What is ACCORDS?

Adult and Child Center for Outcomes Research and Delivery Science

ACCORDS is a ‘one-stop shop’ for pragmatic research:

- A multi-disciplinary, collaborative research environment to catalyze innovative and impactful research
- Strong methodological cores and programs, led by national experts
- Consultations & team-building for grant proposals
- Mentorship, training & support for junior faculty
- Extensive educational offerings, both locally and nationally

ACCORDS Upcoming Events

April 3, 2024 AHSB Room 2002, Zoom	**Ethics, Challenges, & Messy Decisions in Shared Decision Making** Training Clinicians in Shared Decision Making: Lessons from SHARE *Presented by:* Chris Knoepke, PhD, MSW; Laura Scherer, PhD
April 15, 2024 AHSB 2200/2201, Zoom	**Statistical Methods for Pragmatic Research** *Presented by:* Michael Matheny, MD (Vanderbilt University Medical Center)
April 26, 2024 AHSB 2200/2201, Zoom 11am-1pm MT	**ACCORDS/CCTSI Community Engagement Showcase**
May 20, 2024 AHSB 2200/2201, Zoom	**Statistical Methods for Pragmatic Research** Planning a Pragmatic Effectiveness Trial with a Factorial Design by Targeting the Posterior Distribution Variance *Presented by:* Keith Goldfeld, DrPH, MS, MPA/MURP
Last seminars for the 2023-2024 academic year!	

*all times 12-1pm MT unless otherwise noted

COPRH Con

Colorado Pragmatic Research in Health Conference

Innovations in Pragmatic Research Methods

From Data to Equity, Policy, and Sustainability

June 5 - 6, 2024 | 10am-3:30pm MT

Registration is open now at

www.COPRHCon.com

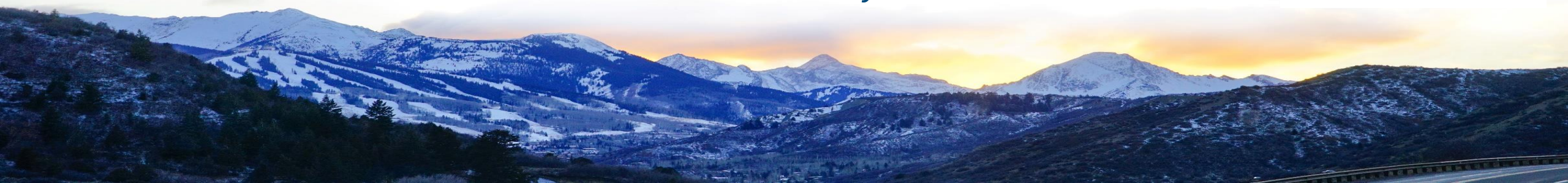
ACCORDS

ADULT AND CHILD CENTER FOR OUTCOMES RESEARCH AND DELIVERY SCIENCE

UNIVERSITY OF COLORADO
CHILDREN'S HOSPITAL COLORADO

Early Bird Registration Ends 3/31

Registration Fees waived for students, staff, and faculty of CU SOM or CHCO



Statistical Methods for Pragmatic Research Seminar Series

2023-2024 seminar series

Ryan Peterson, PhD

Kathryn Colborn, PhD

Pragmatic Statistical Learning:

From Data to Interpretable Insights

colorado school of public health

Pragmatic Statistical Learning: From Data to Interpretable Insights

RYAN PETERSON, ASSISTANT PROFESSOR, BIOSTATISTICS & INFORMATICS

KATIE COLBORN, ASSOCIATE PROFESSOR, MEDICINE

ACCORDS STATISTICAL METHODS IN PRAGMATIC RESEARCH SERIES

MARCH 11, 2024

Outline

The perils of real data

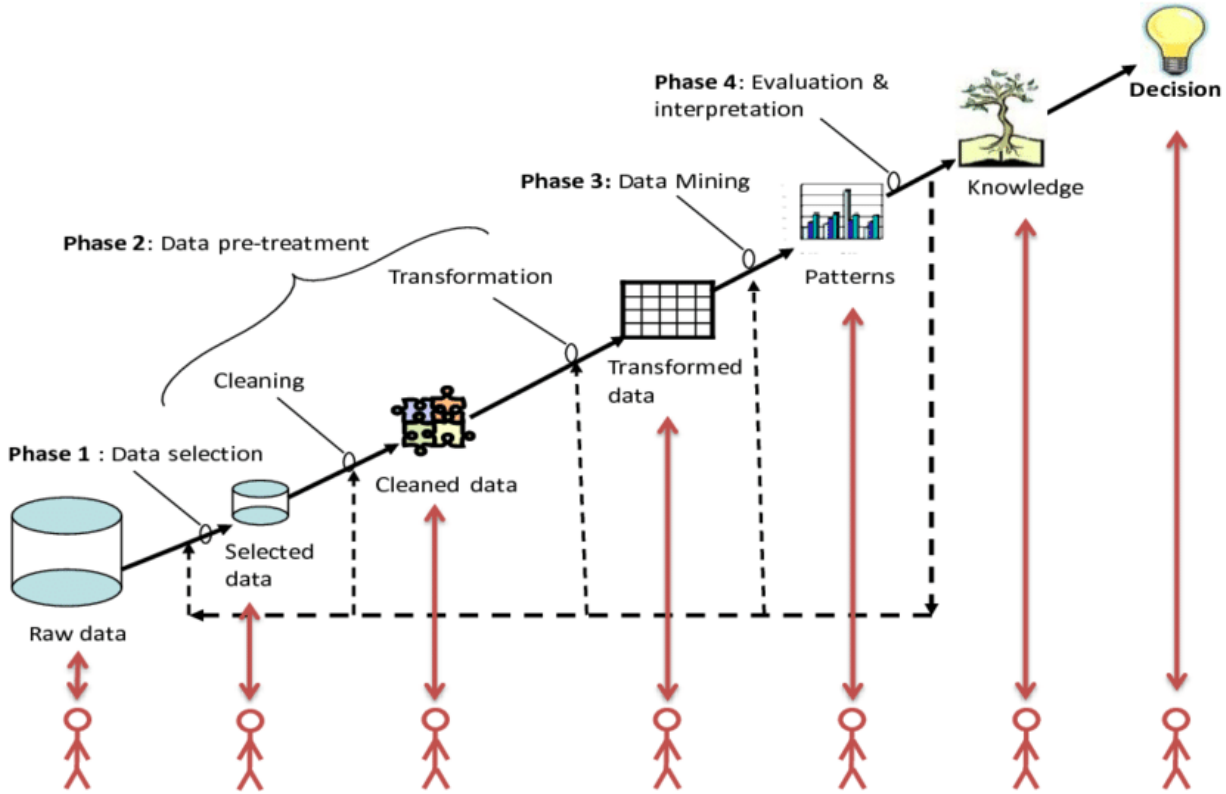
Considerations of machine learning in pragmatic research

A survey of machine learning methods

Case studies

On interpretability

Penalized regression – a pragmatic choice



Classroom data vs real data

CLASSROOM DATA SETS

Showcase how statistical methods are supposed to work

Small

Complete (or filtered to complete cases)

Well-documented

One row per independent observation

REAL DATA SETS

Extremely large (big data)

Frequently missing in complex ways

Complex structure

Not (easily) generalizable

Considerations for BIG data sets

Often cannot be processed on a desktop

Often utilized to solve causal questions, but may provide limited causal insights

Traditional methods used for big data tend to fall victim to bias

Big data might be…

- Extremely *tall,* with many observations of a relatively small set of features
- Extremely *wide,* with many features on a relatively small set of observations
- High throughput, functional, time series
- Some combination of all of these

High-sample or high-dimensional?

