Case Studies in Generating Real World Evidence From Real World Data

Yong Chen, PhD
Rebecca Hubbard, PhD

DEPARTMENT *of*
BIOSTATISTICS
EPIDEMIOLOGY &
INFORMATICS

Perelman
School of Medicine
University *of* Pennsylvania

Course Materials

All course materials can be downloaded from **https://rhubb.github.io/ENAR_short_course/**

This includes:

- Slides
- Reference list
- Tutorials
- Data sets used in tutorials
- R code for tutorial solutions

Schedule

1:00 - 1:15	Introduction
1:15 - 2:00	Case Study 1: EHR data quality
2:00 - 2:30	Tutorial 1
2:30 - 3:15	Case Study 2: Combining RCTs and RWD
3:15 - 3:45	Tutorial 2
3:45 - 4:30	Case Study 3: Distributed analysis
4:30 - 5:00	Tutorial 3

Outline

EHR and the Real World

- EHR are one of the first “Real World Data” sources statisticians have gotten their hands dirty with
- The size of these data sets suggests enormous potential for learning about health and healthcare in real world settings.
  - Worldwide digital healthcare data is expected to reach 25 exabytes (10^{18} bytes) in 2020.
- But... with big data comes big responsibility
- That is, more data, more problems

Electronic Health Records

Definition

"An *Electronic Health Record (EHR)* is an electronic version of a patient's medical history, that is maintained by the provider over time, and may include all of the **key administrative clinical data** relevant to that persons care under a particular provider, including **demographics, progress notes, problems, medications, vital signs, past medical history, immunizations, laboratory data and radiology reports**."
– Centers for Medicare and Medicaid Services

EHR Uptake

- Increased clinical use of EHRs has been driven largely by the Medicare and Medicaid EHR Incentive Program
- Under this program health care providers receiving reimbursement from Medicare are incentivized to adopt EHRs
- The objective of this program was to
  - Improve quality, safety and efficiency of health care and reduce health disparities
  - Engage patients and families in care
  - Improve care coordination
  - Improve population and public health
  - Ensure privacy and security of personal health information
- Regardless of whether these goals have been met or not, the practical implication for researchers is that large amounts of observational medical data are now available.

Objectives

1. Introduce key concepts in the analysis of EHR data
2. Present three example papers (case-studies) to illustrate methods, strengths and limitations
3. Provide hands-on examples using R to implement approaches described in the case studies

Outline

What are the advantages of using EHR data for research?

- No need for patient recruitment or data collection
- Large sample size
- Diverse population
- Generalizability
- Potential for multi-site studies
- Cheap and “easy” access

What are the disadvantages of using EHR data for research?

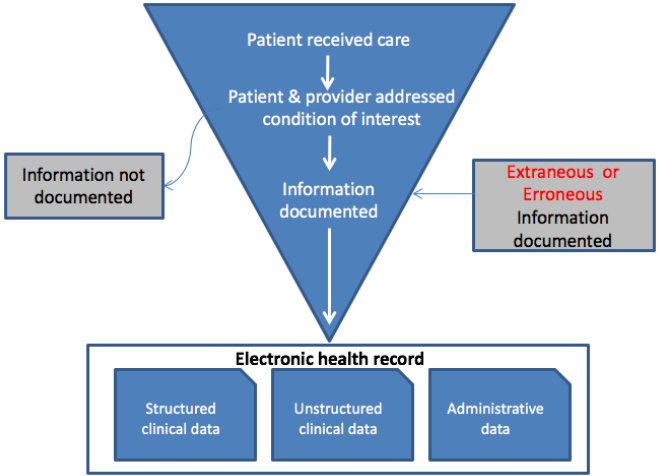
- Data quality may be poor (quantity vs quality tradeoff)
- Data collection is not systematic leading to complex missing data patterns
- Extracting data from text notes is challenging and error-prone
- Privacy protections (HIPAA) limit what data can be accessed and by whom

Dimensions of data quality

- **Completeness**: Is a truth about a patient present in the EHR?
- **Correctness**: Is an element that is present in the EHR true?
- **Concordance**: Is there agreement between elements in the EHR, or between the EHR and another data source?
- **Plausibility**: Does an element in the EHR makes sense in light of other knowledge about what that element is measuring?
- **Currency**: Is an element in the EHR a relevant representation of the patient state at a given point in time?

Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association*. 2013;20(1):144-51.

EHR data provenance



Structured data

- Many types of EHR data are available in *structured* form
- Structured data are standardized, pre-defined, computer readable data elements coded using a closed vocabulary
- For instance, procedure codes provide information on the specific health care procedures a patient has undergone
- Structured data are particularly useful for research because they can be readily manipulated and analyzed using statistical software
- They can also be combined across multiple healthcare systems using the same coding conventions.

Some common data formats

- **Diagnosis codes** - ICD-9/10, SNOMED, issues with rule-outs, ICD-9 to ICD-10 conversion issues (many-to-many mapping, discontinuity in temporal trends spanning switch-over date of 10/1/2015)
- **Procedure codes** - CPT/HCPCS (outpatient), DRG (inpatient), date of administration vs discharge vs test performed vs results
- **Medication codes** - NDC, RxNorm, “homegrown codes”, data for prescriptions (may not be filled or taken) vs dispensings (may not be taken), does not capture OTC use

Unstructured data

- Unstructured data consists of health care providers' narrative clinical notes
- Takes the form of text which typically requires a human reader to understand and interpret
- The Health Story Project estimates that 1.2 billion clinical documents are produced in the US each year, of which 60% are in the form of unstructured notes
- Many data elements potentially of interest for research such as family history and patient behavioral risk factors may be embedded in text notes
- Extracting these data for research use is a challenge but natural language processing (NLP) methods are advancing rapidly

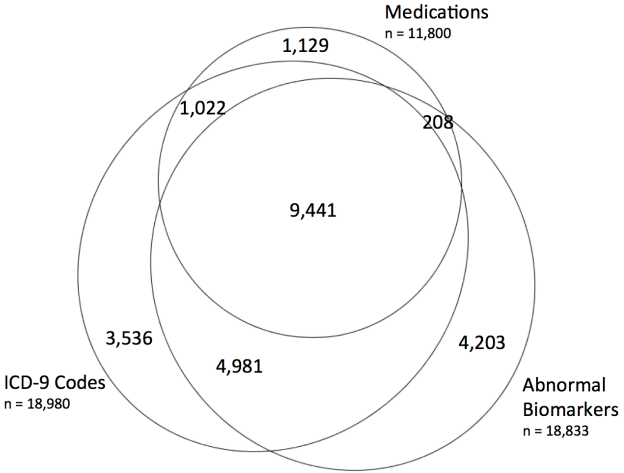
Rule-based Phenotyping

- Most of the existing literature on EHR-derived phenotyping relies on “clinical decision rules”
  - Simple or complex
  - Including one data element or many
  - May include a time component
- Algorithm based on clinical knowledge of the disease and coding practices
- May incorporate structured data as well as unstructured data, often via NLP
- Sometimes validated against gold-standard diagnosis data

Example: Rule-based Phenotyping for T2DM

Variable type	Examples	Format
Diabetes diagnosis	• T2DM • T1DM • DM NOS	ICD-9/10 codes
Medications	• Insulin • Metformin	Prescribing data
Co-morbidities	• PCOS • Obesity	ICD-9/10 codes
Biomarkers	• Glucose • HbA1c	Procedure codes for test administration; numerical results

Agreement among T2DM variables



Adapted from Richesson et al. *J Am Med Inform Assoc* 2013;20:e319-e326.

