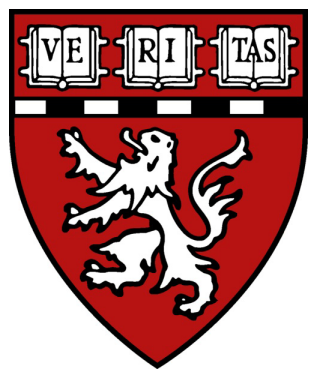
Machine Learning in Biostatistics & Health Policy

Sherri Rose

Associate Professor
Department of Health Care Policy
Harvard Medical School

rose@hcp.med.harvard.edu
drsherrirose.com

July 20, 2016



Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

The New York Times
nytimes.com

September 16, 2007

Do We Really Know What Makes Us Healthy?

By GARY TAUBES

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

The New York Times
nytimes.com

September 16, 2007

Do We Really Know What Makes Us Healthy?

By GARY TAUBES

AMSTATNEWS
The Membership Magazine of the American Statistical Association

Statistics Ready for a Revolution

Next Generation of Statisticians Must Build Tools for Massive Data Sets

Mark van der Laan, Jiann-Ping Hsu/Karl E. Peace Professor in Biostatistics and Statistics at UC Berkeley, and Sherri Rose, PhD candidate at UC Berkeley

variations

Big data and the future

At the beginning of her career **Sherri Rose** discusses big data and stands amazed at its potential.

4 high impact zones in statistical discovery with big data

FierceBigData

September 22, 2014

Clockwise from top left: Dunson, Rudin, McCormic and Rose

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" (pdf) that 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.

STATtr@k A website for new statistics professionals navigating a data-centric

Statisticians' Place in Big Data

– FEBRUARY 1, 2013
POSTED IN: DEVELOPMENT TR@K

Sherri Rose is an NSF mathematical sciences postdoctoral research fellow in the department of biostatistics at the Johns Hopkins Bloomberg School of Public Health.

Big Data has become the new buzz phrase in the world of information collection and analysis. The experiments we conduct and the observational data we collect continue to grow in size, due to rapidly expanding technology.

Large data sets also have drawn the attention of young people, with undergraduate and graduate students choosing computer science, engineering, and statistics for their programs of study. Each of these disciplines brings something unique to the table when discussing the challenges of Big Data, and interdisciplinary collaborations are becoming increasingly common.

statistician

tell me something
interesting (fast)

insert data

big data system

Idreos/Rose

Electronic Health Databases

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

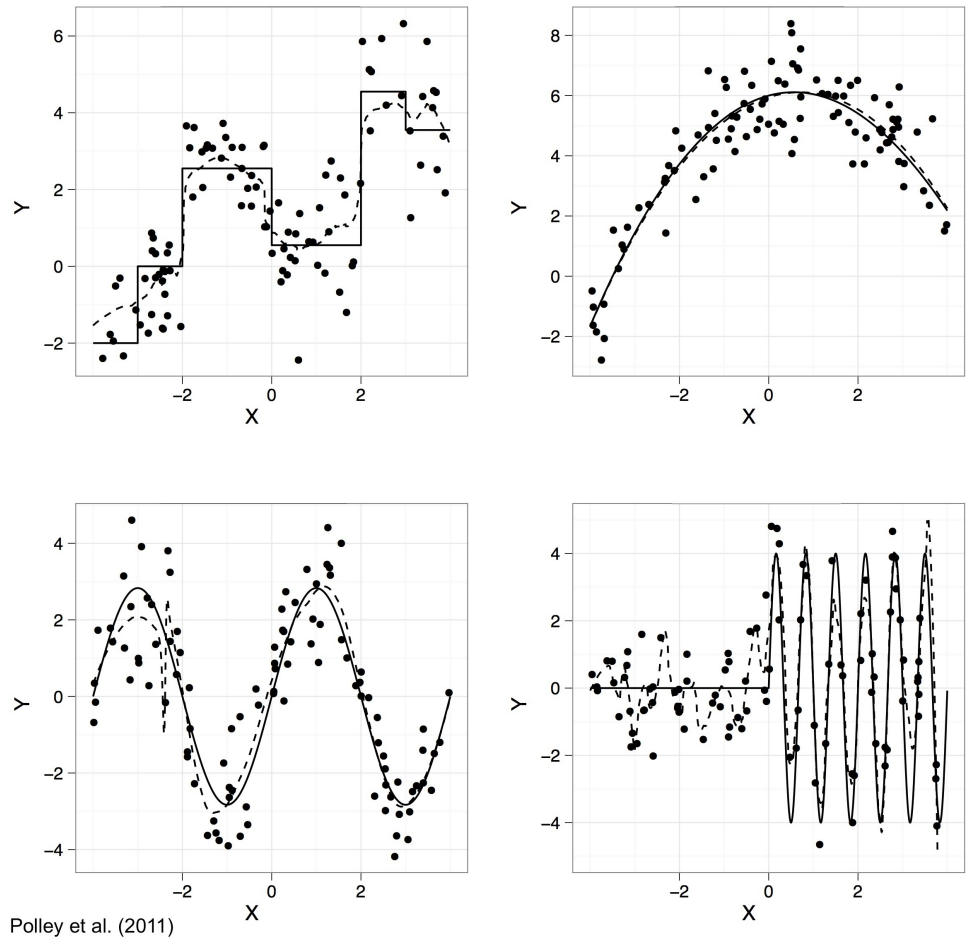
General usefulness of this type of data to answer targeted scientific research questions is an open question.

Need **novel statistical methods** that have desirable statistical properties while remaining computationally feasible.