"High sample size" big data refer to "tall" data sets with an extremely large number of samples (n).

- Common if data collection is cheap relative to value proposition (e.g. advertising, insurance)
- Less common for randomized controlled clinical trials, the gold-standard causal study.

"High-dimensional" data refer to "wide" data sets with an extremely large number of *features* (p).

- Common if measuring many features of a single independent observation is cheap
- 'Omics data are typically high-dimensional, rarely have a n>>p

Additional complexity when big data have both high-n and high-dimensional components (wearables, Google trends, images)

p

n

ID	X1	…	Xp
S1			
…			
Sn			

High-sample or high-dimensional?

HIGH SAMPLE SIZE (N>P)

Produces “stable” insights

Bias of primary concern

Diminishing returns

Black-box algorithms tend to perform well

HIGH DIMENSIONS (P>N)

Tends to produce “unstable” insights

Multiplicity of primary concern

Causal effects difficult to disentangle from correlated features; “Rashomon Effect”

What about when you have both a very high sample size and very high dimensions?

- Not always clear; depends on goals/context.
- An interdisciplinary approach is best; important to include clinician, informatician, data scientist, (bio)statistician, and more. Same goes for complex data structures.

Considerations of machine learning in pragmatic research

Machine learning

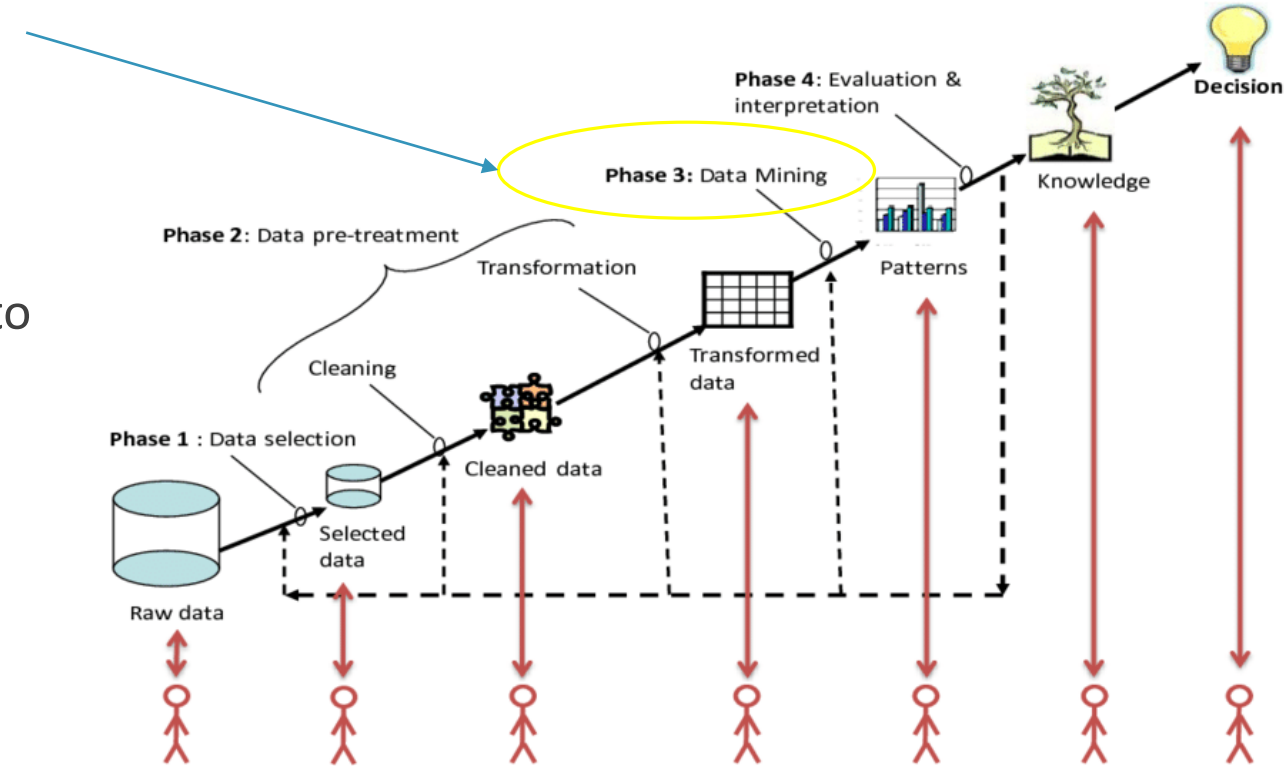
Also called “Data mining”, represents “3rd phase” of knowledge discovery

Supervised learning: using information to predict an outcome

Unsupervised learning: using information to identify subgroups

Other learning types:

- Recommendation systems
- Anomaly detections



Supervised learning

Goal: Given a training data set with covariate X's and outcome Y's, build a model to predict Y given X

Take a new X that was not in the data set and predict Y for that new observation

- Repeating this process is called "validation"

Simple supervised learning techniques can be performed that are *transparent* and *interpretable*:

- Regression!
- Penalized regression (if high-dimensional)
- Interpretable decision trees
- Allow inference on how Y changes with, or is impacted by, X

"Fancier" ML tools: random forests, ensembles/Super learners, K-nearest neighbors, support vector/kernel machines, [deep] neural networks

- Predictions and models are "explainable," but not necessarily interpretable
- Commonly referred to as "black box"

Unsupervised Machine Learning

No outcome of interest

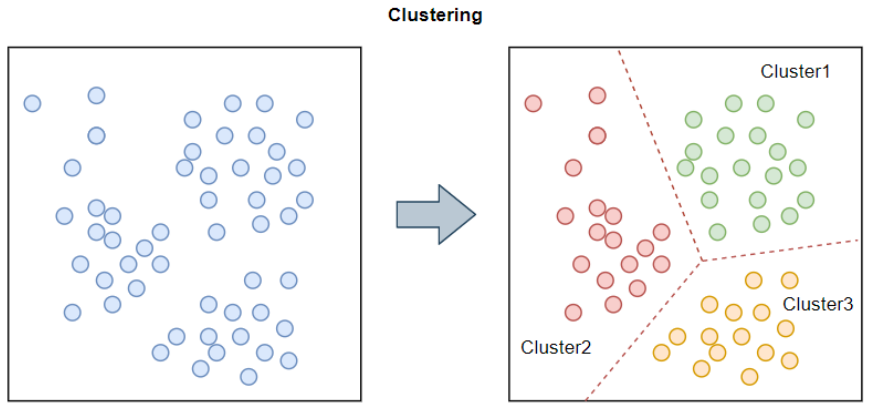
Dimension reduction (going from a high number of variables to a low number of variables)

