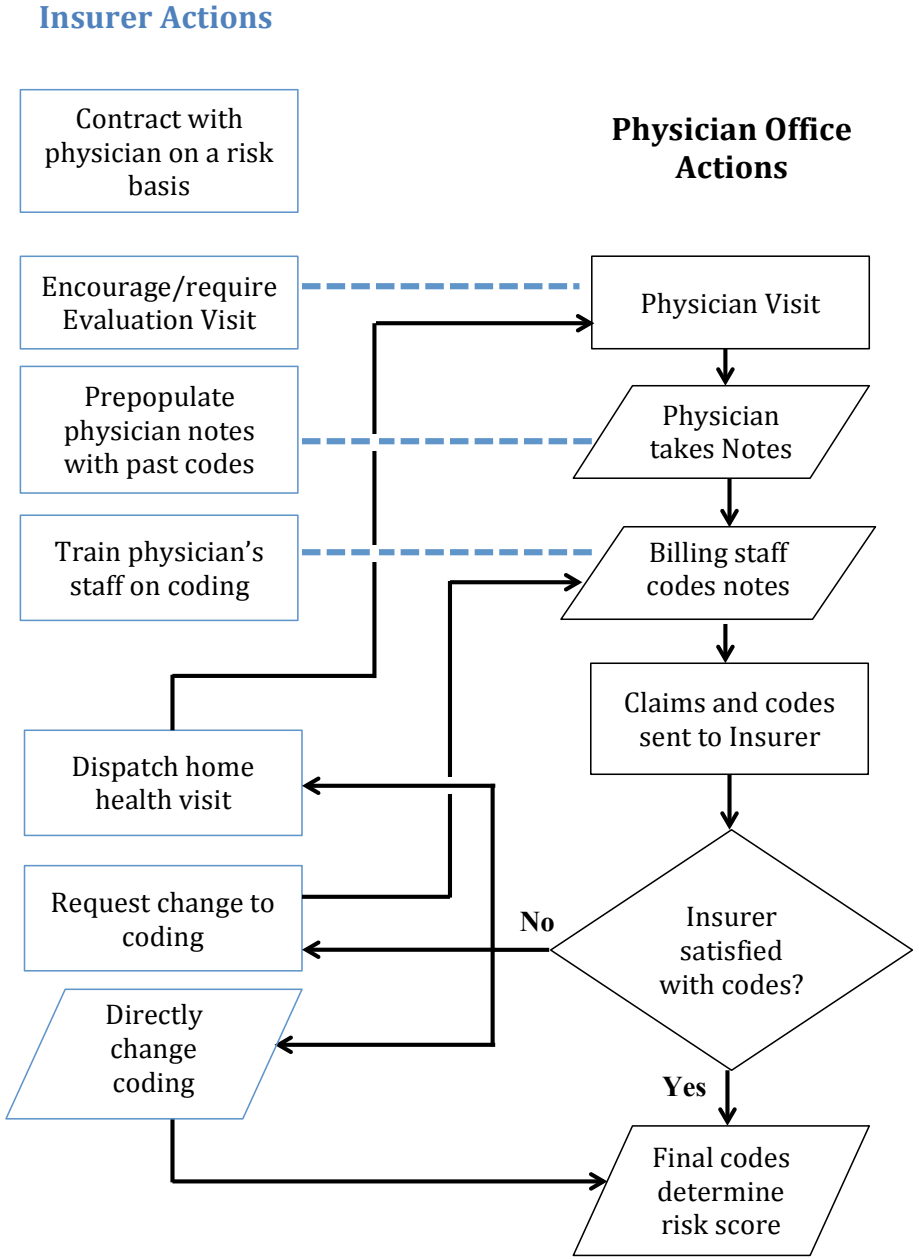
1513	4.163950	3.854244	5.827496
1555	4.277033	3.373982	489.825226
1597	4.390150	3.795142	221.608444
1639	4.503117	3.640379	26.986557
1681	4.616217	3.336954	104.501778
1723	4.729317	3.561723	8.354190
1765	4.842267	3.576960	146.476227
1807	4.955350	3.858309	58.118893
1849	5.068450	3.514176	3.682388
1891	5.181567	3.794615	32.864357
1933	5.294517	3.311670	1.653655
1975	5.407600	3.931615	72.284065
2017	5.520700	4.319901	15.170299
2059	5.633650	3.938955	2.626603
2101	5.746750	3.924497	16.581503
2143	5.859883	3.771340	33.761124
2185	5.972850	3.797512	9.262811
2227	6.085967	3.795501	126.762199
2269	6.199067	3.759673	108.416565
2311	6.312167	3.373145	10.712665
2353	6.425117	3.464702	56.385990
2395	6.538183	3.640879	30.747551
2437	6.651333	3.702649	5.748046
2479	6.764283	3.941036	58.997993
2521	6.877350	3.393778	24.935211
2563	6.990450	3.213435	6.881421
2605	7.103400	3.635089	12.697396
2647	7.216517	3.749416	4.405899
2689	7.329650	3.450428	6.340690
2731	7.442750	3.287580	231.588028