Machine Learning: Big Picture

Machine learning aims to

- "smooth" over the data
- make fewer assumptions



Machine Learning: Options

- Recent health studies have employed newer **algorithms.**

(any mapping from data to a predictor)

Machine Learning: Options

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

Machine Learning: Options

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

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]

Machine Learning: Options

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

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

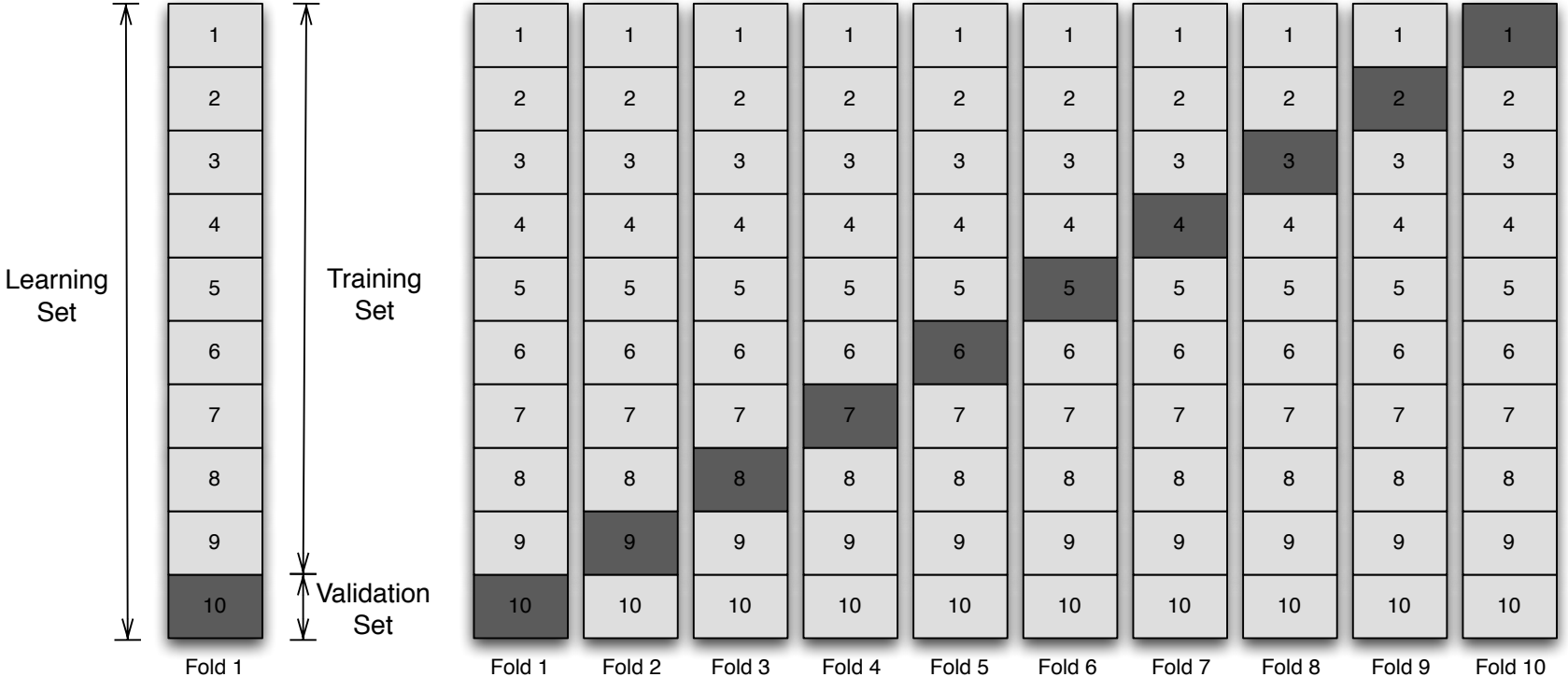
Random forest can predict 30-day mortality of spontaneous intracerebral hemorrhage with remarkable discrimination

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*

Machine Learning: Ensembles

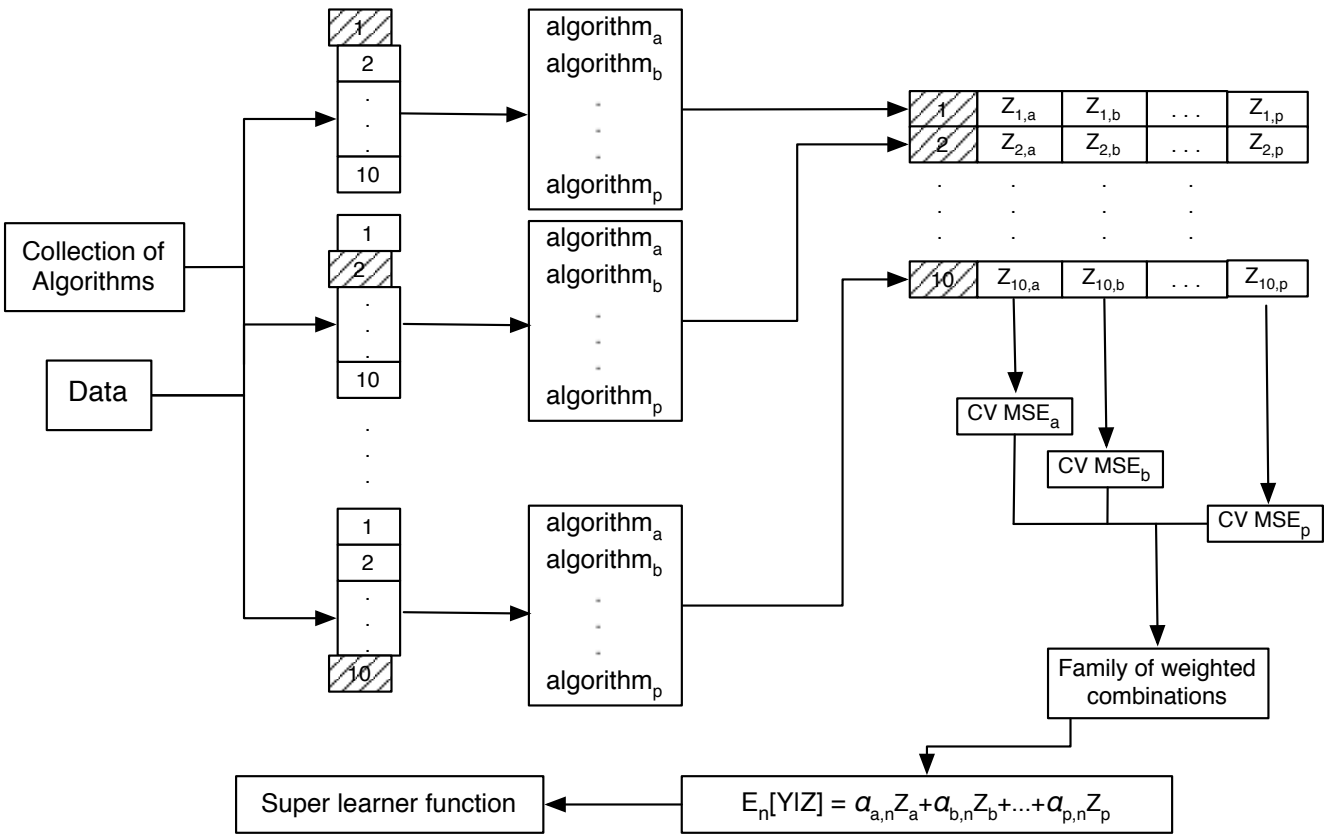
- Ensembling methods allow implementation of multiple algorithms.
- Do not need to decide beforehand which single technique to use; can use several by incorporating **cross-validation**.