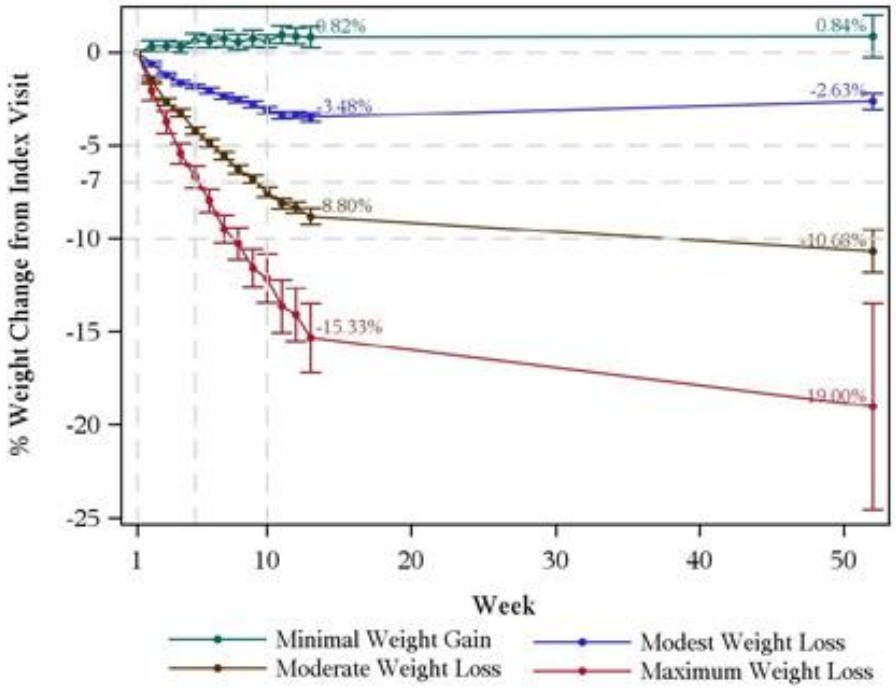
- Without the consideration of the prediction of an outcome

Clustering data points together (top-right)

Example: Weight loss trajectory clustering (bottom plot)

Most pragmatic research ends up a hybrid of unsupervised and supervised methods!





Example: PheCodes

Phecode Map 1.2 with ICD-10cm Codes (beta)

Preprint Paper on bioRxiv

View R markdown code for using this ICD-10cm map in an example PheWAS

Clear Filters | Export All | Export Visible

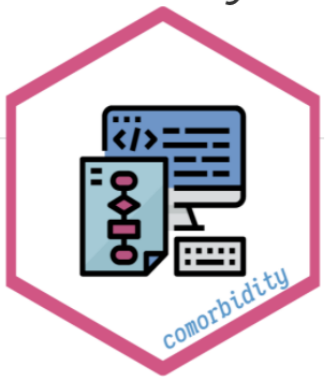
ICD10CM	ICD10CM String	PheCode	Phenotype	Excl. Phecodes	Excl. Phenotypes
icd10	description	080	phenotype	excl. phecode	excl. range name
J95.02	Infection of tracheostomy …	080	Postoperative infection	080-082.99	infectious diseases
K68.11	Postprocedural retroperito…	080	Postoperative infection	080-082.99	infectious diseases
K94.22	Gastrostomy infection	080	Postoperative infection	080-082.99	infectious diseases
K94.32	Esophagostomy infection	080	Postoperative infection	080-082.99	infectious diseases
N98.0	Infection associated with a…	080	Postoperative infection	080-082.99	infectious diseases
T81.4	Infection following a proce…	080	Postoperative infection	080-082.99	infectious diseases
T81.4…	Infection following a proce…	080	Postoperative infection	080-082.99	infectious diseases
T81.4…	Infection following a proce…	080	Postoperative infection	080-082.99	infectious diseases
T81.4…	Infection following a proce…	080	Postoperative infection	080-082.99	infectious diseases

Example: Comorbidity scores

- Charlson score
- Elixhauser comorbidity score

The {comorbidity} Package: Computing Comorbidity Scores

README last updated: 2021-10-19

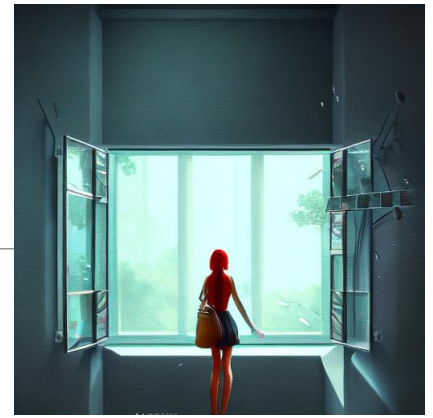


`comorbidity` is an R package for computing comorbidity scores such as the weighted Charlson score and the Elixhauser comorbidity score; both ICD-10 and ICD-9 coding systems are supported.

ML in pragmatic research – what is needed?

PRAGMATIC DATA COMMON ISSUES	DESIRED PROPERTY FOR PRAGMATIC ML
Extremely large (big data)	Scalable (in both p and n)
Missing data	Expectant & respectful of missing data
Complex structure	Respects data structure & dependence
Not (easily) generalizable	Human-interpretable conclusions
Diverse stakeholders; ethical considerations	Transparency & relevance
Noisy (low signal)	Works in low-signal settings
More?	More?

Hotpot.ai: “privacy”



Ethical issues - Privacy

Is privacy of individuals maintained with big data?

The trade-off between privacy and variety of data

- Is it possible to de-identify high-dimensional data?
- Is it possible to de-identify location data?

} Without losing valuable information?

Who can access the stored big data?

Further reading:

- https://www.theperspective.com/debates/businessandtechnology/the-perspective-on-big-data/
- https://hbr.org/2012/08/dont-build-a-database-of-ruin
- https://journalofbigdata.springeropen.com/articles/10.1186/s40537-016-0059-y
- Risk mitigation for synthetic data sets: https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2017/3_risk_mitigation.pdf

Additional ethical concerns

Selection and Sampling Bias

- Researchers may very precisely answer the wrong question
- Biased data sets will have limited generalizability, even with massive sample size
- Example: 2013 Boston pothole detection
  - Used smartphone app to detect potholes using GPS and accelerometers
  - People in lower-income groups less likely to have smartphone
  - Which potholes will be patched?
- Algorithmic bias

Transparency: How can we trust in (black-box) ML when we don't understand how they work?

Who are the stakeholders in pragmatic research? Who is accountable for ML systems (ultimately)?

What ML method for pragmatic research?