Risk Adjustment in Plan Payment = Frozen

$$E[Y \mid W] = \alpha_0 + \alpha_1 W$$

Potentially *$$$* oversight, where it attempts to control for the impact of consumers choosing health plans.

wpb.org



Geruso & Layton (2014)

Data

- **Truven MarketScan** database, those with continuous coverage in 2011-2012; about 11 million people. Variables: age, sex, region, procedures, expenditures, etc. For this paper, extract a random sample of 250,000 people.

TRUVEN HEALTH ANALYTICS
More Than Data. **Answers.**

- Contains information on enrollment and claims from private health plans and employers.

MARKETSCAN® RESEARCH

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

Implications

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

- Possibly allow for simplified formula and introduction of other estimation techniques
- Impact on aggressive diagnostic upcoding or fraud

In Practice, At Scale

Savannah Bergquist, Harvard PhD Student

- Medicare program with 10+ million enrollees
- Practicality of parsimony
- Computational considerations; parallel programming

Stay tuned!

In Practice, At Scale

Akritee Shrestha, Harvard Master's Student

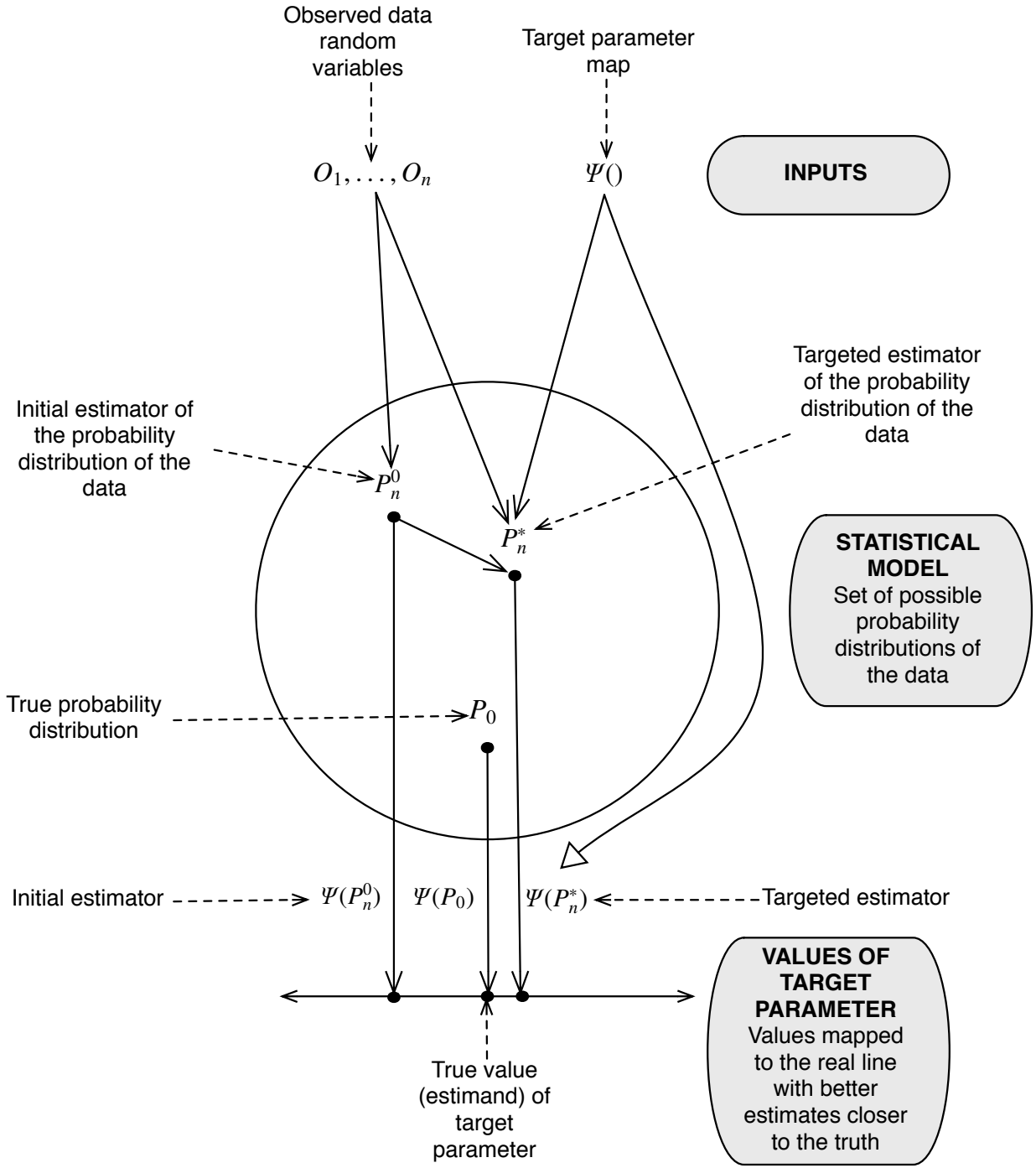
- Commercially insured with 2+ million enrollees
- Practicality of separate risk adjustment for mental health spending
- Computational considerations; trade-offs with full formula

