Toward Standards for Machine Learning Research in Health Care and Policy

The National Academies of SCIENCES ENGINEERING MEDICINE

October 24, 2019

Sherri Rose, Ph.D.

Associate Professor
Department of Health Care Policy
Harvard Medical School

Co-Director
Health Policy Data Science Lab

drsherrirose.org
@sherrirose

VE RI TAS

HPDS Lab

“ Learning two fields takes, surprisingly, twice as long as learning one. But it’s worth the investment because you get to solve real problems for the first time. ”

Barbara Engelhardt | Princeton

“ In both private enterprise and the public sector, research must be reflective of the society we’re serving. ”

Rediet Abebe | Harvard & Cornell

“ …behind every data point there is a human story, there is a family, and there is suffering. ”

Nick Jewell | LSHTM & UC Berkeley

DATA

DATA

Electronic Databases

The increasing availability of electronic health information offers a **resource to health researchers**

Electronic Databases

The increasing availability of electronic health information offers a **resource to health researchers**

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

Electronic Databases

The increasing availability of electronic health information offers a **resource to health researchers**

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

Electronic Databases

The increasing availability of electronic health information offers a **resource to health researchers**

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

Electronic Databases

The increasing availability of electronic health information offers a **resource to health researchers**

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

May need **novel statistical methods** that have desirable properties while remaining computationally feasible

Electronic Databases: EHR $\neq$ EHR

The increasing availability of electronic health information offers a **resource to health researchers**

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

May need **novel statistical methods** that have desirable properties while remaining computationally feasible