Typical process for EHR-based phenotype development

- Clinical experts develop a list of potential variables
  - May include condition of interest, symptoms, co-morbidities, common treatments
- Translate list into corresponding structured codes (e.g., ICD-9/10, CPT)
- NLP experts map terms to UMLS concepts
- Extract all occurrences of demographics, codes of interest, biometric data, and laboratory test results from structured data
- Apply NLP to unstructured (narrative text) data

Classification

- Once data have been extracted from the EHR a classification algorithm can be applied to the individual data elements to create a construct of interest
- Gold standard information for supervised learning approaches extracted via manual chart abstraction
- Classification approaches applied to EHR data range from the very simple to the very complex
  - Dichotomous classification based on presence/absence of data elements based on clinical judgment
  - Prediction modeling, e.g. CART, LASSO
  - Machine learning algorithms, e.g. random forests, neural networks
- Performance is typically evaluated based on PPV and NPV relative to gold standard
- Implications of low prevalence for PPV/NPV

Using validated phenotypes

- Ideally, only validated phenotypes should be used
- Validation requires manual chart abstraction and hence can be costly and slow
- Many validated phenotypes are available, for instance, via PheKb (https://phekb.org)
- However, be cautious about assuming that operating characteristics will be the same in your data set as they were in the derivation data set (i.e., lack of portability)

Issues arising in the analysis of EHR-derived phenotypes

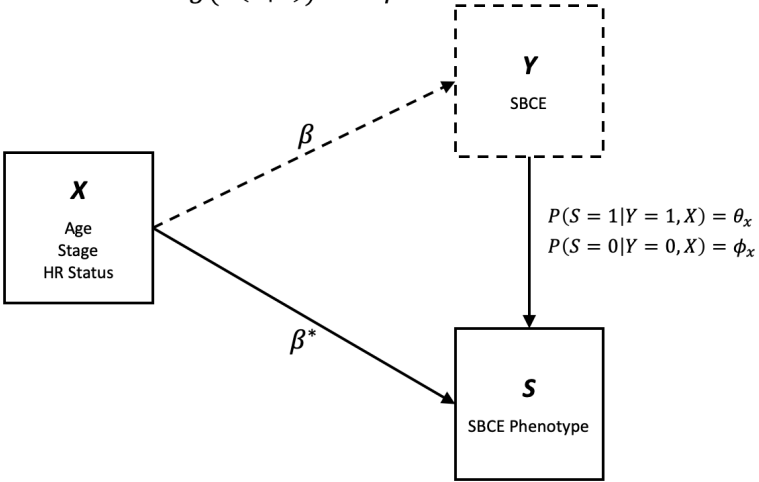
- Regardless of the approach to phenotyping, some residual error will typically remain
- Statistical methods for misclassified outcomes can be adapted to this context
- Some additional challenges in the context of EHR-based research arise due to limited access to validation data
- Accuracy parameters for phenotypes may also exhibit lack of transportability
- We will discuss some alternative approaches to these challenges

Using the EHR to study novel exposures

- EHRs provide the opportunity to identify novel risk factors for disease incidence or outcomes
- EHR allows us to explore a variety of novel risk factors such as treatments for other conditions and co-morbidities
- However, EHR-derived outcomes may exhibit exposure-dependent differences in data quality
  - Only observe the outcome if documented at a healthcare encounter (higher sensitivity among patients with more utilization)
  - Patients interacting with the healthcare system also have more opportunity for erroneous codes to appear in charts (lower specificity among patients with more utilization)

Inference with imperfect phenotypes

$$g\big(E(Y|X)\big) = X^T\beta$$



What can we do about phenotyping error?

- Phenotyping error can lead to substantial bias and inflated type I error
- How can we obtain unbiased estimates in association analyses using EHR-derived phenotypes?
- Challenges in the EHR setting
  - Validation data are costly to obtain and in many cases completely unavailable
  - Phenotyping accuracy is often unknown and may vary widely between derivation data set and other data sets

Classic approach to outcome misclassification

- One binary predictor (X)
- Misclassified binary outcome (S)
- Known sensitivity (θ) and specificity (ϕ); $0 < \theta, \phi < 1$
- Assumes non-differential outcome misclassification, i.e. $\theta = P(S = 1|Y = 1, X) = P(S = 1|Y = 1)$ and $\phi = P(S = 0|Y = 0, X) = P(S = 0|Y = 0)$

	Classified as	
	Diseased	Not diseased
Exposed	a	b
Not exposed	c	d

Classic approach to outcome misclassification

- Naive: $\widehat{OR}_{standard} = (ad)/(bc)$
- Misclassification adjusted: $\widehat{OR} = \frac{a/(a+b)-(1-\phi)}{c/(c+d)-(1-\phi)} \times \frac{d/(c+d)-(1-\theta)}{b/(a+b)-(1-\theta)}$
- Note that

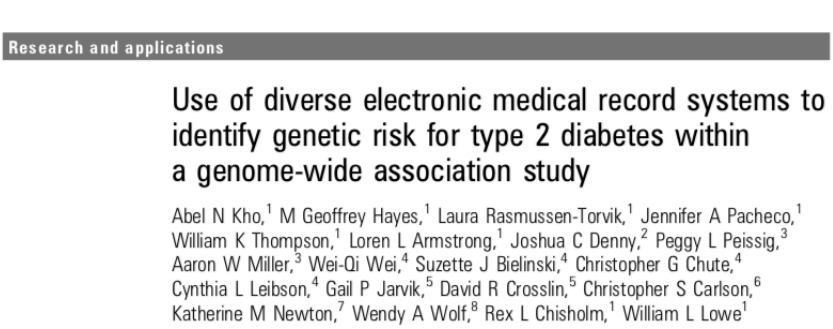
$$\widehat{OR} > \widehat{OR}_{standard} \text{ if } \widehat{OR}_{standard} > 1$$
$$\widehat{OR} < \widehat{OR}_{standard} \text{ if } \widehat{OR}_{standard} < 1$$

Magder LS, Hughes JP. 1997. Logistic regression when the outcome is measured with uncertainty. *Am J Epidemiol.*146(2):195-203.