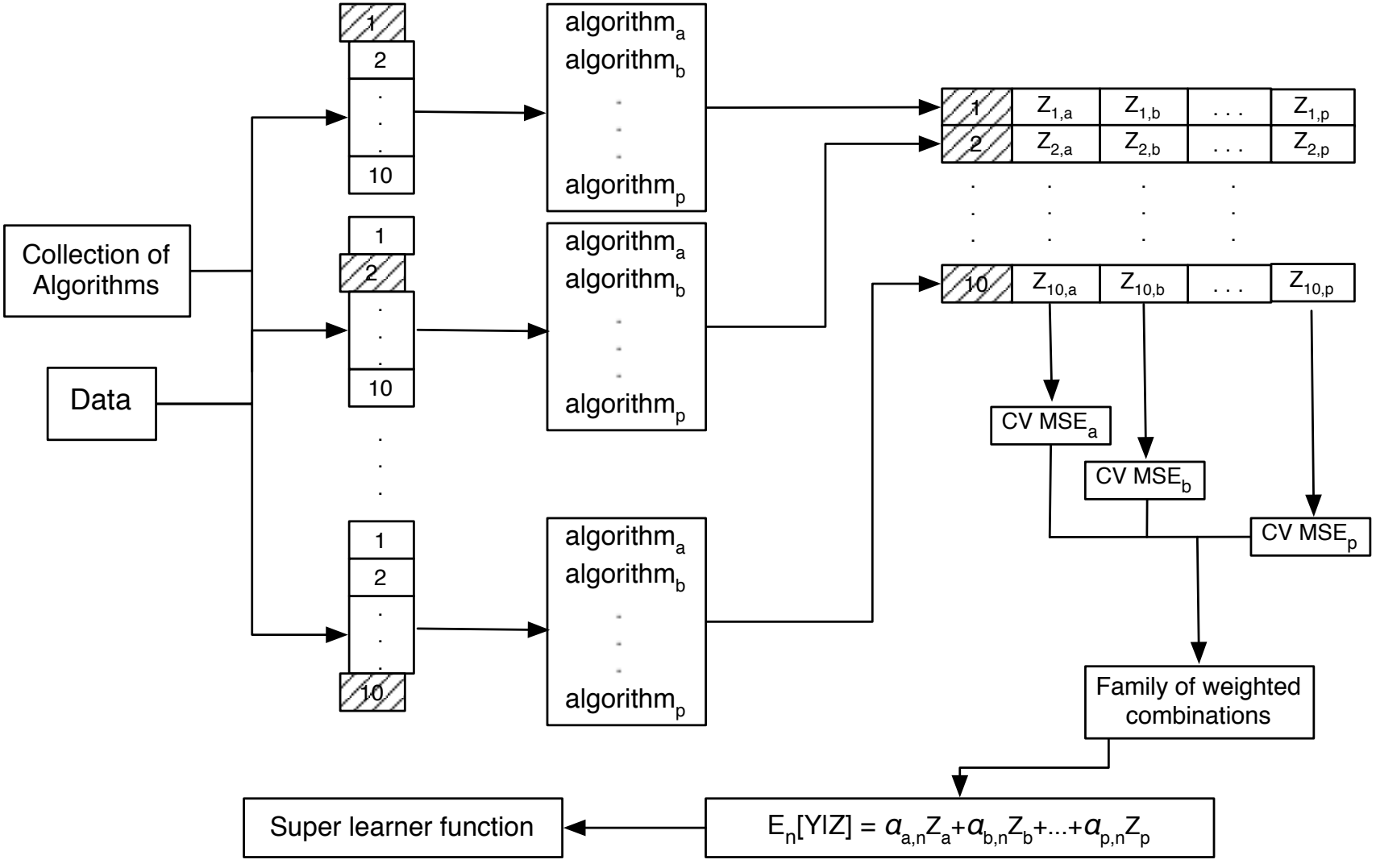
Machine Learning: Ensembles

Super Learner: 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.



Machine Learning: Ensembles



Applications in Health Insurance

1. Risk adjustment in plan payment
2. Effect estimation for variable importance of medical conditions

DATA

- **Truven MarketScan** database, up to 51 million enrollees per year. Variables: age, sex, region, procedures, expenditures, etc.

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

- Enrollment and claims from private health plans and employers.