Desired properties	Neural Networks	Random Forests	Regression	Penalized regression
Scalable (in both p and n)	✓	✓	✗	✓
Expectant & respectful of missing data*	✓	✓	✓ ✓	✓ ✓
Respects data structure & dependence	✓	✓	✓ ✓	✓
Human-interpretable conclusions	✗	✗	✓	✓
Transparency & relevance	✗	✗	✓	✓
Noisy (low-signal)	✗	✗	✓	✓ ✓

Penalized regression is a great prospect for ML in pragmatic research

Machine learning – additional resources

Highly recommend the following textbook if you want to learn more:

- Applied Predictive Modeling by Max Kuhn and Kjell Johnson

Others:

- An Introduction to Statistical Learning by James, Witten, Hastie, and Tibshirani
- The Elements of Statistical Learning by Freidman, Tibshirani, and Hastie
- TRIPOD Statement(s)

Other good readings:

- Leo Breiman. "Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author).” Statistical Science, 16(3) 199-231 August 2001. https://doi.org/10.1214/ss/1009213726
- https://www.fharrell.com/post/stat-ml/

Case studies: Administrative data

Administrative data

Lots of data are collected routinely by organizations: financial data, claims, billing, in addition to EHR data.

Such “administrative” data are not typically collected for the purposes of research, but still can be utilized to answer interesting questions related to research.

Allows for extremely large sample sizes (high-n), typically easy and/or cheap to acquire

Tend to be well-documented, facilitating data management (still hard given massiveness of data).

Important to consider many possible sources of bias!

Examples:

- Medicare/Medicaid data
- Premier data
- Census data
- National Health and Nutrition Examination Survey (NHANES)
- Healthcare Cost and Utilization Project (HCUP) data contains rich information on national and state-level healthcare utilization, access, quality, charges, and outcomes

SSI seasonality in the NRD

Surgical site infections (SSIs) after total knee (TK) and total hip arthroplasty (THA) are devastating to patients and costly to healthcare systems.

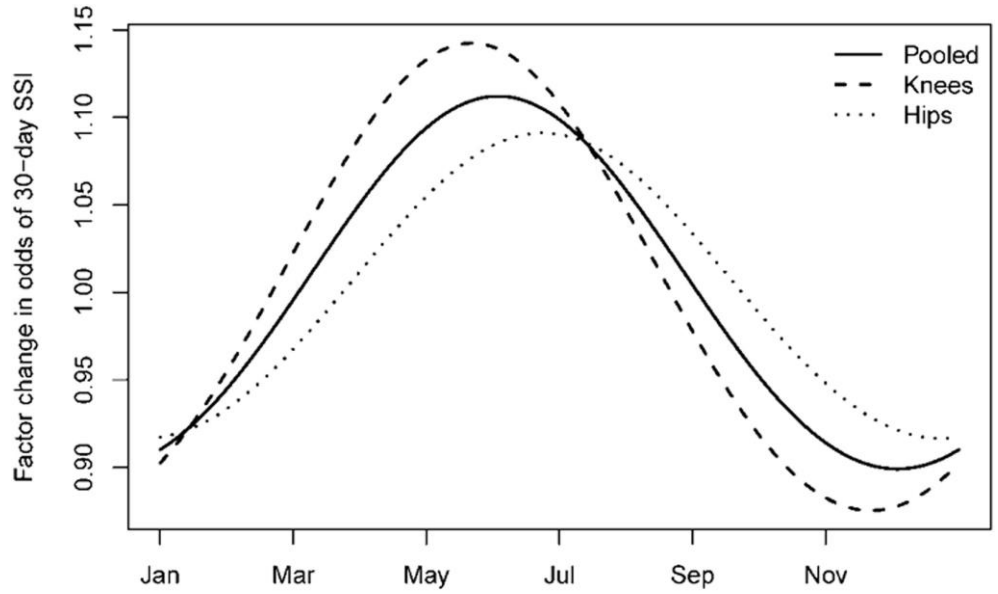
Are SSIs seasonal?

Let's find out with HCUP's National Readmission Database.

>760K procedures over 2 years

Adjusted SSI incidence 24% higher in June vs December

	Nadir Peak, Odds Ratio (95% CI)	Estimated Nadir Month	Estimated Peak Month
TKA	1.305 (1.201-1.418)	December	June
THA	1.19 (1.091-1.298)	January	July
Pooled	1.237 (1.164-1.314)	December	June



Anthony, Peterson et al. 2018

Cellulitis seasonality in the NIS

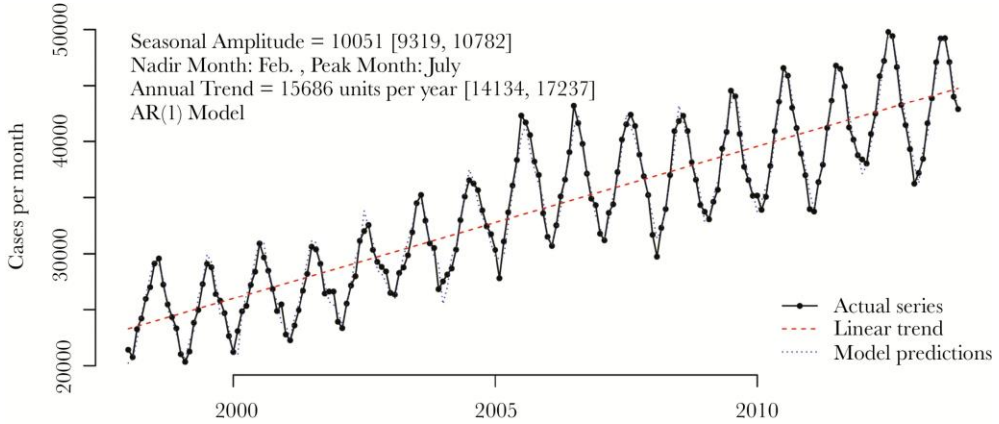
HCUP's National Inpatient Sample is bigger than the NRD (more years available), and is meant to be nationally representative.

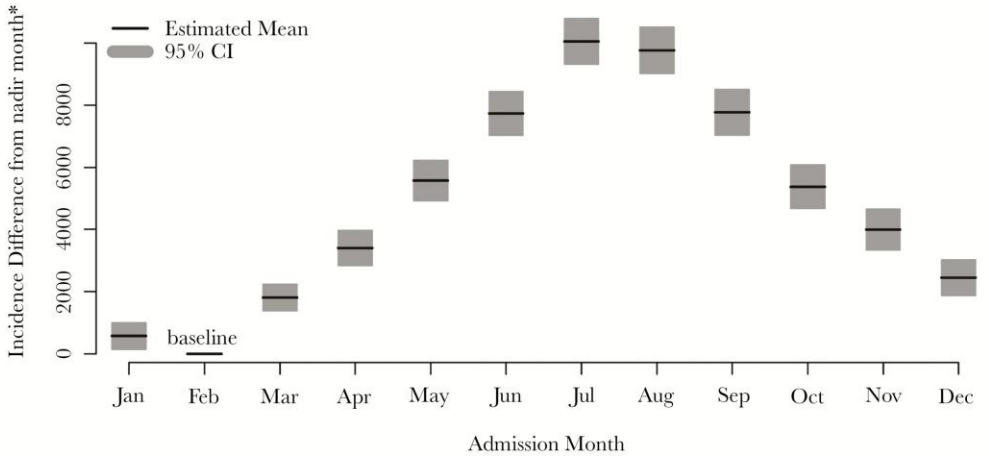
We used data from 1998-2013 to show that hospitalizations for cellulitis, a skin infection, have increased dramatically in incidence and associated costs.