Extension to logistic regression

- Assume non-differential misclassification
- Let $P(Y_i = 1) = expit(\beta^T X_i)$ then by Bayes rules
  - $\hat{P}(Y_i = 1|S_i = 1) = \frac{\theta expit(\beta^T X_i)}{\theta expit(\beta^T X_i)+(1-\phi)(1-expit(\beta^T X_i))}$
  - $\hat{P}(Y_i = 1|S_i = 0) = \frac{(1-\theta) expit(\beta^T X_i)}{(1-\theta) expit(\beta^T X_i)+\phi(1-expit(\beta^T X_i))}$
- An EM algorithm to estimate β
  1. Perform weighted logistic regression, each subject included as both diseased and non-diseased with weights $\hat{P}(Y_i = 1|S = k)$
  2. Update weights using new values for $\hat{\beta}$
  3. Return to (1)

Case Study 1



Research and applications

Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study

Abel N Kho,[1] M Geoffrey Hayes,[1] Laura Rasmussen-Torvik,[1] Jennifer A Pacheco,[1] William K Thompson,[1] Loren L Armstrong,[1] Joshua C Denny,[2] Peggy L Peissig,[3] Aaron W Miller,[3] Wei-Qi Wei,[4] Suzette J Bielinski,[4] Christopher G Chute,[4] Cynthia L Leibson,[4] Gail P Jarvik,[5] David R Crosslin,[5] Christopher S Carlson,[6] Katherine M Newton,[7] Wendy A Wolf,[8] Rex L Chisholm,[1] William L Lowe[1]

- Validation of an EHR-derived type 2 diabetes mellitus (T2DM) phenotype
- Performance compared across multiple healthcare systems
- Investigate association between a genotype and T2DM status

eMERGE

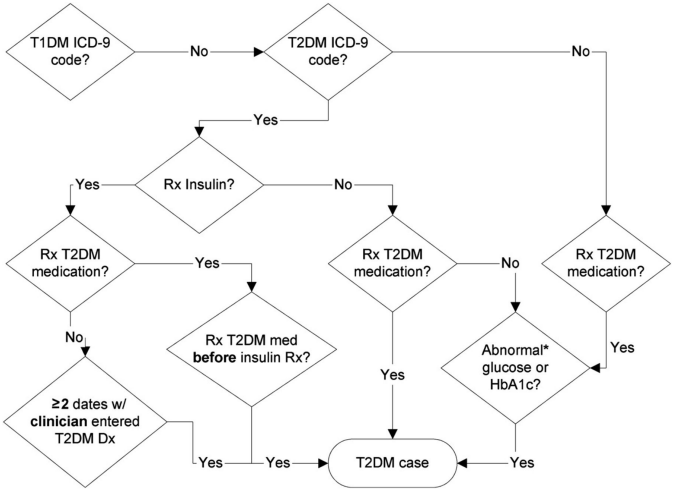
- Multisite consortium consisting of healthcare data linked to biorepository specimens
- Sites differ in denominator populations and structure of healthcare data

Table 1 Overview of participating institutions' EMR and biorepositories and recruitment models

Institution	Biorepository overview	Recruitment model	Repository size	EMR summary
Marshfield Clinic Research Foundation (Marshfield, Wisconsin, USA)	Personalized medicine research project: Geographically defined cohort within an integrated regional healthcare system	Population based	20 000 98% Caucasian	Comprehensive internally developed EMR since 1985 75% participants have 20+ years medical history
Northwestern University (Chicago, Illinois, USA)	Nugene project: Northwestern affiliated hospitals and outpatient clinics	Population based	10 000 12% AA 8% Hispanic	Comprehensive vendor-based inpatient (Cerner, Kansas City, Missouri, USA) and outpatient (Epic Systems, Verona, Wisconsin, USA) EMR since 2000 20+ years ICD-9 data
Vanderbilt University (Nashville, Tennessee, USA)	BioVU: Vanderbilt Clinic, diverse outpatient clinic	Population based	100 000 11% AA	35+ Years medical history data Comprehensive internally developed EMR since 2000
Group Health Cooperative (Seattle, Washington, USA)	GHC Biobank Alzheimer's disease patient registry and adult changes in thought study	Disease-specific cohort	4000 (>96% Caucasian)	Comprehensive vendor-based (Epic Systems) EMR since 2004 20+ years pharmacy data 15+ years ICD-9 data
Mayo Clinic (Rochester, Minnesota, USA)	Vascular diseases biorepository	Disease-specific cohort	3500 (>96% Caucasian)	Comprehensive internally developed EMR since 1995 40-year history of data extraction

AA, African ancestry; EMR, electronic medical record; GHC, Group Health Cooperative; ICD-9, International Classification of Diseases, version 9.

T2DM Phenotype



Phenotype validation

- Algorithm adapted to individual sites' data (e.g. ICD-9 250.00 for T1D and T2D)
- Phenotype for cases and controls validated against manual chart review
- Charts sampled on phenotype status so only PPV and NPV can be estimated

Table 3 Summary of chart review results at three participating sites

	Manual chart review								
	Northwestern University*			Vanderbilt University†			Marshfield Clinic‡		
	Case	Control	Total	Case	Control	Total	Case	Control	Total
EMR prediction									
Case	56	1	57	50	0	50	99	1	100
Control	0	43	43	0	50	50	1	49	50
Total	56	44	100	50	50	100	100	50	150

Genetic risk factor for T2DM

Table 4 Association results for rs7903146 in TCF7L2 with type 2 diabetes

	N	Allele frequencies (T)		OR	L95	U95	p Value
		Cases	Controls				
GHC EA	813	0.327	0.248	1.59	1.25	2.02	0.0002
Marshfield EA	930	0.323	0.243	1.49	1.15	1.93	0.0025
Mayo EA	1159	0.303	0.281	1.17	0.96	1.43	0.1177
NU EA	1229	0.353	0.274	1.51	1.23	1.86	9.27×10^{-5}
VU EA	396	0.334	0.263	1.42	0.99	2.04	0.0601
Cross cohort EA (Meta)		0.328	0.265	1.41			2.98×10^{-10}
NU AA	294	0.353	0.241	0.94	0.24	2.39	0.08672
VU AA	1021	0.353	0.260	1.35	0.11	2.07	2.25×10^{-6}
Cross cohort AA (Meta)		0.353	0.258	1.64			5.30×10^{-7}
Cross-cohort all (Meta)		0.334	0.263	1.46			2.05×10^{-15}

Strengths & Limitations

- EHR data provide the opportunity to look at a large variety of phenotypes
- But EHR-derived phenotypes are imperfect and phenotyping error may vary across sites
- Methods for misclassified outcomes (not used in this case study) can reduce bias in association parameter estimates
- Linkage to health system biobanks facilitates large-scale genetics studies
- Consideration should be given to the source population as this may induce selection bias

Tutorials

- Throughout this course we will use synthetic (i.e., fake) EHR-derived data sets based on the case studies to explore some of the challenges of EHR data and strategies for analyzing them
- The data sets and R code for working with them are available at `https://rhubb.github.io/ENAR_short_course/`
- The tutorials present questions and solutions using R to guide you through an analysis similar to that discussed in each case-study

Data Quality Case Study: PEDSnet

- PEDSnet: A PCORI-funded consortium of 8 children's hospitals
  - Includes data collected in routine clinical encounters for $\sim$5 million children
- Investigated pediatric Type 2 Diabetes Mellitus (T2DM) in a high risk cohort:
  - Children age 10-18 years, at least two clinical encounters between 2001-2017
  - On at least one occasion BMI z-score in excess of the 95th percentile for age and sex
- The tutorial 1 data set was generated by sampling from the distributions of data elements contained in the real PEDSnet data
- The resultant data reflect the features of the PEDSnet data

Tutorial 1

Outline

FDA Framework for RWE

- The 21st Century Cures Act was designed to accelerate medical product development
- This included identifying ways to use RWE to approve a new indication for a drug or satisfy post-approval study requirements
- RWD (principally EHR data) can be used in many ways in medical product development
  - Hypothesis generation
  - Identifying novel risk factors
  - Assessing trial feasibility
  - Defining inclusion/exclusion criteria and baseline characteristics
  - Informing prior probability distributions for Bayesian models
  - Assembling geographically distributed research cohorts for rare diseases
  - Generating evidence on product safety
  - **Generating evidence on product effectiveness**