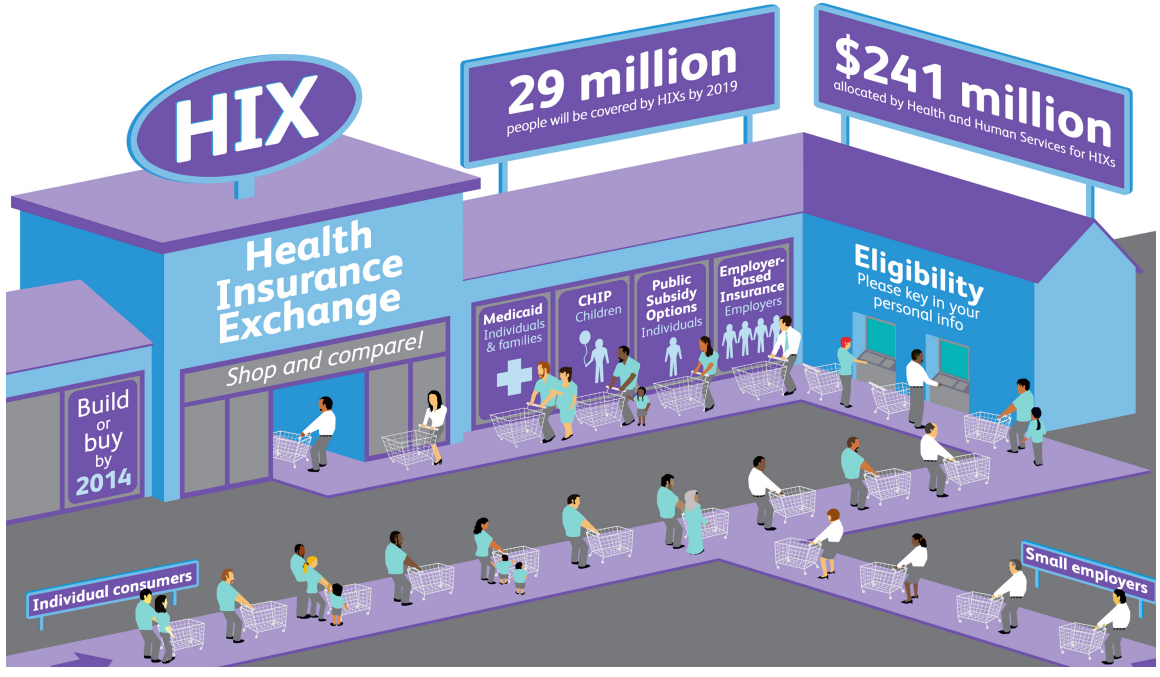
MARKETSCAN® RESEARCH

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 health
- Encourages competition based on efficiency & quality
- Huge financial implications

Health Insurance Marketplace



xerox.com

Risk Adjustment in Plan Payment = Frozen

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

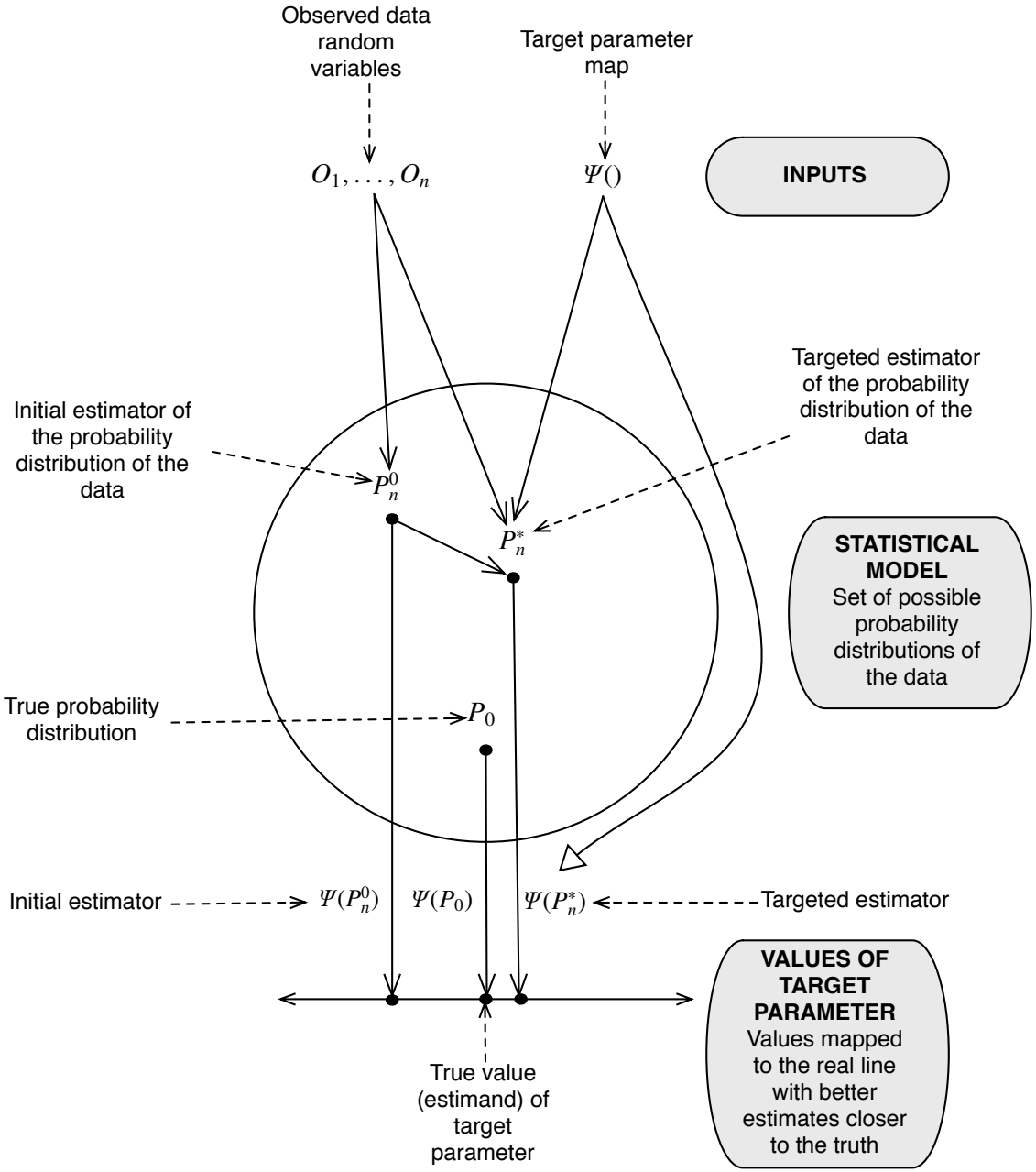
Potentially *$$$* oversight, where it attempts to control for the impact of consumers choosing health plans.