We also showed that incidence was seasonal, which had previously been unknown.

Methods: extract counts of cases by month, then model as a time series

- Efficient for > 5 million records





Peterson et al, 2017

Electronic Health Records (EHR)

The EHR consists of clinic notes, medications, labs, demographics, registries, vitals, billing codes, outpatient encounters, and more.

Databases live as either structured (e.g. ICD-codes), semi-structured (e.g. medications/dosage), or unstructured text (clinical notes).

Our campus has an EMR repository and many resources available for working with this data:

- Health Data Compass: www.healthdatacompass.org
- Colorado Center for Personalized Medicine Biobank: www.cobiobank.org
- CIDA: coloradosph.cuanschutz.edu/research-and-practice/centers-programs/cida
- D2V: https://medschool.cuanschutz.edu/accords/about/d2v

EHR data example – Simon et al. 2021

"Harmonized" EHR data can be used to build machine learning models that can be directly used during clinical care.

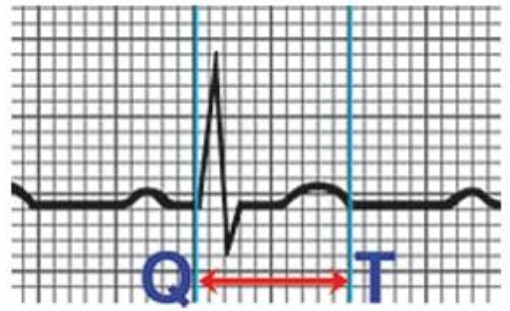
N ≈ 35K patients at risk for drug-induced QT prolongation using the UCHealth EHR, with p ≈ 6500 features derived from the harmonized EHR data

- Included medications, procedure codes, diagnosis codes, labs, and demographic data

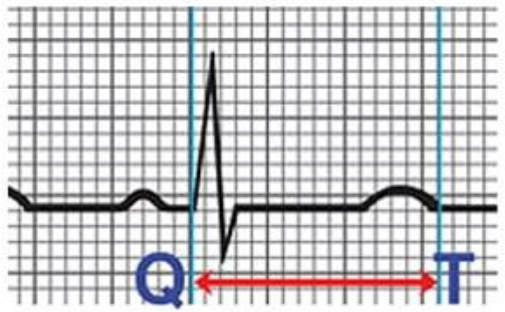
Analyses utilized 96 CPUs and 620 GB of RAM (Google Cloud platform).

Deep neural networks were found to best predict the outcome, achieving an out-of-sample AUC of 0.71 (FPR = 28%, FNR = 30%).

How to interpret??



Normal Q-T Interval



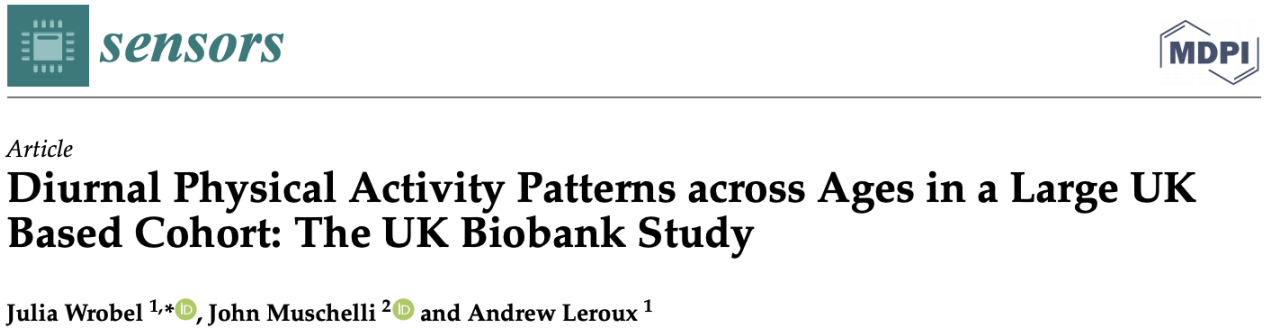
Long Q-T Interval

Source: Cleveland Clinic

Registries as new frontier in pragmatic research

Increasingly, patient registries and large cohort studies are being linked with other types of data:
SEER-Medicare: a cancer registry (35% of all cancer patients) linked to Medicare data

- All of Us: a disease agnostic cohort study gathering data from wearables (heart rate, physical activity, sleep, and more) as well as social determinants of health, family health history, health care access, etc.
- PROGRESS: combines wearable data with electronic health records, biosamples, and patient generated data to identify what influences how your body responds to food.
- NCDB: 70% newly diagnosed cancer patients linked with >34 million historical records
- The UK Biobank study (>80K subjects with granular wearable data, genetic data, and outcomes)



On interpretability

Interpretability in Machine Learning

SIMPLE/INTERPRETABLE

Tend to have parameters which can be interpreted in sentences/words

The process by which predictions are made is mathematically tractable and understandable.

Insights can be confirmed and generalized to new populations.

Examples:

- the older a child, the larger our prediction for their FEV. (linear regression)
- Patients in 3rd class (or crew) on the titanic had the worst odds of survival (logistic regression)

"BLACK BOX" MODELS

If they have parameters, they cannot be well-understood (think: high order interactions)

The process by which predictions are made need not be mathematically tractable (by humans).

Input -> ?? -> predictions. The question marks are "left blank" by the researchers, who only care about the predictions themselves.

Insights only apply to those represented in training data.

Interpretability of specific relationships

How does does X impact Y? Or, how does Y change when X changes?

- In regression, “holding confounders constant” allows us to get closer to a causal interpretation of X -> Y.

Some variables should not be considered “predictors”:

- Mediators
- Not available at time of prediction
- Frequently missing (more soon)
- Near zero variance
- Highly collinear with a better predictor
- Irrelevant
- More?

Explainability tools help uncover variable “importance” and even can describe functional forms for black-box models, but (penalized) regression models are simpler.