BILLING CLAIMS

CLINICAL RECORDS

IMAGING

REGISTRIES

SURVEYS

GOVERNMENT SOURCES

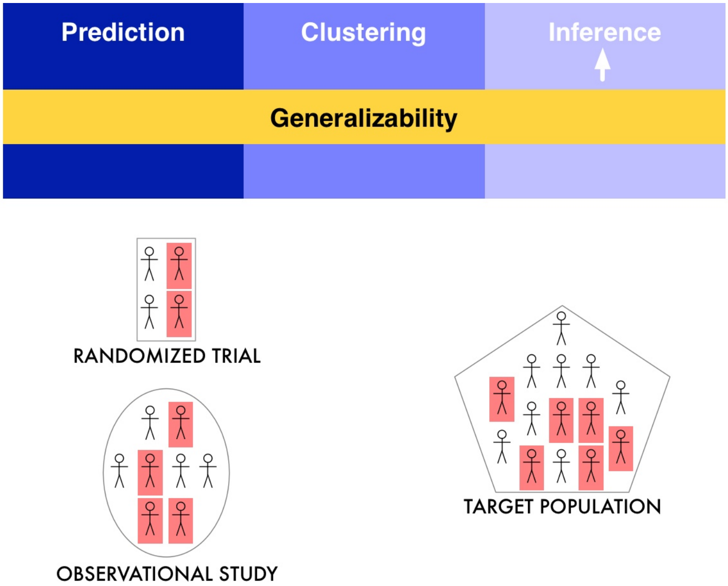
WEARABLE & IMPLANTABLE TECH

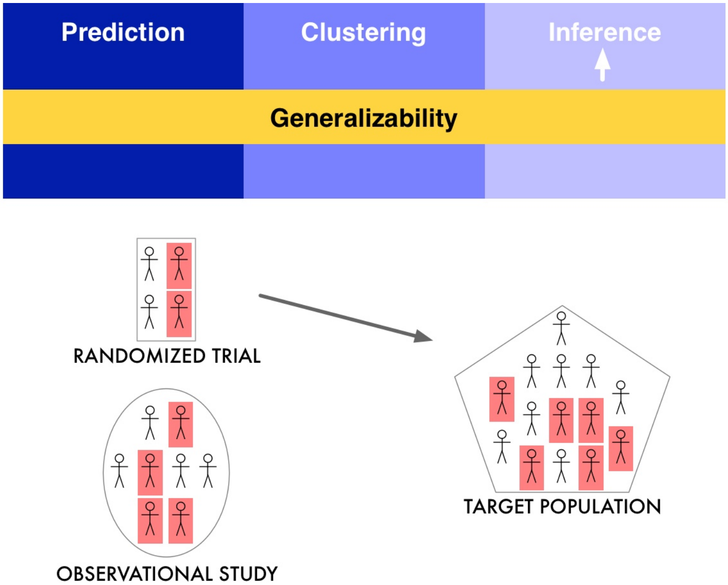
DIGITAL

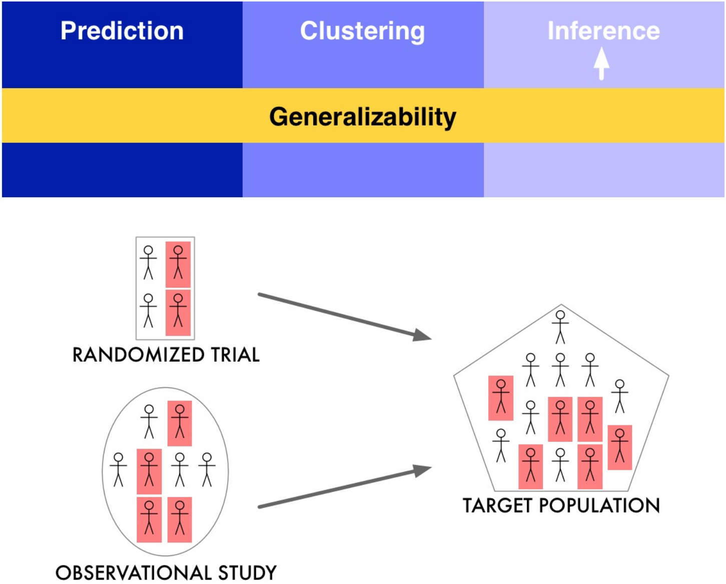
NEWS MEDIA

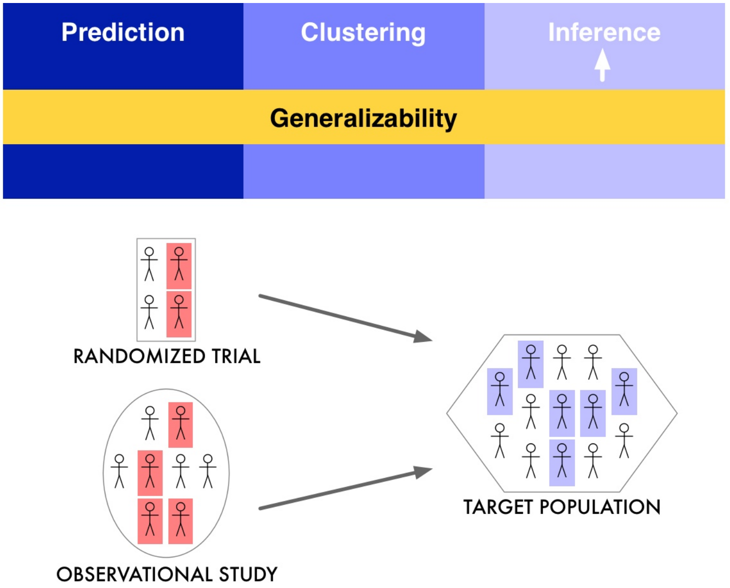
GENERALIZABILITY

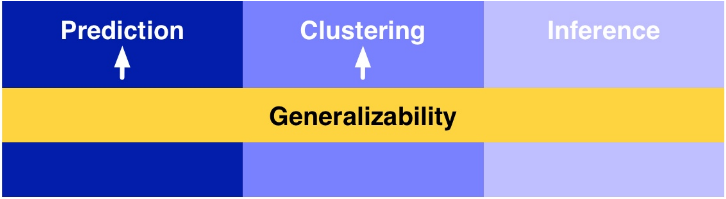
Prediction	Clustering	Inference
Generalizability		

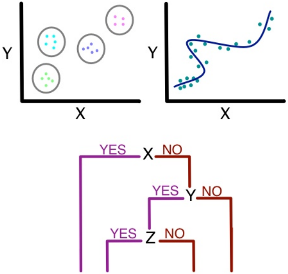

Prediction
Clustering
Inference
Generalizability
RANDOMIZED TRIAL
OBSERVATIONAL STUDY
TARGET POPULATION