wpb.org

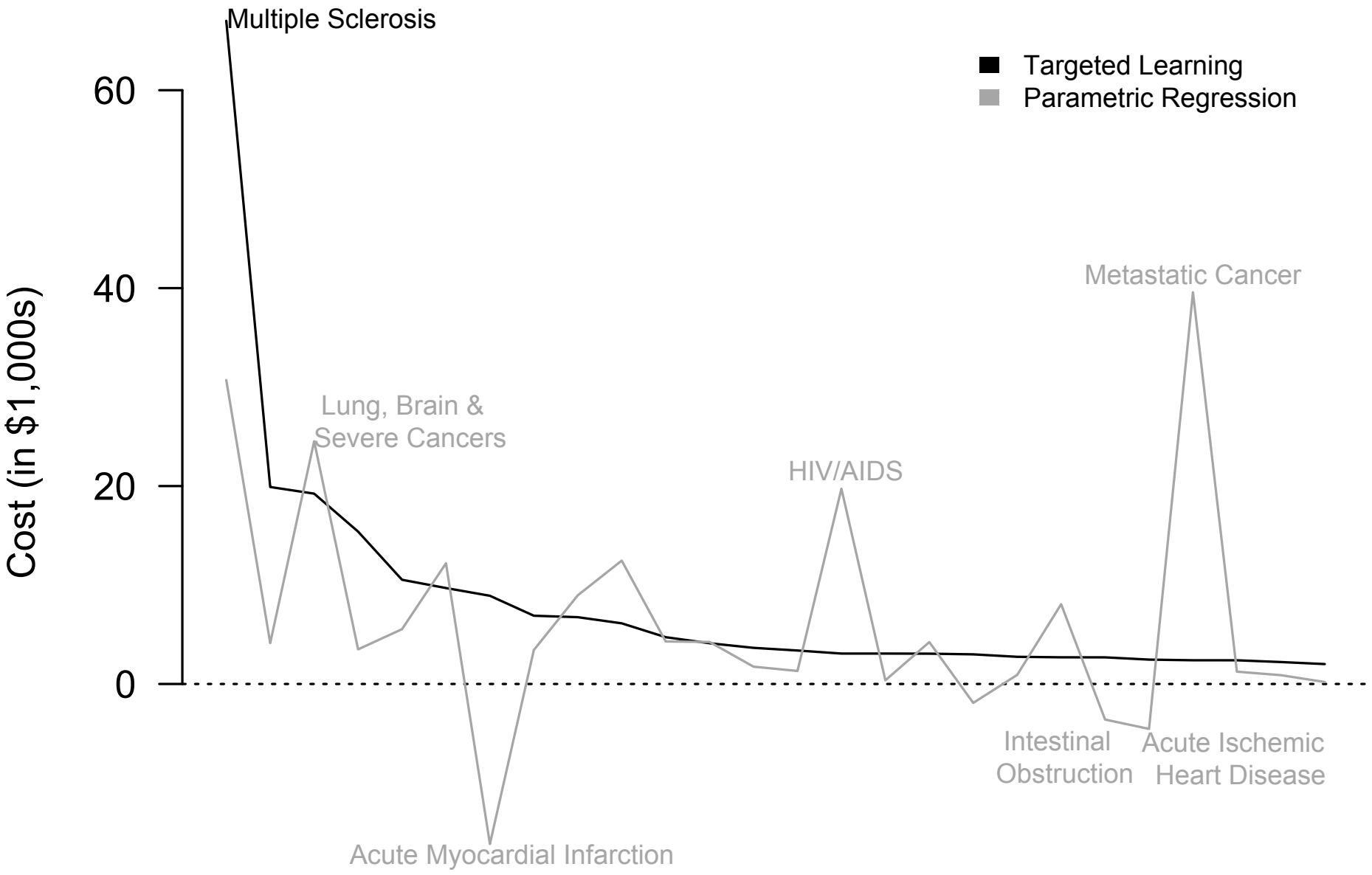
Risk Adjustment in Plan Payment: Results

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

Medical Condition Variable Importance



Medical Condition Variable Importance: Results



Medical Condition Variable Importance: Results

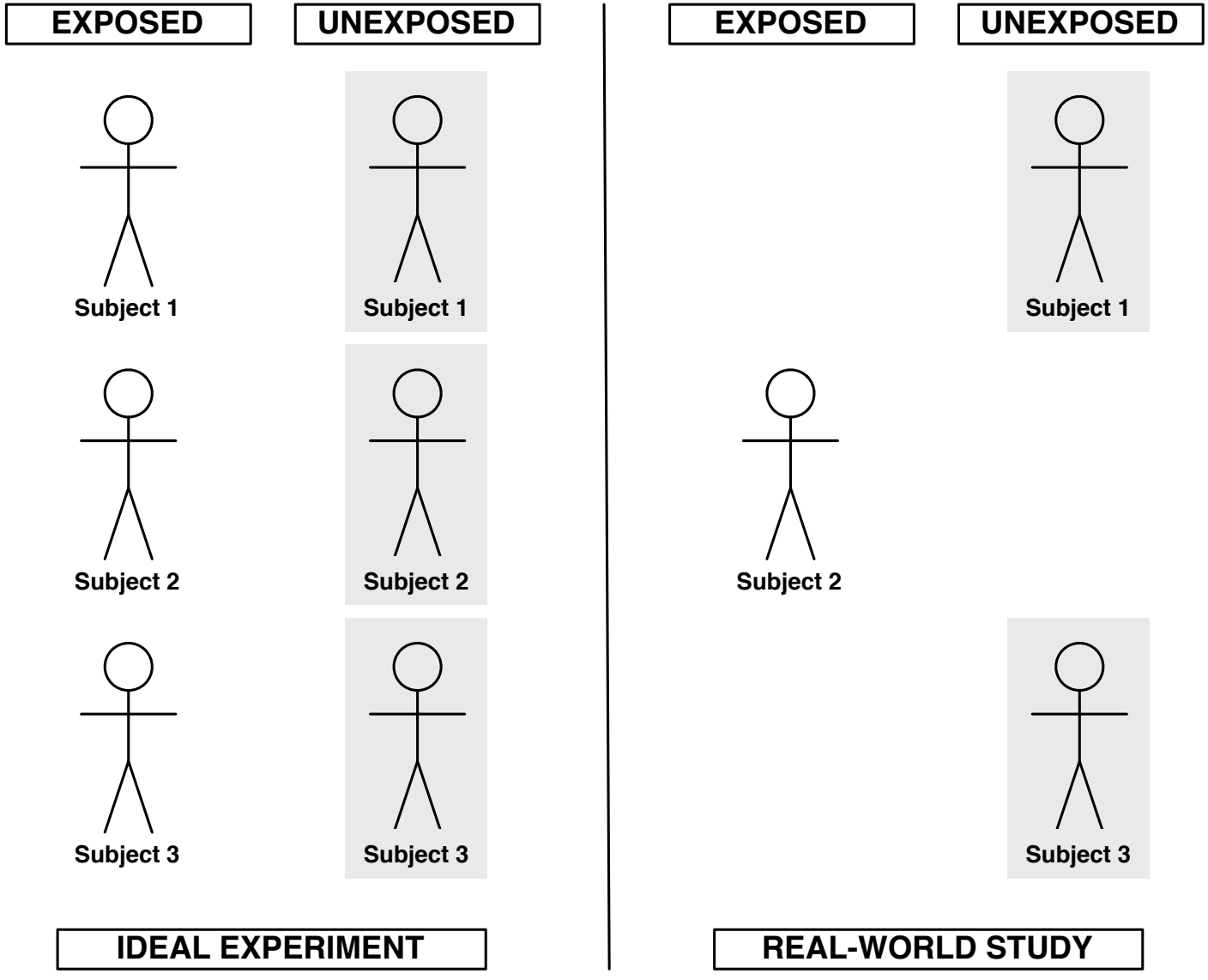
First investigation of the impact of medical conditions on health spending as a variable importance question using double robust estimators.