Stay tuned!

Ensembles for Effect Estimation

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

Targeted Learning: Estimation framework featuring **targeted maximum likelihood estimators (TMLEs)** that incorporate **super learning**, an ensembled machine learning technique, for **effect estimation**.


Observed data random variables
Target parameter map
$O_1, \ldots, O_n$
$\Psi()$
INPUTS
Targeted estimator of the probability distribution of the data
Initial estimator of the probability distribution of the data
P_n^0
P_n^*
STATISTICAL MODEL
Set of possible probability distributions of the data
True probability distribution
P_0
Initial estimator
$\Psi(P_n^0)$
$\Psi(P_0)$
$\Psi(P_n^*)$
Targeted estimator
VALUES OF TARGET PARAMETER
Values mapped to the real line with better estimates closer to the truth
True value (estimand) of target parameter

R Packages

- `tmle` (Gruber): Main point-treatment TMLE package
- `ltmle` (Schwab): Main longitudinal TMLE package

More: `targetedlearningbook.com/software`

MarketScan Data

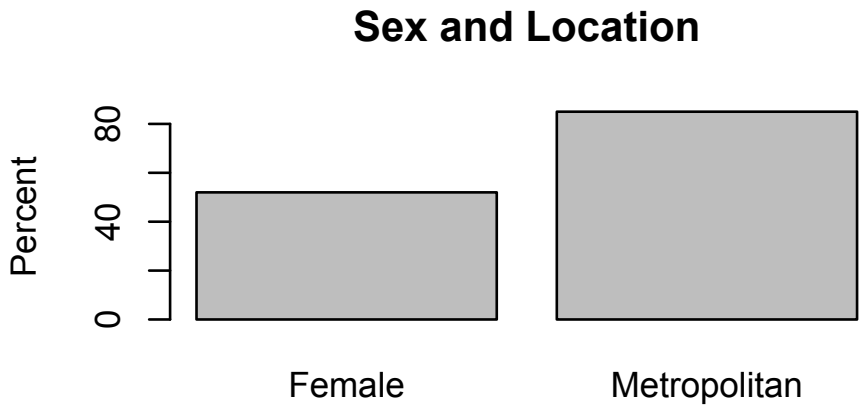
- **Truven MarketScan** database, those with continuous coverage in 2011-2012; 10.9 million people. Variables: age, sex, region, procedures, expenditures, etc.

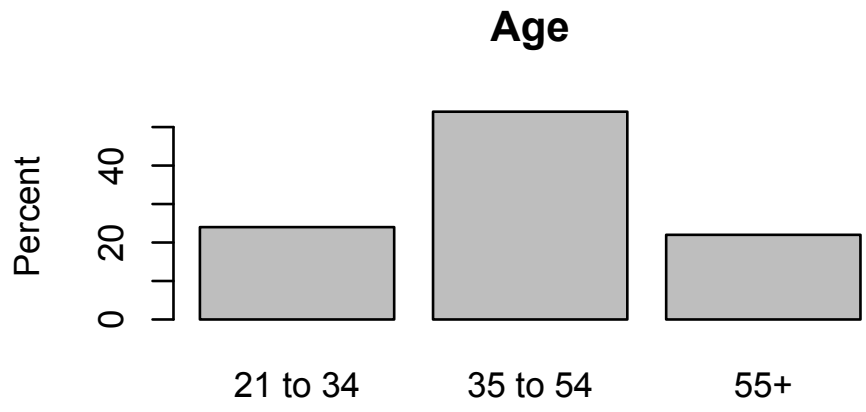
TRUVEN HEALTH ANALYTICS
More Than Data. **Answers.**