Prediction
Clustering
Inference
Generalizability
RANDOMIZED TRIAL
OBSERVATIONAL STUDY
TARGET POPULATION


Prediction
Clustering
Inference
Generalizability
RANDOMIZED TRIAL
OBSERVATIONAL STUDY
TARGET POPULATION

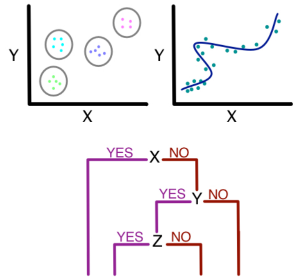

Prediction
Clustering
Inference
Generalizability
RANDOMIZED TRIAL
OBSERVATIONAL STUDY
TARGET POPULATION


Prediction
Clustering
Inference
Generalizability


Y
X
Y
X
YES
X
NO
YES
Y
NO
YES
Z
NO

Prediction | **Clustering** | **Inference**

Generalizability

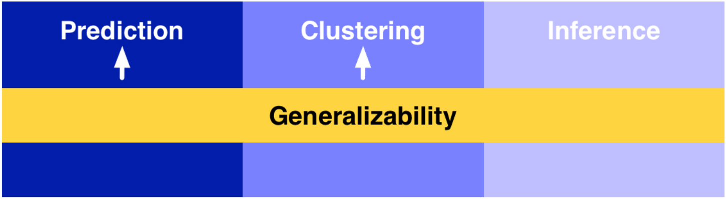


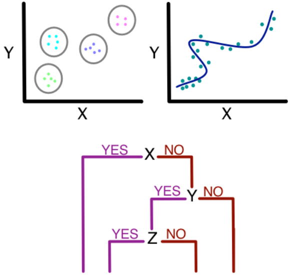
JAMA Network | Open

Invited Commentary | Health Informatics

Machine Learning for Prediction in Electronic Health Data

Sherri Rose, PhD





JAMA Network Open

Invited Commentary | Health Informatics

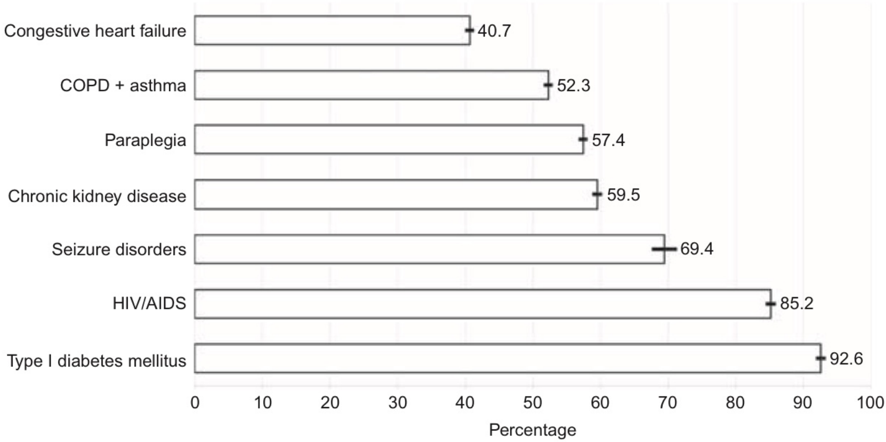
Machine Learning for Prediction in Electronic Health Data

Sherri Rose, PhD

"The machine learning researchers who develop novel algorithms for prediction and the clinical teams interested in implementing them are frequently and unfortunately 2 nonintersecting groups."