Five most expensive medical conditions were

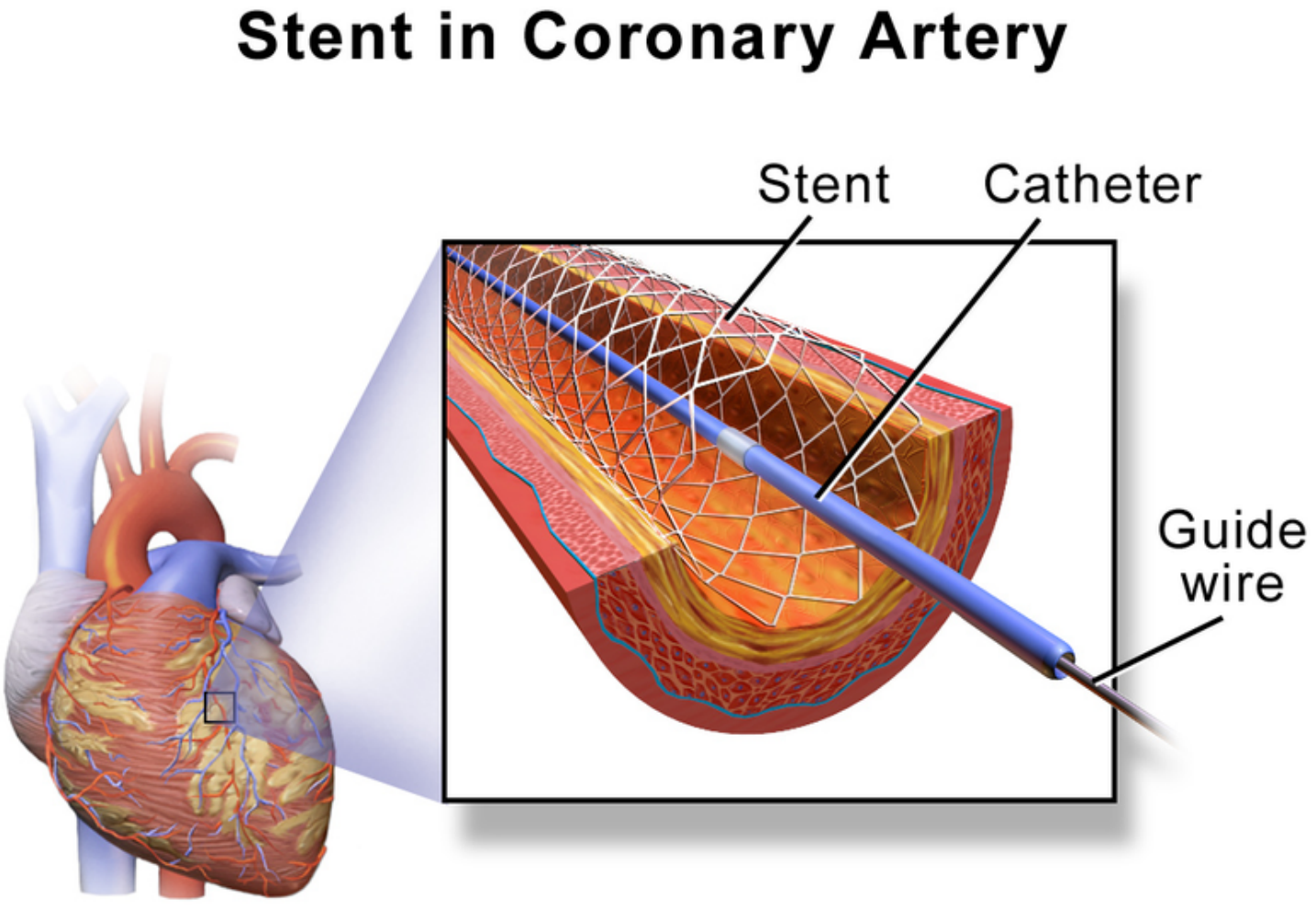
1. multiple sclerosis
2. congestive heart failure
3. lung, brain, and other severe cancers
4. **major depression and bipolar disorders**
5. **chronic hepatitis.**

- Differing results compared to parametric regression.
- What does this mean for incentives for prevention and care?