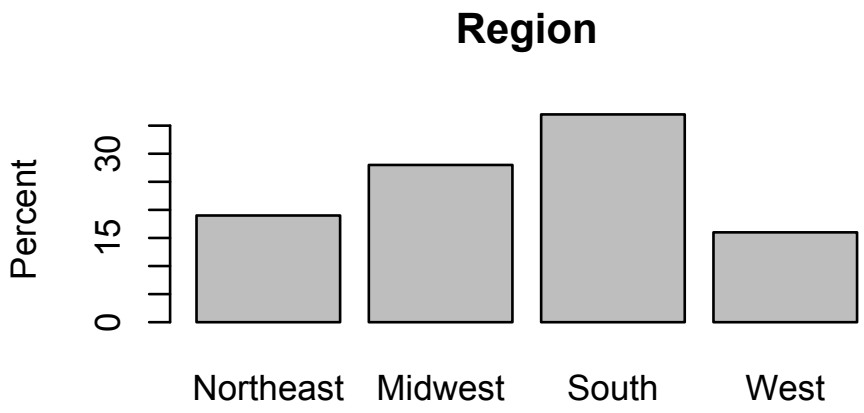
- Enrollment and claims from private health plans and employers.
- Extracted random sample of 1,000,000 people.
- Enrollees were eligible for insurance throughout this entire 24 month period and thus there is no drop-out due to death.

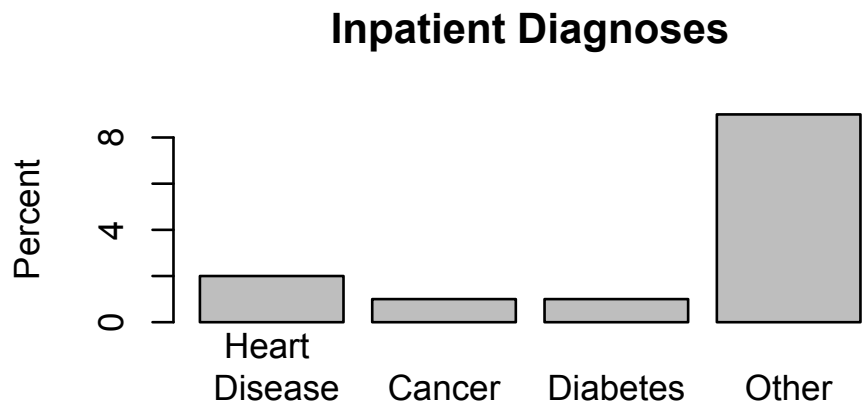
MARKETSCAN® RESEARCH

MarketScan Data Summary (n=1,000,000)