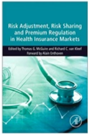
DATASET SHIFT

Chronic Conditions

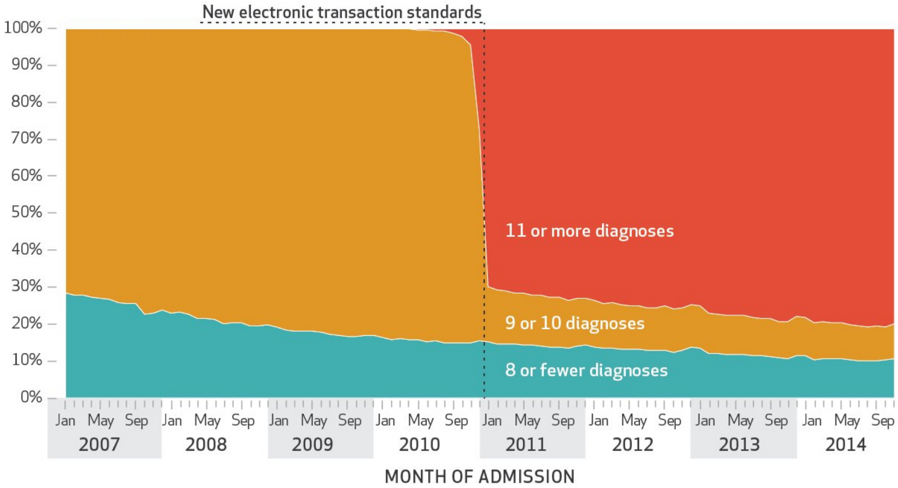


Risk Adjustment for Health Plan Payment

Randall P. Ellis , Bruno Martins and Sherri Rose



Number of Diagnoses Reported

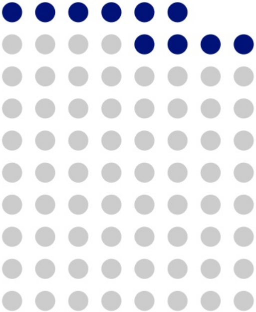


Christopher Ody et al. Health Aff 2019; 38:39

Health Affairs

Variable Selection and Upcoding

Reduced set of 10 variables 92% as efficient

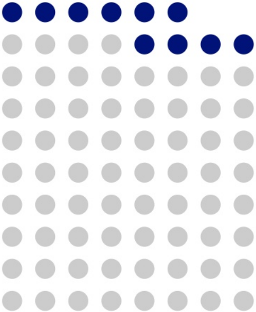


A Machine Learning Framework for Plan Payment Risk Adjustment

Sherri Rose

Variable Selection and Upcoding

~~Reduced set of 10 variables 92% as efficient~~



"…results for the risk adjustment algorithms that considered a limited subset of variables…performed consistently worse across all benchmarks."

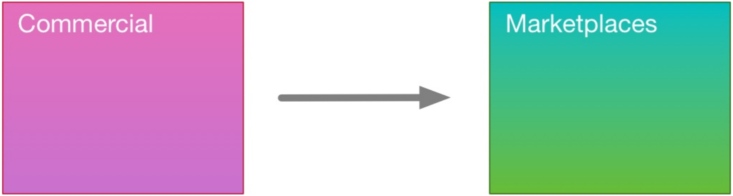
Sample Selection for Medicare Risk Adjustment Due to Systematically Missing Data

Savannah L. Bergquist, Thomas G. McGuire, Timothy J. Layton, and Sherri Rose

A Machine Learning Framework for Plan Payment Risk Adjustment

Sherri Rose

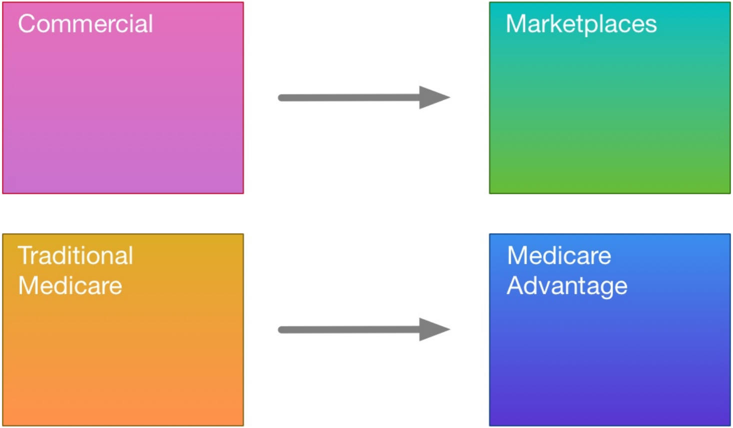
Prediction Using the "Wrong" Data



Matching and Imputation Methods for Risk Adjustment in the Health Insurance Marketplaces