Using RWD to generate evidence on effectiveness

- Establishing evidence of effectiveness using EHR data is more challenging
- Observational studies typically have too many sources of potential bias to be relied upon in the regulatory process, confounding by indication particularly problematic
- Pragmatic clinical trials embedded in healthcare systems can be used
- In rare cases, FDA has approved a new product based on historical control arm data from EHR
  - Only applies in oncology or rare disease settings where randomization is considered unethical or infeasible

Potential uses of RWD in effectiveness studies

- Facilitate pragmatic trials/large simple trials embedded in healthcare systems
- External control arm for a single arm trial
- Additional control data (hybrid control arm) for an RCT
- Observational comparative effectiveness analyses

Challenges to validity of RWE from a regulatory perspective

- Lack of alignment of trial and RWD inclusions/exclusion criteria
- Lack of alignment of trial and RWD definitions for key data elements (treatments, outcomes, confounders, etc)
- Lack of capture of endpoints
- Missing data
- Channeling bias and confounding

FDA considerations for use of RWD

- Are the data fit for use?
  - Consideration of data quality issues
  - Missing data due to porosity of EHR
- Is the study design adequate to provide evidence?
  - Sources of bias
- Does the study conduct meet FDA regulatory requirements?
  - Study monitoring
  - Data integrity

`https://www.fda.gov/media/120060/download`

External control arms

- EHR data on patients receiving standard of care can be used as a comparator arm for single arm trials
- This may be a stronger design than the single arm trial alone which relies on historical comparisons
- However, selection bias and information bias can arise in this setting

Motivation for the external control design

- Rare diseases or small sub-groups may be difficult to enroll trials; the EC design allows all eligible patients to be allocated to intervention arm
- Lack of randomization may be appealing to some patients and clinicians, speeding enrollment
- Addresses ethical issues on randomization to placebo

Selection of external controls

- Characteristics of patients represented in EHR may differ from those participating in trials
  - For instance, patients participating in oncology trials tend to be healthier than the general patient population
- Must be able to harmonize inclusion/exclusion criteria between trial and EHR
- This can be challenging if trial inclusion criteria include characteristics not typically assessed (or recorded) in routine care

Selection of external controls

- Differences in recorded patient characteristics can be accounted for in analysis
- Can be achieved via standardization, matching, weighting, or regression adjustment often using propensity score approaches
- Important to consider which patient characteristics are available in EHR data for adjustment as well as differences in ascertainment/quality of data elements between trial and EHR

Multivariate matching

- In matching, the first impulse is to try to match each exposed subject to an unexposed subject who appears nearly the same in terms of observed covariates.
- However, this is impractical when there are many covariates. For instance, with 20 binary covariates, there are 2^{20} or about a million patterns of covariates.
- Randomization produces covariate balance, in expectation, not perfect matches. Perfect matches are not needed to balance observed covariates.
- Multivariate matching methods attempt to produce matched pairs or sets that balance observed covariates, so that distributions of observed covariates are similar in exposed and unexposed groups.

The propensity score

- The propensity score is a fundamental tool for constructing matched sets (Rosenbaum PR, Rubin DB. *Biometrika*. 1983;70(1):41-55.)
- The *propensity score* $e(\mathbf{x})$ is the probability that a person with observed covariates $\mathbf{X} = \mathbf{x}$ is in exposure group $Z = 1$

$$e(\mathbf{x}) = P(Z = 1|\mathbf{X} = \mathbf{x})$$

- In context of trials with external control arms, Z will represent membership in the trial cohort
- The propensity score is a scalar function of $\mathbf{x}$ that summarizes the information required to balance the multivariate distribution of the covariates between the two treatment groups.

The propensity score

- Rather than stratifying or matching exactly on $\mathbf{x}$, form matched sets based on probability of being in the trial cohort, $e(\mathbf{x})$.
- *Exact matching on the propensity score*: within a stratum or matched set, units may have different values of $\mathbf{x}$ but have the same propensity score.
- Assuming there are no unobserved confounders, exact matching on the propensity score yields a conditional distribution of Z given $\mathbf{x}$ that is the same as a randomized experiment.
- The propensity score can be estimated by predicting treatment assignment from the observed covariates using a prediction model such as logistic regression.

Propensity score matching

1. Estimate the propensity score using a model for binary outcomes such as logistic regression including all key confounders, interactions and higher order terms. Variables associated only with outcome should be included. Variables associated only with exposure (trial membership) should not.
2. We form matched pairs to minimize the absolute propensity score differences between the matched pairs. A variety of matching algorithms are implemented in the *MatchIt* package in R.
3. We examine the standardized mean differences (SMD) on the matched sample of all covariates, as well as interactions and higher order terms. Variables wiith an SMD < 0.05 or 0.1 are typically considered to be "balanced".

$$\text{SMD} = \frac{(\bar{x}_{treatment} - \bar{x}_{control})}{s_{treatment}}$$

Mahalanobis matching

- While the propensity score acheives covariate balance, on average, an alternative multivariate matching strategy would be to try to directly match subjects based on similarity in their covariates
- If $\hat{\Sigma}$ is the sample covariace matrix of $\mathbf{x}$, then the estimated Mahalanobis distance between $\mathbf{x}_k$ and $\mathbf{x}_l$ is

$$(\mathbf{x}_k - \mathbf{x}_l)^T \hat{\Sigma}^{-1} (\mathbf{x}_k - \mathbf{x}_l)$$

- The Mahalanobis distance works well for multivariate normal data
- When the data are not normal, Mahalanobis distance can perform undesirably, especially in the presence of outliers.
- If one covariate contains extreme outliers, its standard deviation will be inflated, and the Mahalanobis distance will tend to ignore that covariate in matching.

Exposure (treatment) ascertainment

- Details of treatment must be adequately captured in EHR
- Supportive care may differ between trial and routine care and may not be well-recorded
- Care received outside of the health system providing data will not be captured
- For pharmacologic exposures, information on adherence generally not available
- Typically not possible to implement intent to treat analysis; Analyses usually as treated

Effect of missing outcome data

Carrigan G, et al. 2019. An evaluation of the impact of missing deaths on overall survival analyses of advanced non–small cell lung cancer patients conducted in an electronic health records database. *Pharmacoepidemiology and drug safety*. doi:10.1002/pds.4758

- Used data from the Flatiron Health database of community oncology centers and gold-standard death data from National Death Index to investigate effect of missing information on mortality on hazard ratio estimates
- Missing outcomes substantially inflated estimates of median survival time but had little effect on hazard ratios
- However, using EHR data as external control arm resulted in significant bias
- Implications for between-site comparisons if loss to follow-up patterns differ

Hybrid control trials

- The hybrid control trial is similar to the external control design but includes both randomized controls and external controls
- Adding an EHR-based external control arm to an RCT has many of the same advantages as the external control design
- An additional strength is that similarity between EHR and randomized controls can be directly assessed to evaluate possible bias in estimating treatment effects attributable to differences between the external control and trial patients

Case Study 2

Breast Cancer Research and Treatment (2020) 184:161–172
https://doi.org/10.1007/s10549-020-05838-5