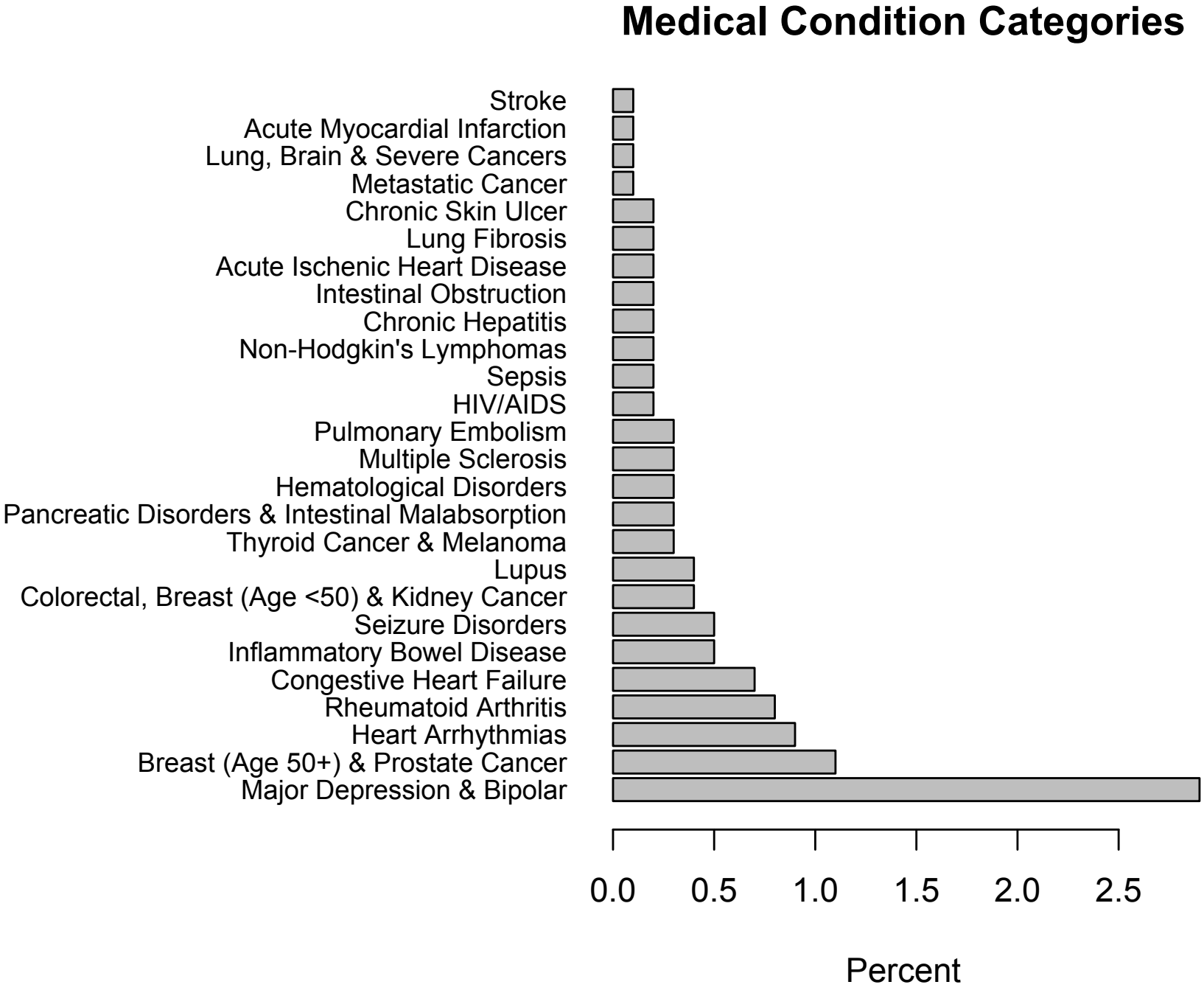








MarketScan Data Summary



Ensembles for Effect Estimation: Parameter

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

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

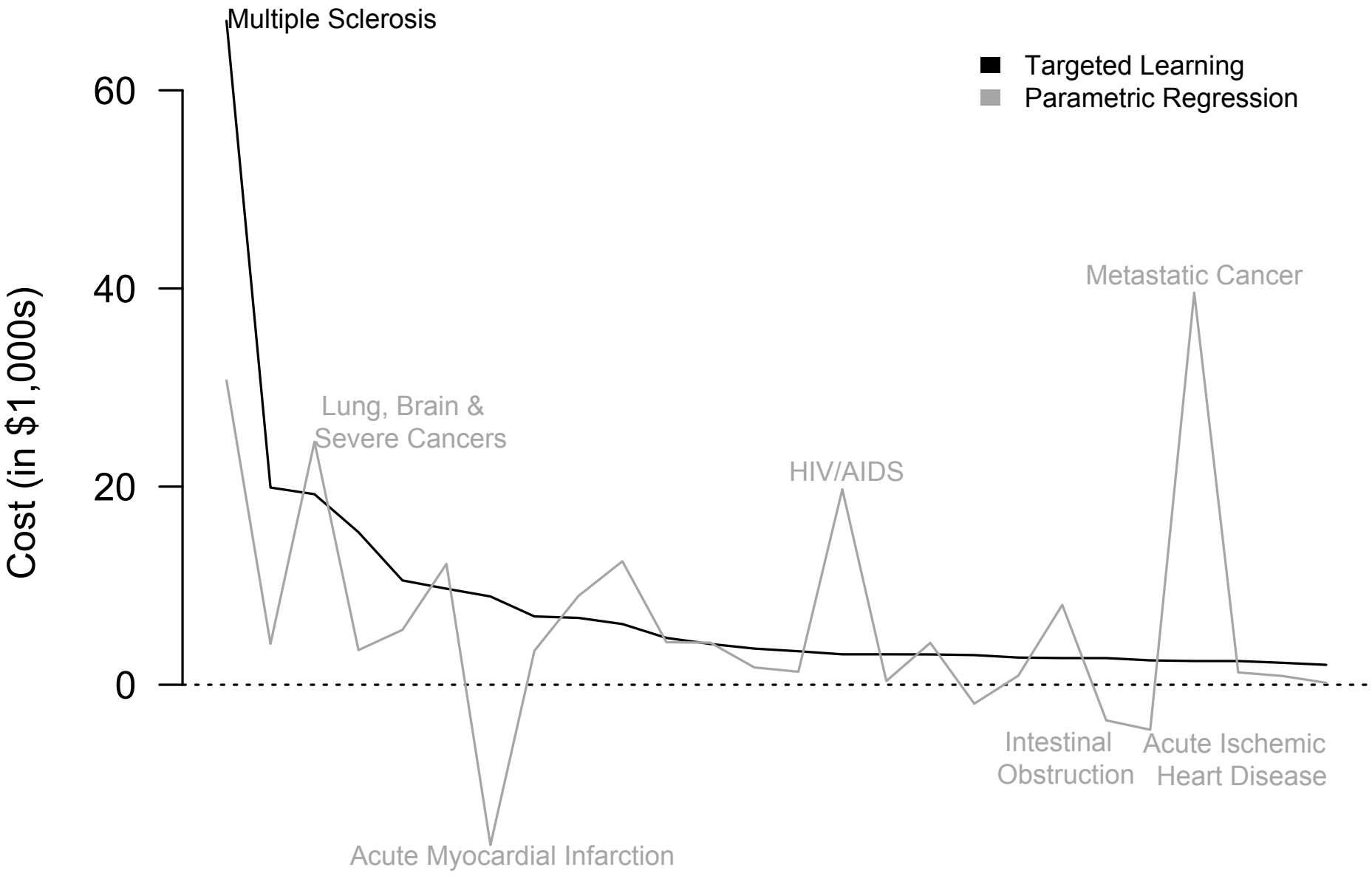
Ensembles for Effect Estimation