Sherri Rose Julie Shi Thomas G. McGuire
Sharon-Lise T. Normand

Prediction Using the “Wrong” Data



Sample Selection for Medicare Risk Adjustment Due to Systematically Missing Data

Savannah L. Bergquist, Thomas G. McGuire, Timothy J. Layton, and Sherri Rose

Matching and Imputation Methods for Risk Adjustment in the Health Insurance Marketplaces

Sherri Rose   Julie Shi   Thomas G. McGuire
Sharon-Lise T. Normand

FAIRNESS

Who decides the research question?

Who is in the target population?

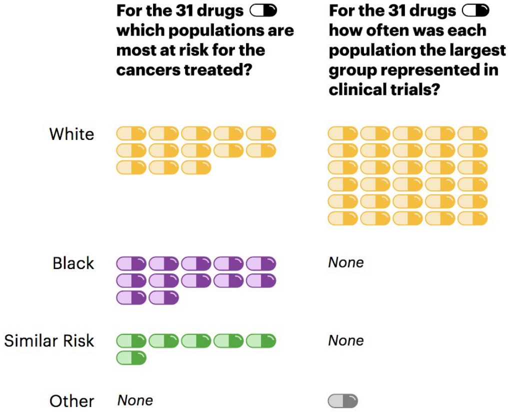
What do the data reflect?

How will the algorithm be assessed?


Prediction
Clustering
Inference
Generalizability
Fairness

Black Patients Miss Out On Promising Cancer Drugs

A ProPublica analysis found that black people and Native Americans are under-represented in clinical trials of new drugs, even when the treatment is aimed at a type of cancer that disproportionately affects them.



Note: Drugs are labeled "Similar Risk" if black Americans are at least 80 percent as likely as white Americans to be diagnosed with the cancer treated.

Chen and Wong (2018)

The NEW ENGLAND JOURNAL of MEDICINE

Perspective

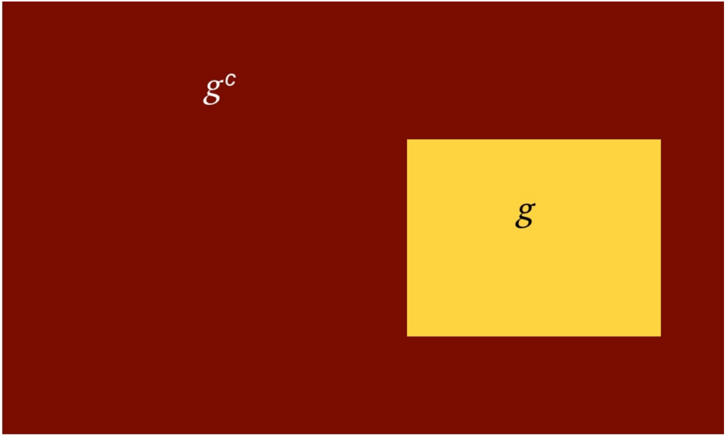
Machine Learning and Prediction in Medicine — Beyond the Peak of Inflated Expectations

Jonathan H. Chen, M.D., Ph.D., and Steven M. Asch, M.D., M.P.H.

Yet there are problems with real-world data sources. Whereas conventional approaches are largely based on data from cohorts that are carefully constructed to mitigate bias, emerging data sources are typically less structured, since they were designed to serve a different purpose (e.g., clinical care and billing). Issues ranging from patient self-selection to confounding by indication to inconsistent availability of outcome data can result in inadvertent bias, and even racial profiling, in machine predictions. Awareness of such challenges may keep the hype from outpacing the hope for how data analytics can improve medical decision making.