CLINICAL TRIAL

Check for updates

Real-world survival outcomes of heavily pretreated patients with refractory HR+, HER2−metastatic breast cancer receiving single-agent chemotherapy—a comparison with MONARCH 1

Hope S. Rugo[1] · Veronique Dieras[2] · Javier Cortes[3,4,5] · Debra Patt[6,7] · Hans Wildiers[8] · Joyce O'Shaughnessy[9] · Esther Zamora[5] · Denise A. Yardley[10] · Gebra Cuyun Carter[11] · Kristin M. Sheffield[11] · Li Li[11] · Valerie A. M. Andre[11] · Xiaohong I. Li[11] · Martin Frenzel[11] · Yu-Jing Huang[11] · Maura N. Dickler[11] · Sara M. Tolaney[12]

- Single arm trial (MONARCH) of abemaciclib in treatment refractory HR+, HER2- metastatic breast cancer (MBC)
- Compared to RWD cohort of patients receiving single agent chemotherapy in the Flatiron Health oncology EHR database
- Flatiron data included 2.1 million cancer patients, of whom 15,000 were MBC patients

Inclusion/Exclusion Criteria

	Trial	**RWD**
HR status	Yes	Yes
HER2 status	Yes	Yes
Prior endocrine therapy	Yes	Not required
Prior taxane therapy	Yes	No
Number of prior chemo regimens	1-2	1-2
ECOG Performance Status	0-1	0-1 (exclude missing)
Prior CDK4&6 therapy	Not permitted	Not permitted (if documented)
CNS metastases	Not permitted	Not permitted (based on ICD codes)

- After applying I/E criteria, RWD cohort N = 281

Difference in Patient Characteristics Before Matching

	Trial (N = 132) N (%)	**RWD (N = 281)** N (%)
Age <65 years	90 (68.2)	159 (56.6)
White race	124 (93.9)	192 (68.3)
Prior chemotherapy		
1 Regimen	67 (50.8)	159 (56.6)
2 Regimens	65 (49.2)	122 (43.4)
Prior endocrine therapy		
0 Regimen	17 (12.9)	114 (40.6)
1 Regimens	48 (36.4)	77 (27.4)
1 Regimen	25 (18.9)	54 (19.2)
2 Regimens	42 (31.8)	36 (12.8)
PR-	35 (26.5)	99 (35.2)

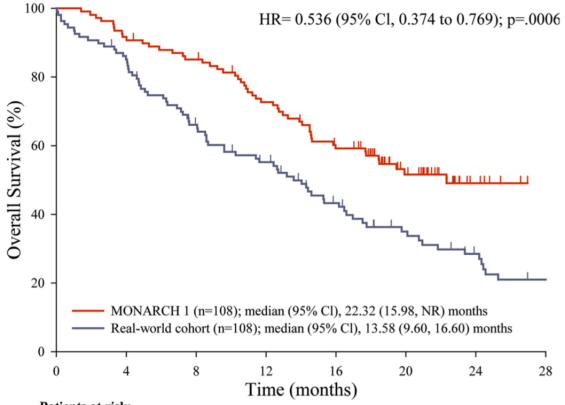
Difference in Patient Characteristics After Matching

	Trial (N = 108) N (%)	**RWD (N = 108)** N (%)
Age <65 years	72 (66.7)	71 (65.7)
White race	100 (92.6)	103 (95.4)
Prior chemotherapy		
1 Regimen	61 (56.5)	64 (59.3)
2 Regimens	47 (43.5)	44 (40.7)
Prior endocrine therapy		
0 Regimen	17 (15.7)	16 (14.8)
1 Regimens	40 (37.0)	77 (38.9)
1 Regimen	23 (21.3)	24 (22.2)
2 Regimens	28 (25.9)	26 (24.1)
PR-	29 (26.9)	31 (28.7)

Statistical Analysis

- Primary outcome: overall survival (OS)
- Analyzed via Kaplan-Meier and Cox PH
- Differences in patient characteristics addressed via Mahalanobis matching (N = 108 per arm in matched cohort)
  - Matched on: age group, race group, number of prior chemo regimens, number of prior endocrine therapy regimens, prior capecitabine use, and PR status

Results


HR= 0.536 (95% CI, 0.374 to 0.769); p=.0006
Overall Survival (%)
100
80
60
40
20
0
0
4
8
12
16
20
24
28
Time (months)
MONARCH 1 (n=108); median (95% CI), 22.32 (15.98, NR) months
Real-world cohort (n=108); median (95% CI), 13.58 (9.60, 16.60) months
Patients at risk:
MONARCH 1
108 99 91 76 60 32 8 0
Real-world cohort
108 91 67 54 38 27 19 13

Strengths & Limitations

- EHR data contain relatively rich information (compared to claims) on patient characteristics, facilitating harmonization of many I/E criteria
- Some characteristics not available or frequently missing in EHR data (e.g. ECOG PS)
- Information on some characteristics is available but incomplete (e.g. prior chemotherapy use)
- Some information is of poor quality (e.g. site of metastasis based on ICD codes)
- Outcome ascertainment in RWD may be incomplete
- Care context may differ between trial and RWD due to differences in participating centers, time period, patient populations

Tutorial 2

Outline

Motivation: a large-scale CER research using observational data

THE LANCET
Volume 394, Issue 10211, 16–22 November 2019, Pages 1816-1826

Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis

Marc A Suchard, Martijn J Schuemie, Harlan M Krumholz, Seng Chan You, Ruijun Chen, Nicole Pratt, Christian G Reich, Jon Duke, David Madigan, George Hripcsak, Patrick B Ryan

Lancet 2019; 394: 1816-26
Published Online October 24, 2019 https://doi.org/10.1016/S0140-6736(19)32317-7

See **Comment** page 1782

Department of Biostatistics, Fielding School of Public Health (Prof M A Suchard MD, M J Schuemie PhD), **and Department of Biomathematics, David Geffen School of Medicine at UCLA** (Prof M A Suchard), **University of California, Los Angeles, CA, USA; Epidemiology Analytics, Janssen Research &**

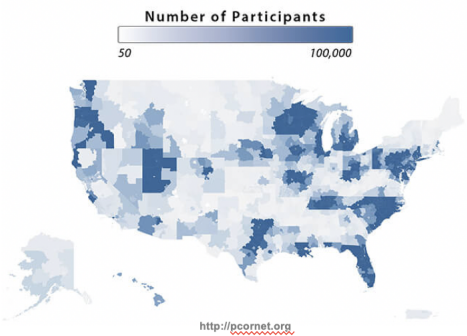
Summary

Background Uncertainty remains about the optimal monotherapy for hypertension, with current guidelines recommending any primary agent among the first-line drug classes thiazide or thiazide-like diuretics, angiotensin-converting enzyme inhibitors, angiotensin receptor blockers, dihydropyridine calcium channel blockers, and non-dihydropyridine calcium channel blockers, in the absence of comorbid indications. Randomised trials have not further refined this choice.

Methods We developed a comprehensive framework for real-world evidence that enables comparative effectiveness and safety evaluation across many drugs and outcomes from observational data encompassing millions of patients, while minimising inherent bias. Using this framework, we did a systematic, large-scale study under a new-user cohort design to estimate the relative risks of three primary (acute myocardial infarction, hospitalisation for heart failure, and stroke) and six secondary effectiveness and 46 safety outcomes comparing all first-line classes across a global network of six administrative claims and three electronic health record databases. The framework addressed residual confounding, publication bias, and p-hacking using large-scale propensity adjustment, a large set of control outcomes, and full disclosure of hypotheses tested.