Interpretability of *inferences*

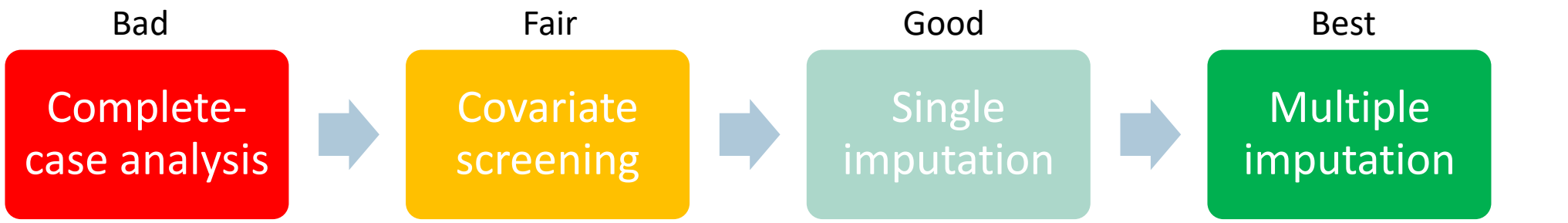
In statistics, we not only ask "does X impact Y", we ask whether X impacts Y *significantly.*

- Is the relationship strong enough that we cannot attribute it to chance?

We traditionally use p-values and confidence intervals for inferences, which are generally well-understood and interpretable (though they have their issues). Unfortunately, they are only available for regression!

Desired properties	Neural Networks	Random Forests	Regression	Penalized regression
Scalable (in both p and n)	✓	✓	✗	✓
Expectant & respectful of missing data*	✓	✓	✓ ✓	✓ ✓
Respects data structure & dependence	✓	✓	✓ ✓ *	✓
Human-interpretable conclusions	✗	✗	✓	✓
Transparency & relevance	✗	✗	✓	✓
Noisy (low-signal)	✗	✗	✓	✓ ✓
Valid p-values & CIs for variables	✗	✗	✓	✗ ✓

Respectfulness to missing data



-throws away lots of data
-assumes MCAR

-throws away less data
-assumes MCAR

-throws away no data
- Overly precise
- Assumes MAR

-throws away no data
- Respects uncertainty in missing data
- Assumes MAR

Multiple imputation with penalized regression

Obtain M data sets with missing data stochastically imputed

On each data set, perform penalized regression method of choice and store fitted model coefficients

Take the median coefficient estimates across M data sets

- Each data set “votes” on whether a coefficient is positive, negative, or zero
- If a majority say positive, the median coefficient will be positive (vice versa)

Final model is parameterized by median coefficients

As described in Peterson (2021).

Making interpretable inferences after penalized regression

Some options for post-selection inference after penalized regression:

- Knockoff filters
- Stability selection
- Bootstrapping (sometimes)
- Marginal false discovery rates
- Selective inference

When candidate predictors are highly correlated with each other, modern tools become unstable and imprecise. Considering such features as *substitutable* can help.

- Example: BMI and waist circumference are both potential predictors.
- Irrelevant question: Is BMI important conditional on waist circumference?
- Relevant question: Is either BMI or waist circumference important?

Case study: ASPIN

AUTOMATED **S**URVEILLANCE OF **P**OSTOPERATIVE **IN**FECTIONS

Automated Surveillance of Postoperative Infections (ASPIN)

- 5-year R01 funded by AHRQ (multiple PI: Colborn and Meguid 1R01HS027417)
- Development and validation of models for preoperative risk estimation and detection of postoperative infections
- Goal: reduce postoperative infections through audit and feedback to surgeons

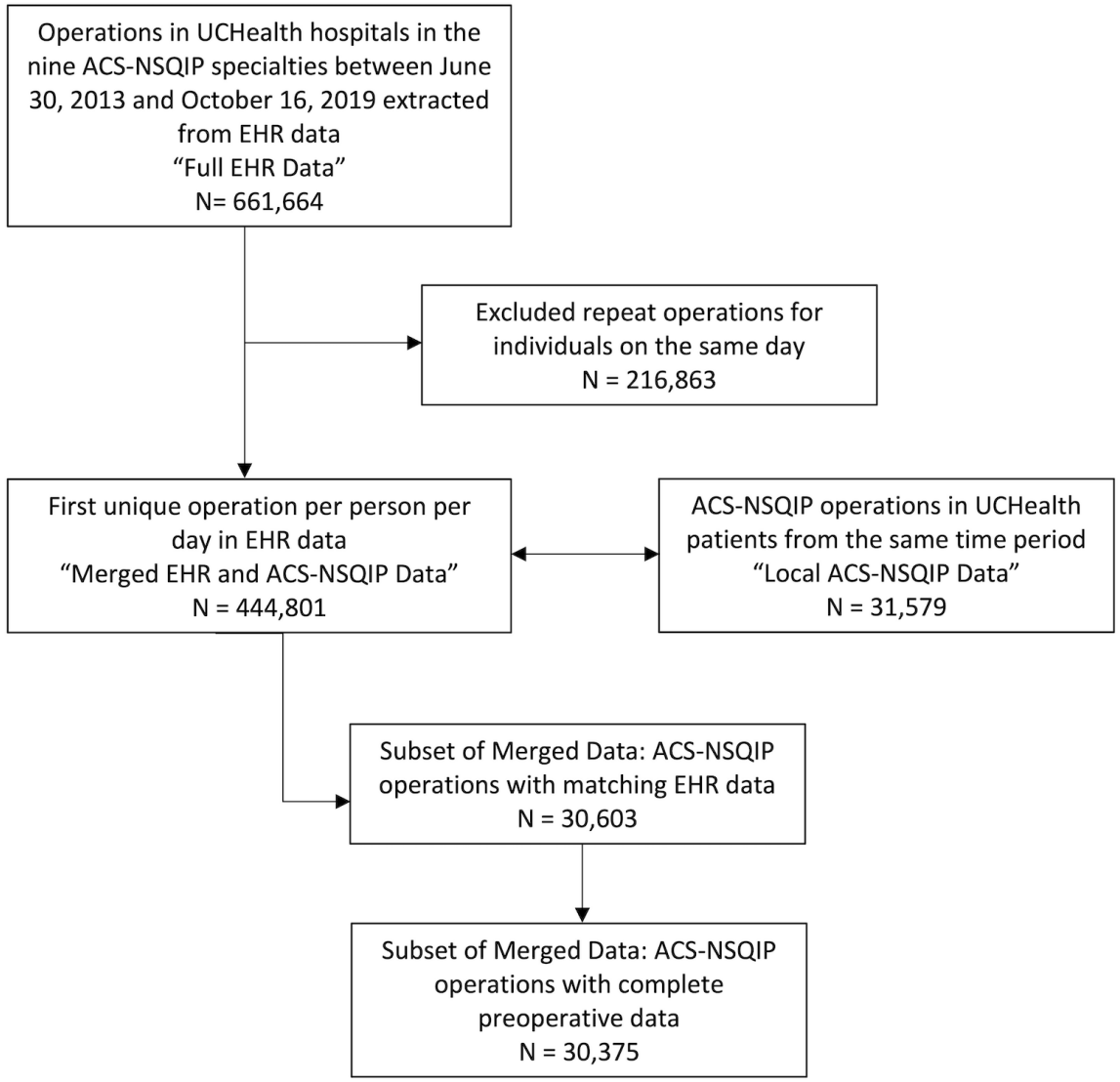
Population and Data

Obtain all operation data from UCHealth hospitals between 2013-2019

Combine ACS NSQIP registry and EHR data

EHR variables 365 days prior to operation up to 30 days after:

- ICD-9 and ICD-10 diagnosis codes
- Current Procedural Terminology (CPT) codes
- Procedure names
- Medication therapeutic and pharmaceutical classes
- Patient and operative characteristics



Preoperative and Postoperative Models

Preoperative models:

- *Objective:* Estimate the preoperative probability of each infectious complication for each patient
- *Data:* ACS NSQIP outcomes data & EHR structured predictor variables 365 days prior to the operation up to one day prior to the operation

Postoperative models:

- *Objective:* Estimate the postoperative probability of each infectious complication for each patient (i.e., likelihood the patient experienced the event, given the data).
- *Data:* ACS NSQIP outcomes data & EHR structured & unstructured predictor variables 0-30 days postoperative (2-30 days for medications)

ACS NSQIP outcomes data

Epic data

Patient ID	NSQIP SSI
1	1
2	0
3	0

Patient ID	CPT
1	87070
1	87077
1	83605
2	80053
2	96365
3	49424
3	80053

Patient ID	ICD10
1	T81.4XXA
1	F41.9
1	K21.9
2	G89.18
2	K21.9
3	R50.9
3	E87.1

Patient ID	Medication
1	ACETAMINOPHEN
1	OXYCODONE
1	AMOXICILLIN
2	ACETAMINOPHEN
2	HEPARIN
3	SODIUM CHLORIDE
3	OXYCODONE

Patient ID	NSQIP SSI	080	87070	80053	AMOXICILLIN
1	1	1	1	0	1
2	0	0	0	1	0
3	0	0	0	1	0

~30K x 15K dim

Model Development

Divide data into training (70%) and test (30%) sets, split temporally

Apply lasso plus the knockoff filter for controlled variable selection

- Generate the knockoff variables without looking at the outcomes
- Estimate penalized regression coefficients (lasso)
- Use false discovery rate (FDR) correction to select variables whose coefficients are sufficiently larger than their knockoff's

Further filter variable selection with input from surgeons

Fit logistic regression in training set using selected features and apply to test set to estimate performance

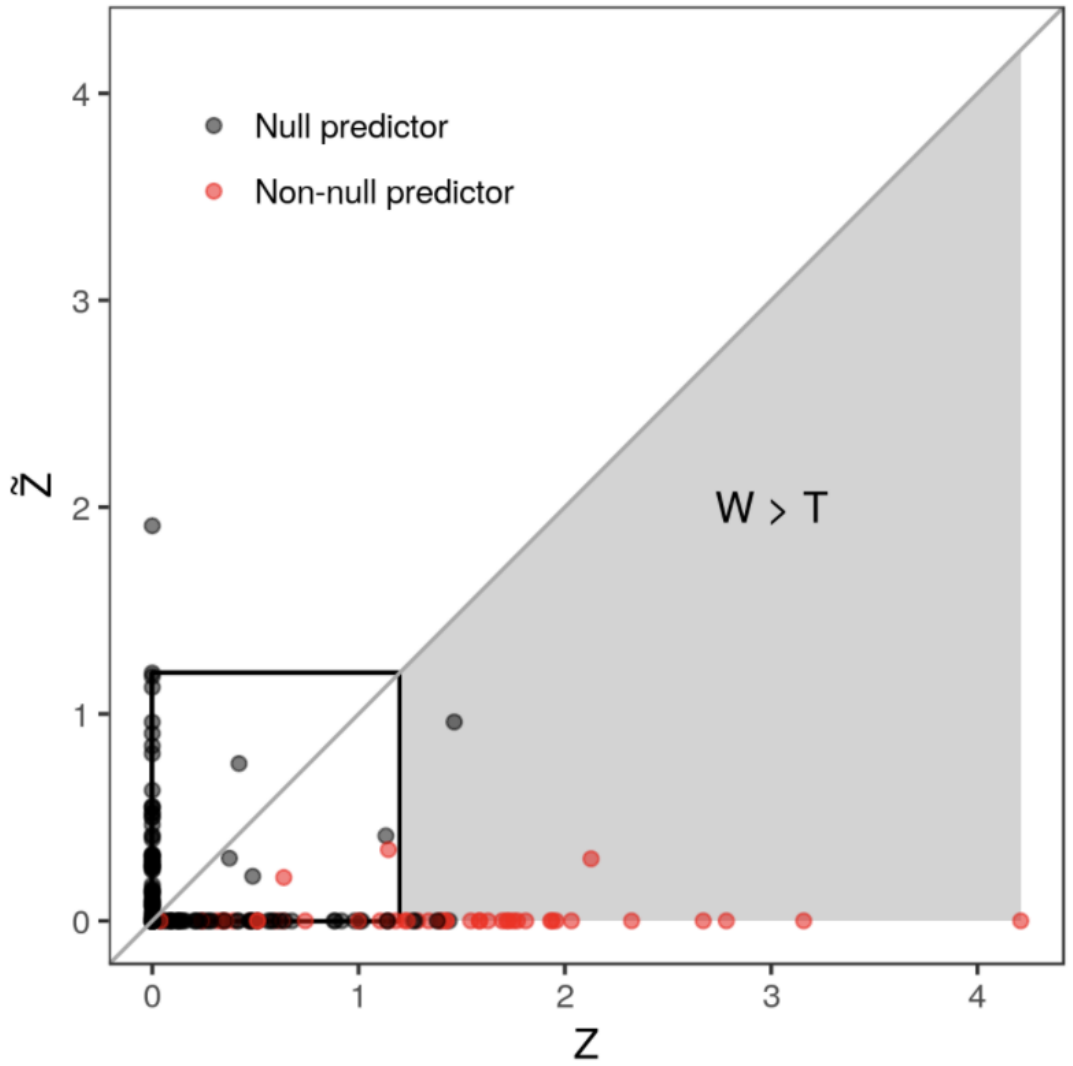


Image source:
https://web.stanford.edu/group/candes/knockoffs/outline.html

Results

	Training (N=21,450)	Testing (N=9,189)
Age (mean [SD])	55.0 (16.5)	54.2 (17.0)
Female	12,055 (56.2%)	5,134 (55.9%)
Surgeon specialty		
Orthopedic Surgery	7,247 (33.8%)	2,950 (32.1%)
General Surgery	5,835 (27.2%)	2,569 (28.0%)
Gynecology	2,455 (11.4%)	1,095 (11.9%)
Urology	1,628 (7.6%)	728 (7.9%)
Neurosurgery	1,579 (7.4%)	711 (7.7%)
Postoperative infections		
Overall infections	1,551 (7.2%)	567 (6.2%)
Surgical Site Infections	784 (3.7%)	285 (3.1%)
Urinary tract infection	322 (1.5%)	144 (1.6%)
Sepsis/septic shock	584 (2.7%)	184 (2.0%)
Pneumonia	222 (1.0%)	63 (0.7%)