Targeted Learning: In Practice

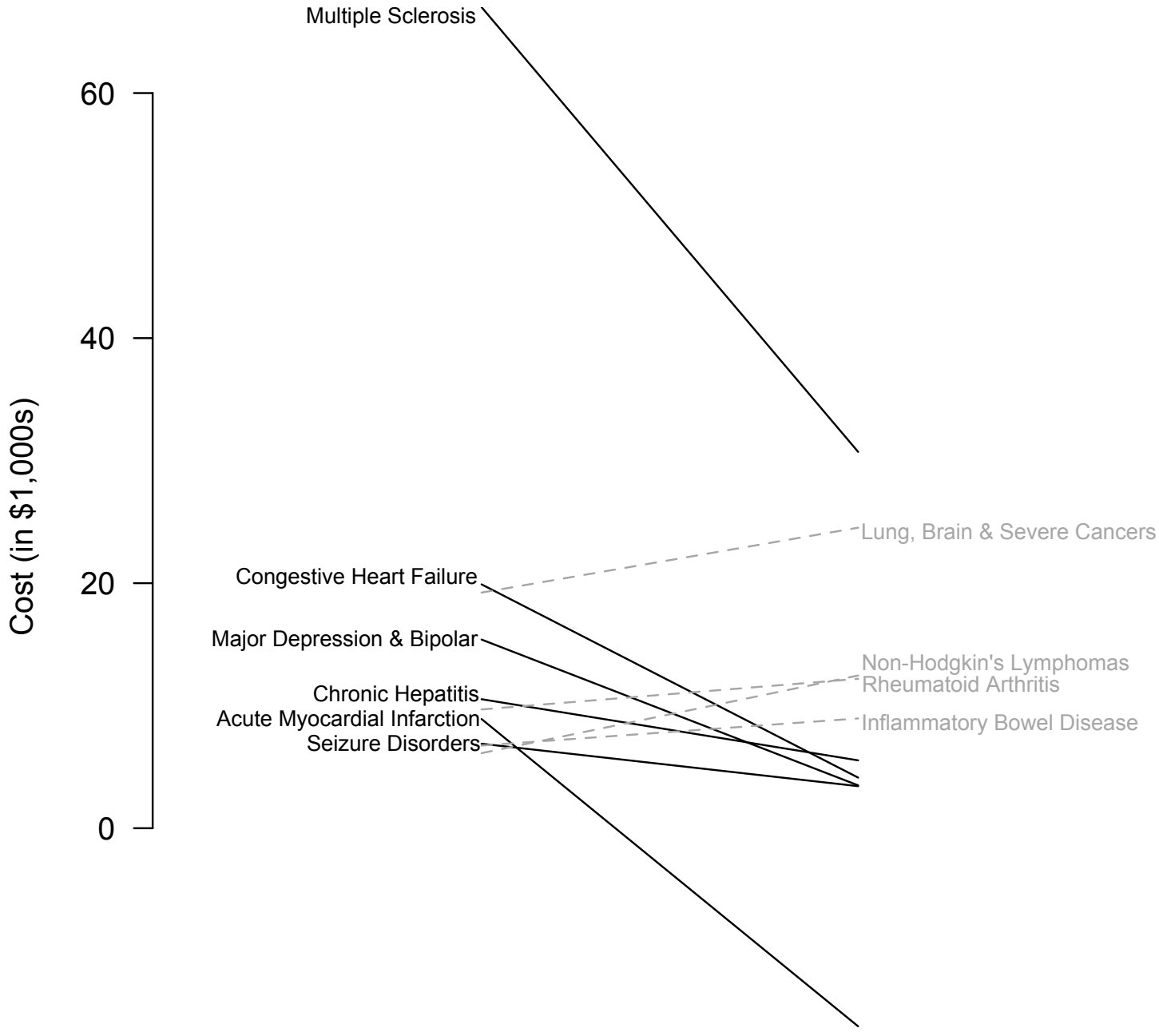
Allows the incorporation of machine learning methods for the estimation of both $E(Y \mid A, W, M^{-})$ and $P(A = 1 \mid W, M^{-})$ so that we do not make assumptions about the probability distribution P we do not believe.