Motivation: Distributed Health Data Networks

- No data centralization: data holders maintain control over data
  - Participants adopt common data model (CDM)
  - Analyses performed distributively (email, central server, etc.)
  - No patient-level data transfer
- Sentinel Initiative
  - FDA: Post-market safety surveillance
  - 16 data partners contribute billing data, EHRs
- PCORnet
  - National patient-centered clinical research network for comparative effectiveness research
  - 68 million patients



OHDSI's global research community

Promise and challenges of multi-site studies

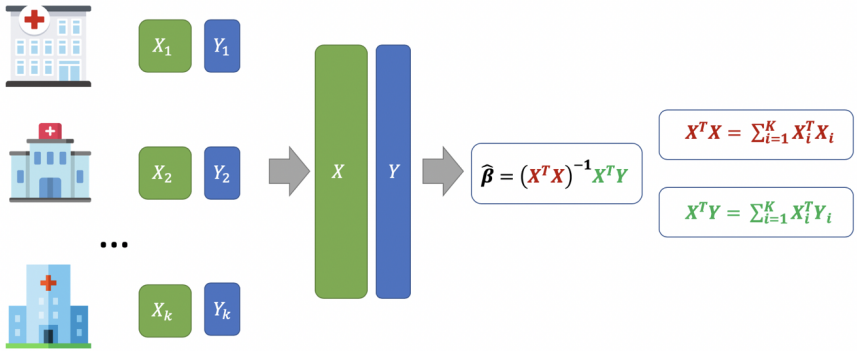
- Covers broader population
  - Results are more generalizable
  - Better statistical power
  - Opportunities of studying rare diseases
- Ecosystem of bringing together expertise from different investigators

- Sharing individual-patient level data is challenging

- Need iterative collaboration/coordination among investigators from different institutes
  (consuming lots of time, effort, funding)

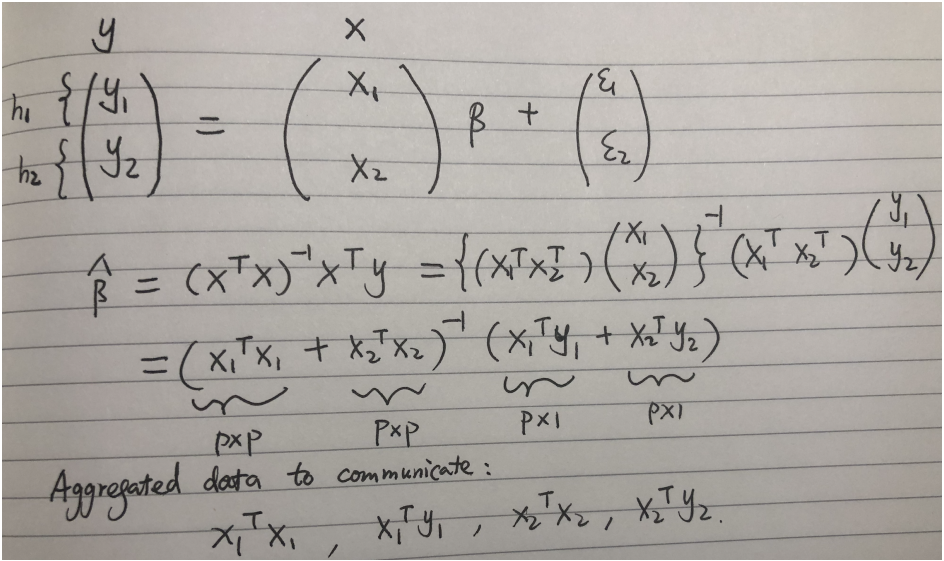
Distributed Linear Regression - a useful result

- Do we really need to share patient-level data?



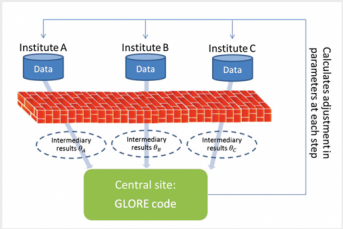
Chen et al. (2006) Regression cubes with lossless compression and aggregation.

A bit more detail



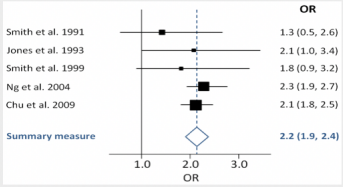
Current algorithms of distributed analysis

- **Lossless** --- obtain the identical results as the combined analysis
  - For simple linear regression
    - **Distributed linear regression** (Chen et al. 2006)
  - For non-linear regressions, iterative algorithms are needed
    - **Binary outcome**: Grid Binary LOgistic REgression (GLORE) (Wu et al. 2012)
    - **Time-to-event outcome**: WebDISCO: a Web service for distributed Cox model (Lu et al. 2015)



Wu et al. 2012, *JAMIA*

- **One-shot (non-iterative)** --- only requires the collaborative sites to exchange aggregated data ***once***
  - Averaging local estimates
    - Meta-analysis, distributed PCA (Fan et al. 2018), distributed LDA (Lu et al. 2019), ...
  - Surrogate likelihood (Jordan et al. 2018)

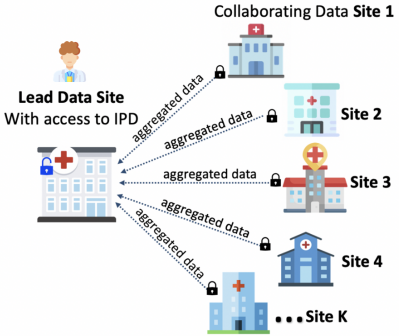


Our Goal

- **Safe** --- protect patient-level information
- **Accurate** --- provide estimates which are close to the pooled analysis results
- **Efficient** --- no iterative communication across collaborating sites

Inspiration: an interesting observation – unique architecture of DRN

- An investigator at a hospital do have access to the patient-level data of that hospital
- Collaborating hospitals can share aggregated data



JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2019, VOL. 114, NO. 526, 668–681: Theory and Methods
https://doi.org/10.1080/01621459.2018.1429274

Taylor & Francis
Taylor & Francis Group

Communication-Efficient Distributed Statistical Inference

Michael I. Jordan[a], Jason D. Lee[b], and Yun Yang[c]

[a]Department of Statistics, University of California Berkeley, Berkeley, CA; [b]Institute of Computational and Mathematical Engineering, Stanford University, Cupertino, CA; [c]Statistical Science, Duke University, Durham, NC

ABSTRACT

We present a *communication-efficient surrogate likelihood* (CSL) framework for solving distributed statistical inference problems. CSL provides a communication-efficient surrogate to the global likelihood that can be used for low-dimensional estimation, high-dimensional regularized estimation, and Bayesian inference. For low-dimensional estimation, CSL provably improves upon naive averaging schemes and facilitates the construction of confidence intervals. For high-dimensional regularized estimation, CSL leads to a minimax-optimal estimator with controlled communication cost. For Bayesian inference, CSL can be used to form a communication-efficient quasi-posterior distribution that converges to the true posterior. This quasi-posterior procedure significantly improves the computational efficiency of Markov chain Monte Carlo (MCMC) algorithms even in a nondistributed setting. We present both theoretical analysis and experiments to explore the properties of the CSL approximation. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received December 2016
Revised December 2017

KEYWORDS
Communication efficiency;
Distributed inference;
Likelihood approximation

An idea

Can we use local hospital's patient-level data and collaborating hospitals' aggregated data to do multi-site data analysis?