	Preoperative Models		Postoperative Models	
Infection type	No. variables	AUC Test Set	No. variables	AUC Test Set
SSI	7	0.73	4	0.91
UTI	6	0.76	7	0.93
Sepsis/septic shock	6	0.89	7	0.95
Pneumonia	6	0.84	8	0.96

Preoperative Models: Variables

SSI variables:

- wound class, outpatient, ortho, urology, PheCodes 080: "Postoperative infection" & 560: “Other intestinal obstruction”, blood culture, outside hospital CT scan

UTI variables:

- sex, outpatient, gynecology, urology, PheCode 591: “Urinary tract infection”, outside hospital CT scan

Sepsis variables:

- wound class, ASA, outpatient, ortho, PheCodes 567: ”Peritonitis/retroperitoneal infections” & 994: “Sepsis and SIRS”

Pneumonia variables:

- comorbidities, ASA, outpatient, PheCodes: 150: “Cancer of esophagus” & 480: “Pneumonia” & 501: “Pneumonitis due to inhalation of food or vomitus”

Postoperative Models: Variables

SSI variables:

- PheCodes 080: "Postoperative infection" & 1011: “Complications of surgical proc.”, blood culture, >=1 antibiotic

UTI variables:

- PheCode 591: “Urinary tract infection” & 590: “Pyelonephritis” & 592.X: “Cystitis/Urethritis” & 599.X: “Symptoms of urinary system”, >=1 antibiotic, urine culture, C.difficile PCR

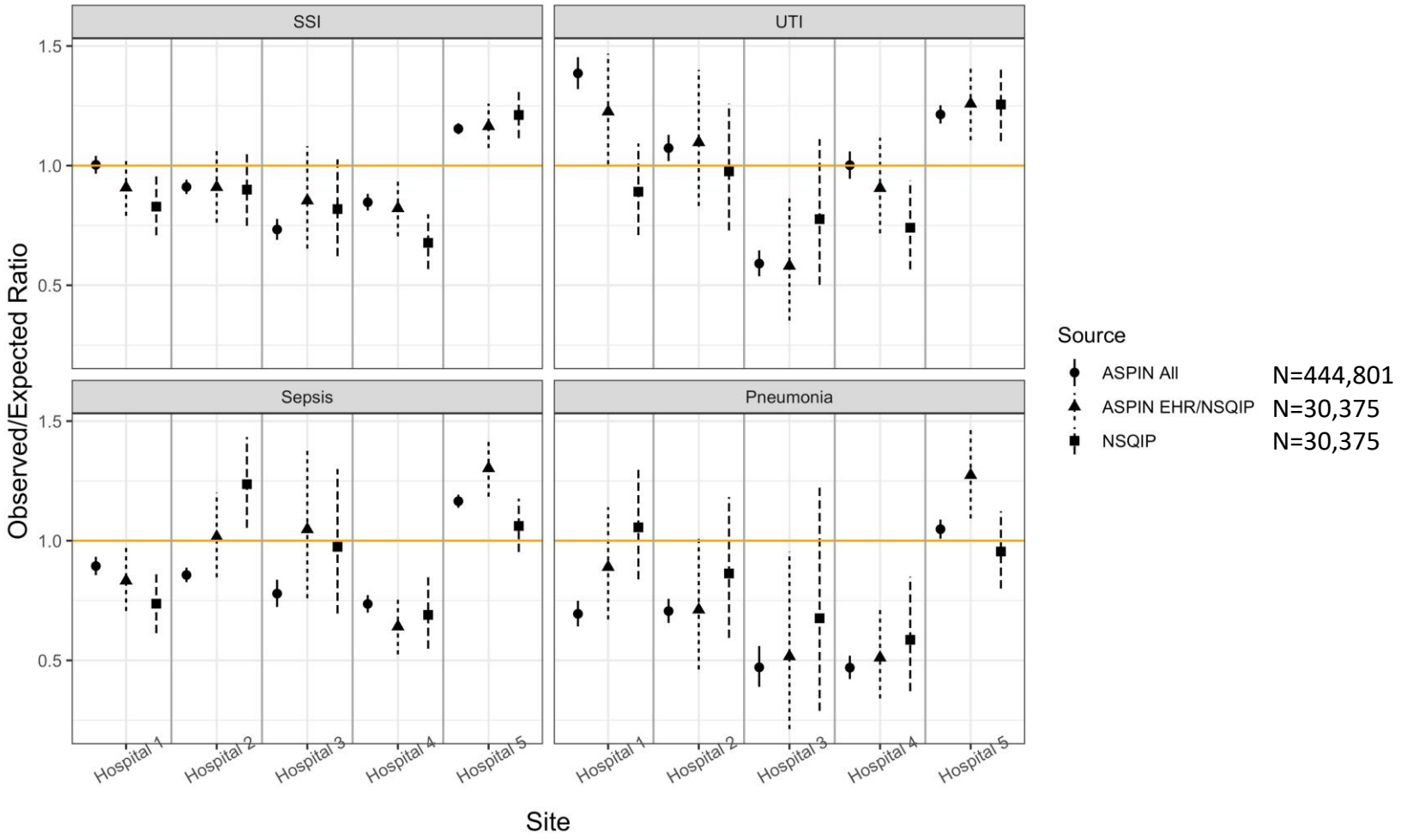
Sepsis variables:

- PheCodes 540.X: ”Appendicitis” & 994: “Sepsis and SIRS”, >=1 antibiotic, CBC auto diff, blood culture, magnesium serum, peripheral blood smear

Pneumonia variables:

- PheCodes 480.X: “Bacterial/viral pneumonias & 501: “Pneumonitis due to inhalation of food or vomitus” & 1013: ”Asphyxia and hypoxemia”, >=1 antibiotic, magnesium serum, vancomycin trough, respiratory culture, blood gasses

OE Ratio Comparisons by Hospital



Thank you!

QUESTIONS?

Transparent vs black box learning

AN INTERPRETABLE STATISTICAL MODEL MAY BE THE BETTER CHOICE IF…

Uncertainty is inherent in the outcome

The signal-to-noise ratio is low

One wants to isolate effects of a small number of variables

Uncertainty in an overall prediction or the effect of a predictor is sought

Additivity is the dominant way that predictors affect the outcome (interactions are relatively rare and can be pre-specified)

The sample size isn't huge

One wants to isolate the effects of "special" variables such as treatment or a risk factor

One wants the entire model to be interpretable and ethically tractable

A BLACK-BOX ML MODEL MAY BE THE BETTER CHOICE IF…

The outcome doesn't have a strong component of randomness

The signal-to-noise ratio is large

Overall prediction is the goal, without being able to succinctly describe the impact of any one variable (e.g., treatment)

One is not very interested in estimating uncertainty in forecasts or in effects of selected predictors

Non-additivity is expected to be strong and can't be isolated to a few pre-specified variables

The sample size is huge

One does not care that the model is a "black box"

Source: F. Harrell Regression Blog