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

Variable Importance Estimates



Variable Importance Estimates



Implications

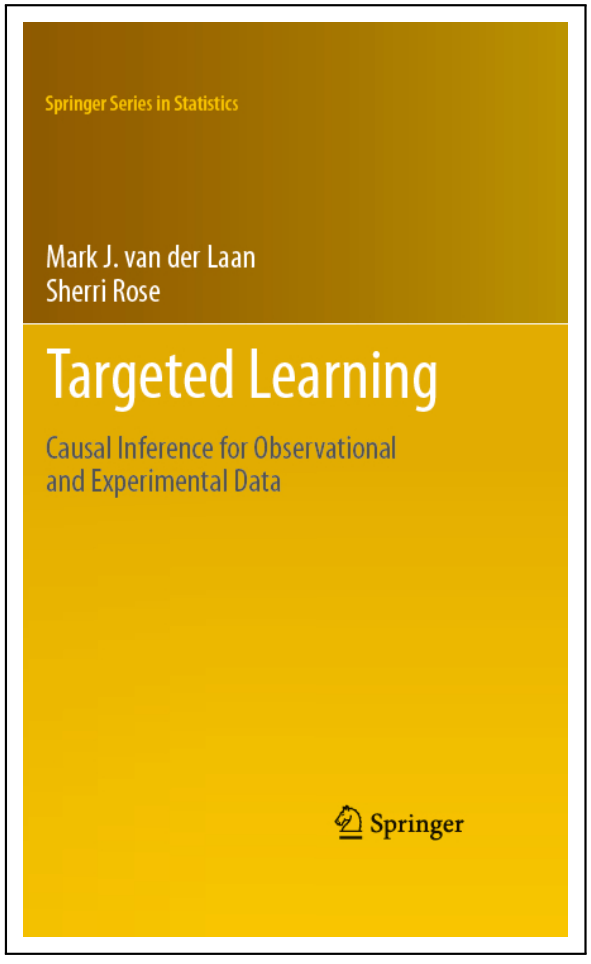
First investigation of the impact of medical conditions on health spending as a variable importance question using double robust estimators.

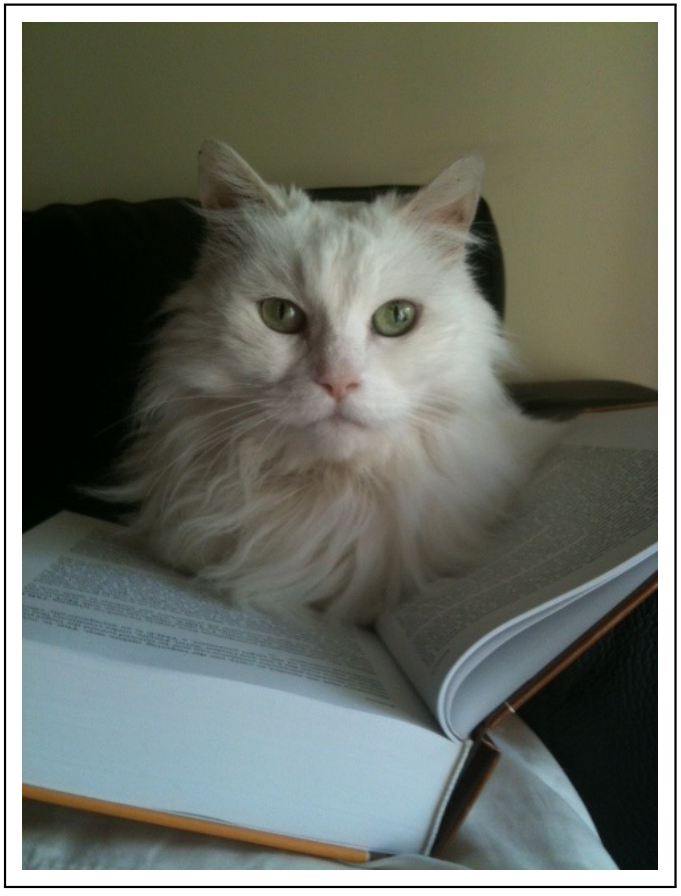
Five most expensive medical conditions were

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?