Causal Questions

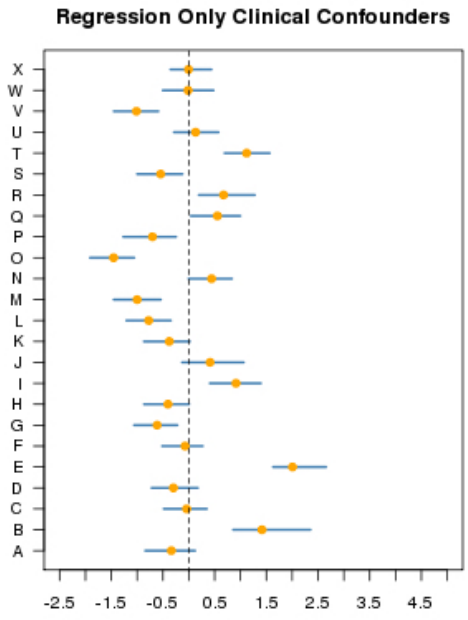


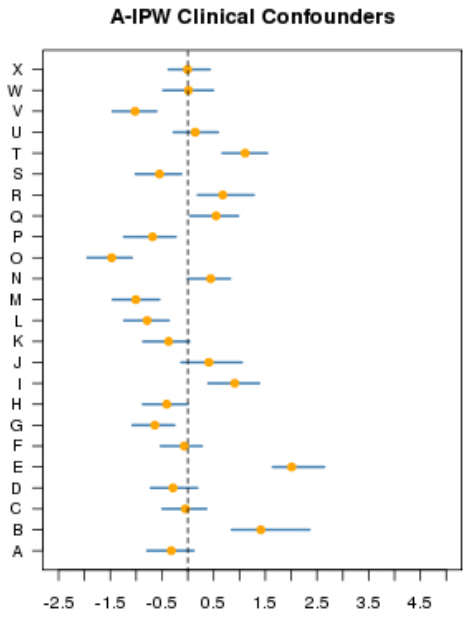
Comparative Effectiveness

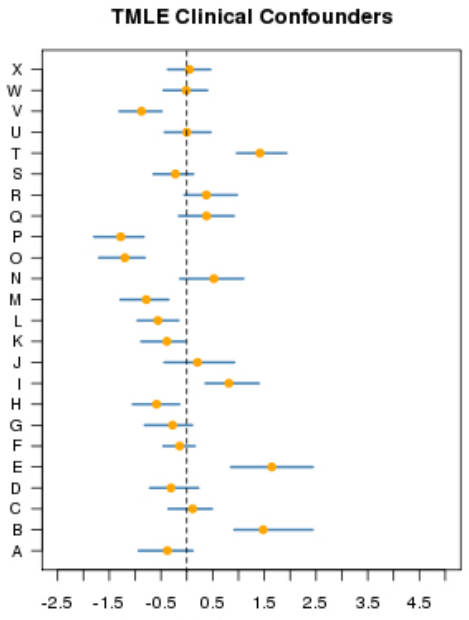


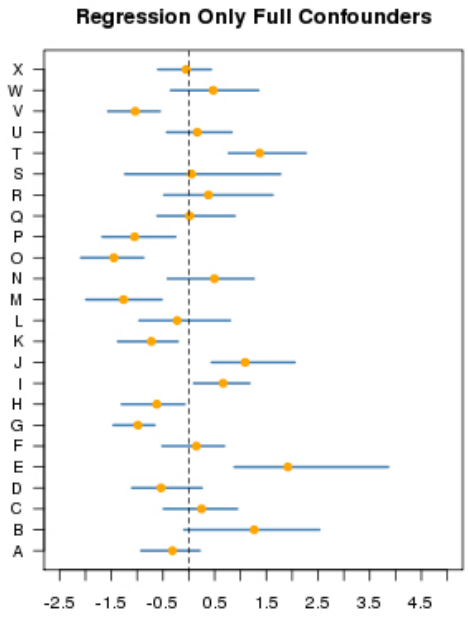
HatfieldBlausen.com staff. "Blausen gallery 2014"