ODAL

One-shot Distributed Algorithm for Logistic Regression

Journal of the American Medical Informatics Association, 27(3), 2020, 376–385
doi: 10.1093/jamia/ocz199
Advance Access Publication Date: 9 December 2019
Research and Applications

Research and Applications

Learning from electronic health records across multiple sites: A communication-efficient and privacy-preserving distributed algorithm

Rui Duan [1], Mary Regina Boland [1], Zixuan Liu[2], Yue Liu,[3] Howard H. Chang,[4] Hua Xu,[5] Haitao Chu,[6] Christopher H. Schmid,[7] Christopher B. Forrest,[8] John H. Holmes,[1] Martijn J. Schuemie [9], Jesse A. Berlin,[9] Jason H. Moore,[1] and Yong Chen [1]

The likelihood functions

- **Combined likelihood function** (if data could be shared)

$$L(\beta) = \frac{1}{nK}\sum_{j=1}^{K}\sum_{i=1}^{n}\{y_{ij}x_{ij}^{T}\beta - \log\{1+\exp(x_{ij}^{T}\beta)\}\}$$

- **Local likelihood function** (assume local site to be the first site, j=1)

$$L_1(\beta) = \frac{1}{n}\sum_{i=1}^{n}\{y_{i1}x_{i1}^{T}\beta - \log\{1+\exp(x_{i1}^{T}\beta)\}\}$$

How to borrow aggregated information from other sites to make $L_1(\beta)$ more like $L(\beta)$?

The surrogate likelihood (SL) approach

- For an initial value $\bar{\beta}$,

$$L(\beta) = L(\bar{\beta}) + \nabla L(\bar{\beta})^T(\beta - \bar{\beta}) + \sum_{t=2}^{\infty} \frac{1}{t!} \nabla^t L(\bar{\beta})(\beta - \bar{\beta})^{\otimes t}$$

$$L_1(\beta) = L_1(\bar{\beta}) + \nabla L_1(\bar{\beta})^T(\beta - \bar{\beta}) + \sum_{t=2}^{\infty} \frac{1}{t!} \nabla^t L_1(\bar{\beta})(\beta - \bar{\beta})^{\otimes t}$$

$$\sum_{t=2}^{\infty} \frac{1}{t!} \nabla^t L_1(\bar{\beta})(\beta - \bar{\beta})^{\otimes t} = L_1(\beta) - L_1(\bar{\beta}) - \nabla L_1(\bar{\beta})^T(\beta - \bar{\beta})$$

First-order SL function

$$\tilde{L}^1(\beta) = L_1(\beta) + \{\nabla L(\bar{\beta}) - \nabla L_1(\bar{\beta})\}^T \beta$$

$$\nabla L(\bar{\beta}) = \frac{1}{K} \sum \nabla L_j(\bar{\beta}); \quad \nabla L_j(\bar{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \{y_{ij} - \text{expit}(x_{ij}^T \bar{\beta})\} x_{ij}$$

Increase the approximation accuracy

$$L(\beta) = L(\bar{\beta}) + \nabla L(\bar{\beta})^T(\beta - \bar{\beta}) + \frac{1}{2}\nabla^2 L(\bar{\beta})(\beta - \bar{\beta})^{\otimes 2} + \sum_{t=3}^{\infty} \frac{1}{t!}\nabla^t L(\bar{\beta})(\beta - \bar{\beta})^{\otimes t}$$

$$L_1(\beta) = L_1(\bar{\beta}) + \nabla L_1(\bar{\beta})^T(\beta - \bar{\beta}) + \frac{1}{2}\nabla^2 L_1(\bar{\beta})(\beta - \bar{\beta})^{\otimes 2} + \sum_{t=3}^{\infty} \frac{1}{t!}\nabla^t L_1(\bar{\beta})(\beta - \bar{\beta})^{\otimes t}$$

Second-order SL function

$$\tilde{L}^2(\beta) = L_1(\beta) + \{\nabla L(\bar{\beta}) - \nabla L_1(\bar{\beta})\}^T\beta + \frac{1}{2}(\beta - \bar{\beta})^T\{\nabla^2 L(\bar{\beta}) - \nabla^2 L_1(\bar{\beta})\}^T(\beta - \bar{\beta})$$

$$\nabla^t L(\bar{\beta}) = \frac{1}{K}\sum \nabla^t L_j(\bar{\beta}), \text{for } t = 1, 2.$$

$$\nabla L_j(\bar{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\{y_{ij} - \text{expit}(x_{ij}^T\bar{\beta})\}x_{ij};\ \nabla^2 L_j(\bar{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\text{expit}(x_{ij}^T\bar{\beta})\{1 - \text{expit}(x_{ij}^T\bar{\beta})\}x_{ij}\ x_{ij}^T.$$

Surrogate likelihood estimates

- First-order algorithm (ODAL1)

$$\tilde{\beta}^1 = \text{argmax}_\beta \, \tilde{L}^1(\beta)$$

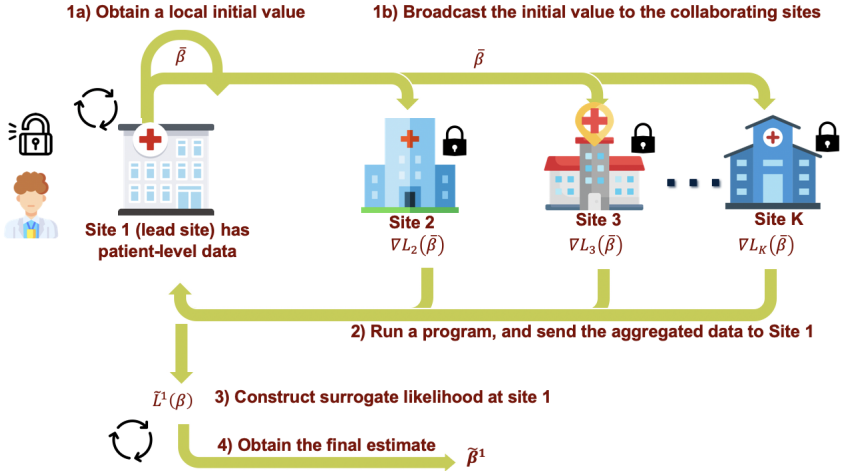
- Second-order algorithm (ODAL2)

$$\tilde{\beta}^2 = \text{argmax}_\beta \, \tilde{L}^2(\beta)$$

- Initial estimator:

$$\bar{\beta} = \text{argmax}_\beta \, L_1(\beta)$$

ODAL step-by-step illustration:



Theorem

Under mild regularity conditions, the proposed estimators $\tilde{\beta}^1$, and $\tilde{\beta}^2$ satisfy,

$$\sqrt{Kn}(\tilde{\beta}^t - \beta^*) \to N(0, \{\mathbb{E}\nabla^2 f(y, x; \beta^*)\}^{-1})$$

for $t = 1, 2$, as $Kn \to \infty$, where

$$f(y, x; \beta) = yx^T\beta - \log\{1 + \exp(x^T\beta)\}.$$

- Inference at local site (no extra communication)

$$\hat{V}_1 = \frac{1}{Kn}\{\nabla^2 L_1(\tilde{\beta}^1)\}^{-1}; \hat{V}_2 = \frac{1}{Kn}\{\nabla^2 L_1(\tilde{\beta}^2)\}^{-1}$$

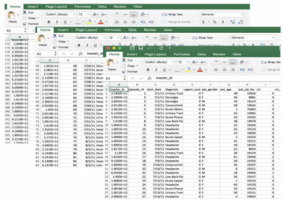
ODAL advantages:

Research question:
What are the risk factors of acute myocardial infarction?