Algorithmic Fairness

Common measures of fairness are based on the notion of **group fairness**, striving for similarity in predicted outcomes or errors for groups



Global vs. Group Fit

	R^2	MHSUD Net Compensation
1. baseline formula	13.1%	-$2,822

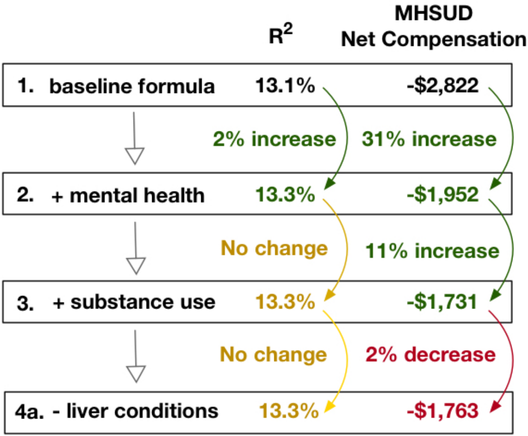
Global vs. Group Fit

	R^2	MHSUD Net Compensation
1. baseline formula	13.1%	-$2,822
	2% increase	31% increase
2. + mental health	13.3%	-$1,952

Limitations of *P*-Values and *R*-Squared for Stepwise Regression Building: A Fairness Demonstration in Health Policy Risk Adjustment

Sherri Rose and Thomas G. McGuire

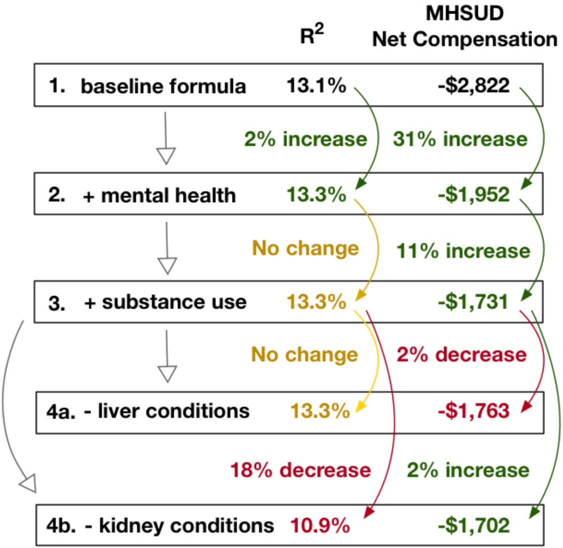
Global vs. Group Fit



Limitations of *P*-Values and *R*-Squared for Stepwise Regression Building: A Fairness Demonstration in Health Policy Risk Adjustment

Sherri Rose and Thomas G. McGuire

Global vs. Group Fit



Limitations of *P*-Values and *R*-Squared for Stepwise Regression Building: A Fairness Demonstration in Health Policy Risk Adjustment

Sherri Rose and Thomas G. McGuire

MORE ON METRICS

How Do We Evaluate Classifiers?

Area Under the Receiver Operating Characteristic Curve (AUC):

Summary metric of the predictive discrimination, specifically measuring the ranking performance for random discordant pairs

- Assessing prediction performance primarily using AUC can be misleading
- **Leaderboard AUC:** Despite many published warnings, machine learning competitions and articles often assign their leaderboard and winners solely on a single metric — often AUC for classification

How Do We Evaluate Classifiers?

$$\text{True Positive Rate} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

also known as:
Sensitivity and Recall

$$\text{False Positive Rate} = \frac{\text{False Positives}}{\text{False Positives} + \text{True Negatives}}$$

also known as:
1-Specificity

$$\text{Positive Predictive Value} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

also known as:
Precision

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{n}$$

... and more, including **calibration**.

Aortic Valves Study

$$\text{True Positive Rate} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

also known as:
Sensitivity and Recall

$$\text{False Positive Rate} = \frac{\text{False Positives}}{\text{False Positives} + \text{True Negatives}} = 0\%$$

also known as:
1-Specificity

$$\text{Positive Predictive Value} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

also known as:
Precision

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{n}$$

Aortic Valves Study

$$\text{True Positive Rate} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

also known as:
Sensitivity and Recall

$$\text{False Positive Rate} = \frac{\text{False Positives}}{\text{False Positives} + \text{True Negatives}} = 0\%$$

also known as:
1-Specificity

$$\text{Positive Predictive Value} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} = 100\%$$