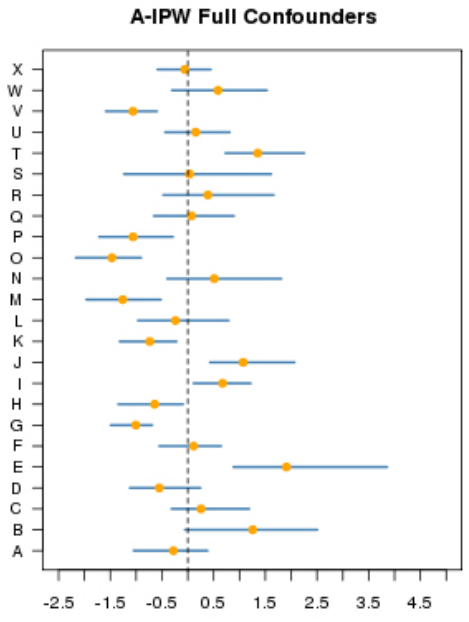
Hospital Profiling

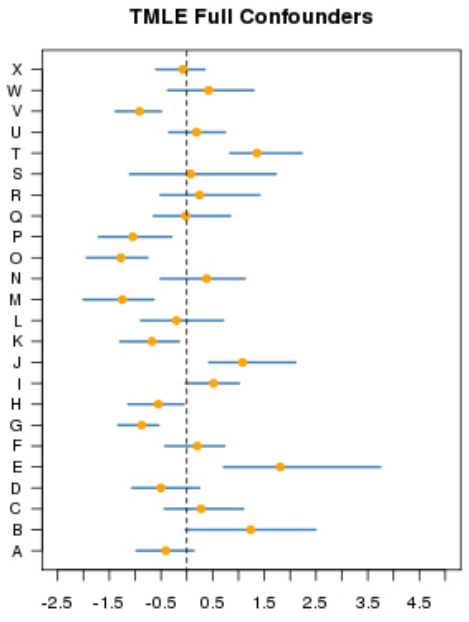




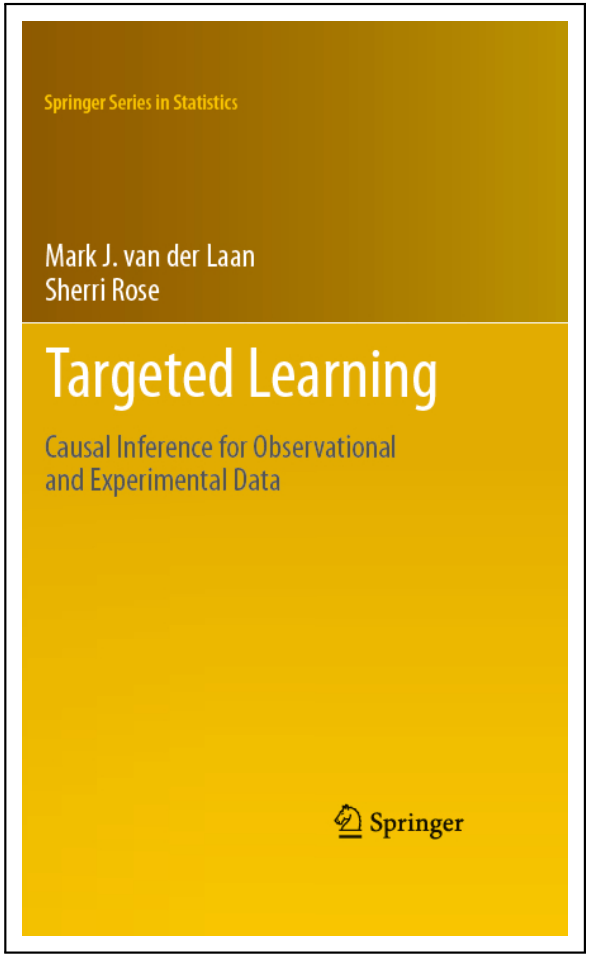


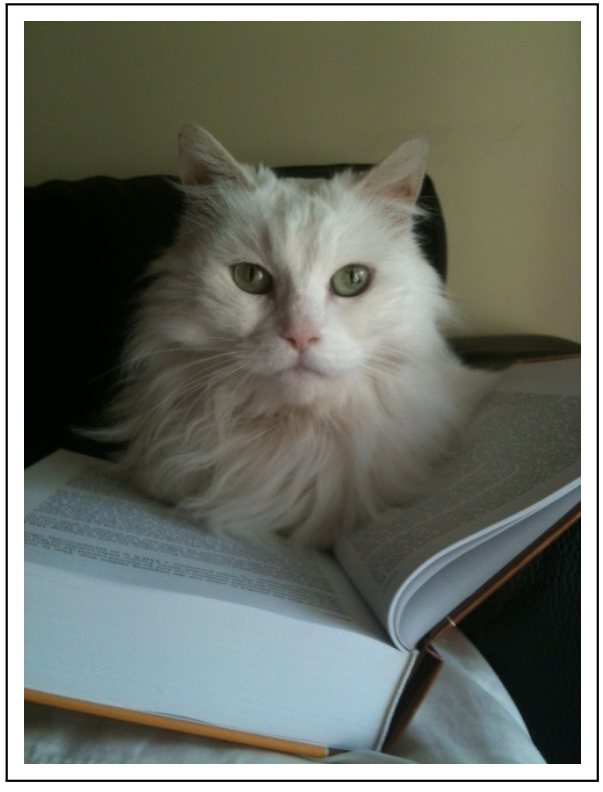






Targeted Learning Methods

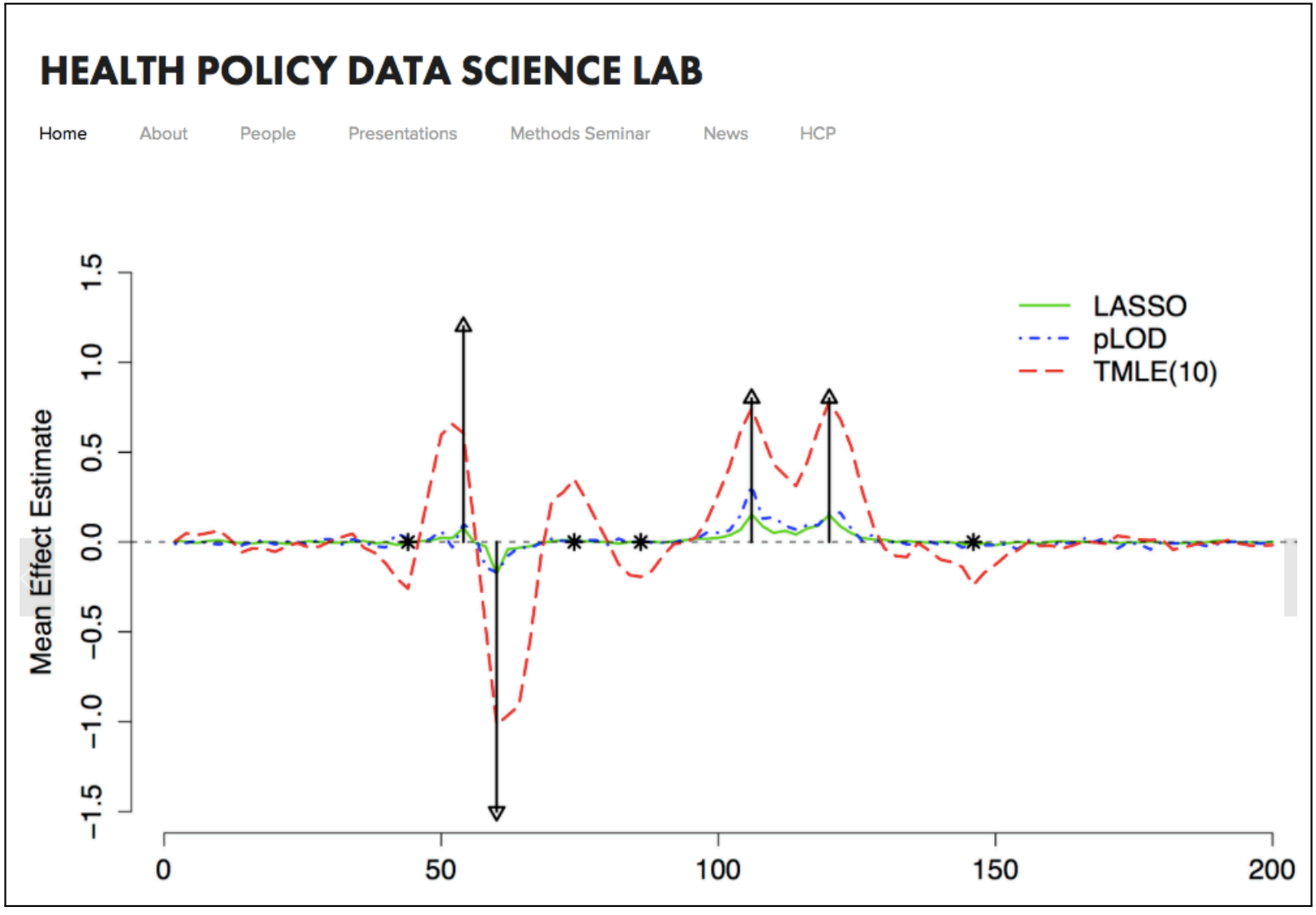




van der Laan & Rose, *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York: Springer, 2011.

`targetedlearningbook.com`

Health Policy Data Science Lab



healthpolicydatascience.org