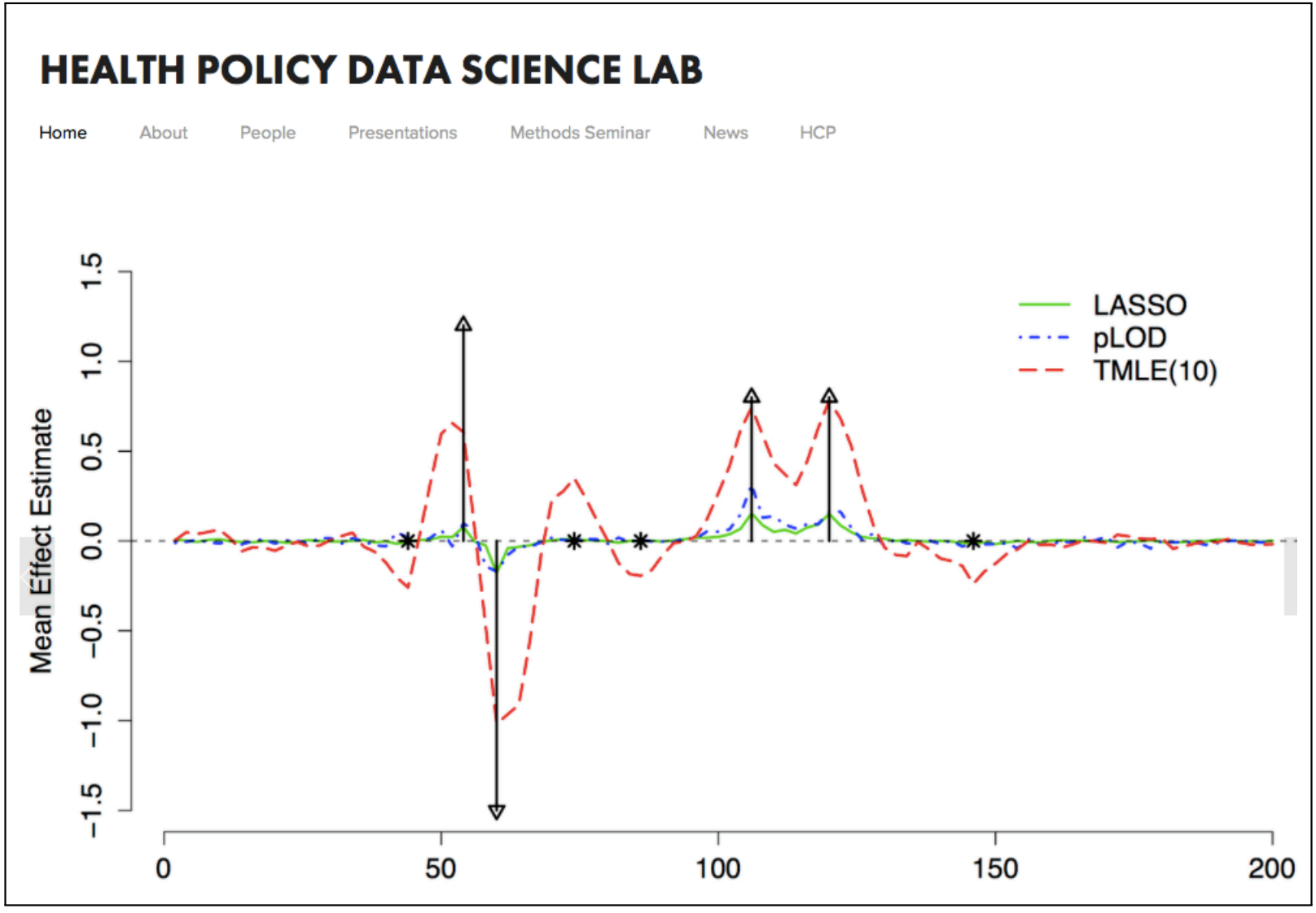
More on these methods targetedlearningbook.com





van der Laan & Rose, *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York: Springer, 2011.

More on health policy healthpolicydatascience.org



@HPDSLab

Ensembling Literature

- The super learner is a generalization of the stacking algorithm (Wolpert 1992, Breiman 1996) and has optimality properties that led to the name "super" learner.
- LeBlanc & Tibshirani (1996) discussed the relationship of stacking algorithms to other algorithms.
- Additional methods for ensemble learning have also been developed (e.g., Tsybakov 2003; Juditsky et al. 2005; Bunea et al. 2006, 2007; Dalayan & Tsybakov 2007, 2008).
- Refer to a review of ensemble methods (Dietterich 2000) for further background.
- van der Laan et al. (2007) original super learner paper.

Effect Estimation Literature

- Maximum-Likelihood-Based Estimators: g-formula, Robins 1986
- Estimating equations: Robins and Rotnitzky 1992, Robins 1999, Hernan et al. 2000, Robins et al. 2000, Robins 2000, Robins and Rotnitzky 2001.
- Additional bibliographic history found in Chapter 1 of van der Laan and Robins 2003.
- van der Laan and Rubin (2006) original TMLE paper.
- For even more references, see Chapter 4 of *Targeted Learning*.