also known as:
Precision

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{n}$$

Aortic Valves Study

$$\text{True Positive Rate} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

also known as:
Sensitivity and Recall

$$\text{False Positive Rate} = \frac{\text{False Positives}}{\text{False Positives} + \text{True Negatives}} = \mathbf{0\%}$$

also known as:
1-Specificity

$$\text{Positive Predictive Value} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} = \mathbf{100\%}$$

also known as:
Precision

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{n} = \mathbf{98\%}$$

Aortic Valves Study

$$\text{True Positive Rate} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

also known as:
Sensitivity and Recall

$$\text{False Positive Rate} = \frac{\text{False Positives}}{\text{False Positives} + \text{True Negatives}} = 0\%$$

also known as:
1-Specificity

$$\text{Positive Predictive Value} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} = 100\%$$

also known as:
Precision

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{n} = 98\%$$

$$\text{AUC} = 73\%$$

Aortic Valves Study

$$\text{True Positive Rate} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} = 1\%$$

also known as:
Sensitivity and Recall

$$\text{False Positive Rate} = \frac{\text{False Positives}}{\text{False Positives} + \text{True Negatives}} = 0\%$$

also known as:
1-Specificity

$$\text{Positive Predictive Value} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} = 100\%$$

also known as:
Precision

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{n} = 98\%$$

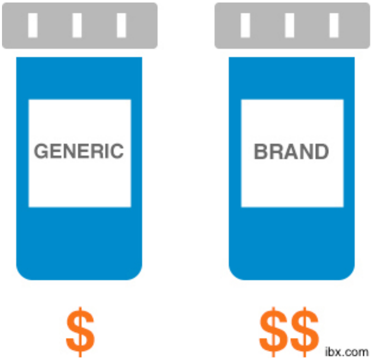
AUC = 73%

HUMANS + MACHINES

Predicting Unprofitability

Profit-Maximizing Insurer:

- Design plan to attract profitable & deter unprofitable enrollees
- Cannot discriminate based on pre-existing conditions
- Raise/lower out of pocket costs of drugs for some conditions
- Distortions make it difficult for unprofitable groups to find acceptable coverage



Demonstrate drug formulary identifies unprofitable enrollees

Computational health economics for identification of unprofitable health care enrollees

Sherri Rose, Savannah L. Bergquist, Timothy J. Layton

Predicting Unprofitability

- Limit to ∼10 non-zero variables
- Augment with therapeutic classes for HIV & multiple sclerosis drugs

```r
# lasso screener that always retains classes for HIV and MS drugs
var.index <- c(which(colnames(newdat)=="tcls14"), which(colnames(newdat)=="tcls251"))

screen.glmnet10 <- function(Y, X, family, alpha = 1, minscreen = 2, nfolds = 10, nlambda = 100,fixed.var.index=var.index,...) {
  # .SL.require('glmnet')
  if(!is.matrix(X)) {
    X <- model.matrix(~ -1 + ., X)
  }
  fitCV <- glmnet::cv.glmnet(x = X, y = Y, lambda = NULL, type.measure = 'deviance',
                             nfolds = nfolds, family = family$family, alpha = alpha,
                             nlambda = nlambda, pmax=10, parallel=T)
  whichVariable <- (as.numeric(coef(fitCV$glmnet.fit, s = fitCV$lambda.min))[-1] != 0)
  # the [-1] removes the intercept; taking the coefs from the fit w/ lambda that gives minimum cvm
  if (sum(whichVariable) < minscreen) {
    warning("fewer than minscreen variables passed the glmnet screen,
            increased lambda to allow minscreen variables")
    sumCoef <- apply(as.matrix(fitCV$glmnet.fit$beta), 2, function(x) sum((x != 0)))
    newCut <- which.max(sumCoef >= minscreen)
    whichVariable <- (as.matrix(fitCV$glmnet.fit$beta)[, newCut] != 0)
  }
  whichVariable[c(var.index)] <- TRUE
  return(whichVariable)
}
```

sl-bergquist.github.io/unprofits

IN CLOSING

Bryan Cantrill
@bcantrill

How about a conference called "In Retrospect" in which presenters revisit talks they've given years prior -- and describe how their thinking has evolved since?

7:01 PM - 28 Jun 2018

1,036 Retweets **5,714** Likes

nature International weekly journal of science

Publish houses of brick, not mansions of straw

Papers need to include fewer claims and more proof to make the scientific literature more reliable, warns William G. Kaelin Jr.

23 May 2017 NATURE | COLUMN: WORLD VIEW

"…goal of a paper seems to have shifted from validating specific conclusions to making the broadest possible assertions."

Role of Tutorials

Practice of Epidemiology

Mortality Risk Score Prediction in an Elderly Population Using Machine Learning

Sherri Rose*

Practice of Epidemiology

Targeted Maximum Likelihood Estimation for Causal Inference in Observational Studies

Megan S. Schuler and Sherri Rose*

International Journal of Epidemiology, 2019, 1–7
doi: 10.1093/ije/dyz132
Education Corner

Reflection on modern methods: when worlds collide—prediction, machine learning and causal inference

Tony Blakely,[1]* John Lynch,[2] Koen Simons,[1] Rebecca Bentley[1] and Sherri Rose[3]

Preprints, Data, and Code

Does Your Algorithm Have a Social Impact Statement?

Responsibility

Explainability

Accuracy

Auditability

Fairness

fatml.org/resources/principles-for-accountable-algorithms

1. Improvements to research infrastructure needed

2. Types of training most important for this research

3. Future research needs

1. **Improvements to research infrastructure needed**

    `Developing and maintaining software`

2. **Types of training most important for this research**

3. **Future research needs**

1. Improvements to research infrastructure needed

`Developing and maintaining software`

2. Types of training most important for this research

“Learning two fields takes, surprisingly, twice as long as learning one. But it’s worth the investment because you get to solve real problems for the first time.”

Barbara Engelhardt | Princeton

3. Future research needs

1. Improvements to research infrastructure needed

`Developing and maintaining software`

2. Types of training most important for this research

> “Learning two fields takes, surprisingly, twice as long as learning one. But it’s worth the investment because you get to solve real problems for the first time.”

Barbara Engelhardt | Princeton

3. Future research needs

Does Your Algorithm Have a Social Impact Statement?

Responsibility Explainability
Accuracy Auditability
Fairness

1. Improvements to research infrastructure needed

`Developing and maintaining software`

2. Types of training most important for this research

> "Learning two fields takes, surprisingly, twice as long as learning one. But it's worth the investment because you get to solve real problems for the first time."

Barbara Engelhardt | Princeton

3. Future research needs

Does Your Algorithm Have a Social Impact Statement?

Responsibility Explainability
Accuracy Auditability
Fairness

Machine learning for causal inference in *Biostatistics*

SHERRI ROSE
Department of Health Care Policy, Harvard Medical School
rose@hcp.med.harvard.edu
and
DIMITRIS RIZOPOULOS
Department of Biostatistics, Erasmus University Medical Center

1. Improvements to research infrastructure needed

`Developing and maintaining software`

2. Types of training most important for this research

"Learning two fields takes, surprisingly, twice as long as learning one. But it's worth the investment because you get to solve real problems for the first time."

Barbara Engelhardt | Princeton

3. Future research needs

Does Your Algorithm Have a Social Impact Statement?

Responsibility Explainability
Accuracy Auditability
Fairness

Machine learning for causal inference in *Biostatistics*

SHERRI ROSE
Department of Health Care Policy, Harvard Medical School
rose@hcp.med.harvard.edu
and
DIMITRIS RIZOPOULOS
Department of Biostatistics, Erasmus University Medical Center

Standards and guidelines adopted by the community followed by buy-in from journals and grantors

Acknowledgements


Sam Adhikari, PhD
NYU


Austin Denteh, PhD
Tulane


Savannah Bergquist, PhD
Berkeley Haas


Akritee Shrestha, MS
Wayfair


Maia Majumder, PhD
Harvard


Alex McDowell
Harvard


Anna Zink
Harvard


Toyya Pujol
Georgia Tech


Irina Degtiar
Harvard


Christoph Kurz
University of Munich

Funding:

NIH Director's New Innovator Award (DP2-MD012722)
Laura and John Arnold Foundation
NIH R01-GM111339