Fit a logistic model with 4 risk factors.

How to perform multicenter analysis?

Strategy 1: pooled analysis



Strategy 2: ODAL

Two results are nearly the same. (Duan el al. 2019)

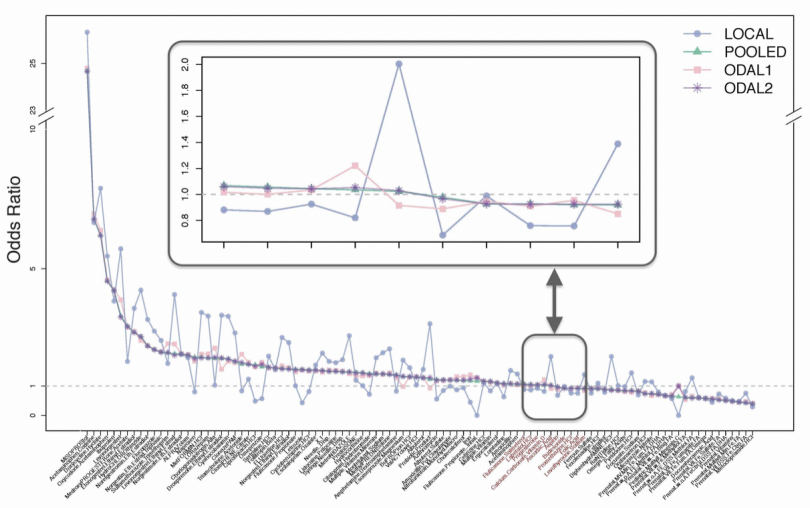
Evaluation through EHR data from Penn Medicine

- Outcome: normal pregnancy (Z34 ICD-10 codes or a V22 ICD-9 code) or a fetal loss (ICD-9 code 630-639 or ICD-10 code O00-O08)
- Exposure: 100 most common medications prescribed during pregnancy, prevalence ranging from 0.05% - 20%.

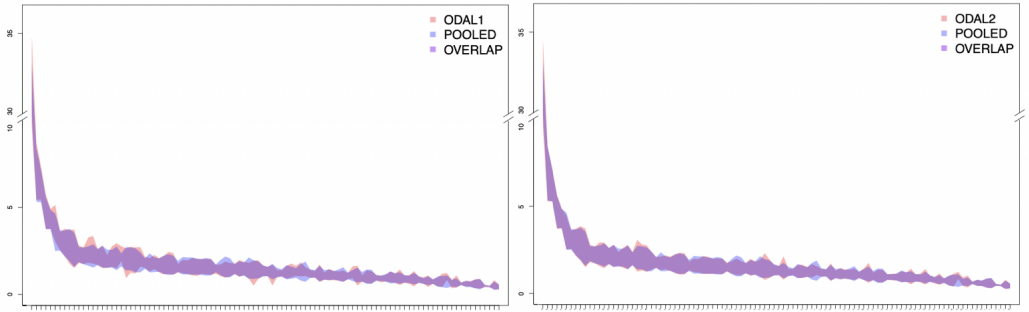
Table 1. Demographics of Pregnancies Treated at UPenn Health System (UPHS)

Demographics	Normal Pregnancy (N=30,810)	Fetal Loss (M=4,763)	P-value
Race			
White *	13911 (45.2%)	2291 (48.1%)	
African American	12918 (41.9%)	1871 (39.3%)	
Other	1916 (6.2%)	274 (5.8%)	
Asian	2065 (6.7%)	327 (6.9%)	
Age	29.40	32.15	<0.001
Weight (pounds)	123.45	115.43	<0.001
Body Mass Index	16.95	16.61	0.043

Results


LOCAL
POOLED
ODAL1
ODAL2
Odds Ratio
25
23
10
5
1
0
2.0
1.8
1.6
1.4
1.2
1.0
0.8

Results - inference



- ODAL2 needs to transfer an extra hessian matrix.
- ODAL2 provides more accurate estimation than ODAL1.

Results - inference

- ODAL: communication-efficient distributed algorithm for logistic regression
- What's next?

Distributed algorithm for time to event outcomes

ODAC

One-shot **D**istributed **A**lgorithm for **C**ox Proportional Hazards Model

Journal of the American Medical Informatics Association, 0(0), 2020, 1–9
doi: 10.1093/jamia/ocaa044
Research and Applications

AMIA INFORMATICS PROFESSIONALS, LEADING THE WAY.

OXFORD

Research and Applications

Learning from local to global: An efficient distributed algorithm for modeling time-to-event data

Rui Duan,[1,†] Chongliang Luo,[1,†] Martijn J. Schuemie,[2,†] Jiayi Tong,[1] C. Jason Liang,[3] Howard H. Chang,[4] Mary Regina Boland,[1] Jiang Bian,[5,6] Hua Xu,[7] John H. Holmes,[1] Christopher B. Forrest,[8] Sally C. Morton,[9] Jesse A. Berlin,[10] Jason H. Moore,[1] Kevin B. Mahoney,[11] and Yong Chen[1]

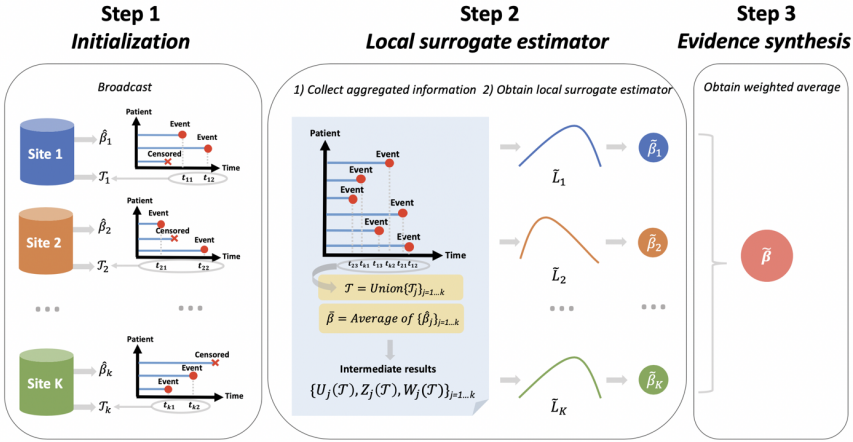
New technical challenges

- For the i-th observation in the j-th site
  - y_{ij} --- time to event
  - δ_{ij} ---censoring indicator
  - x_{ij} --- observed risk factors
- Combined log partial likelihood function

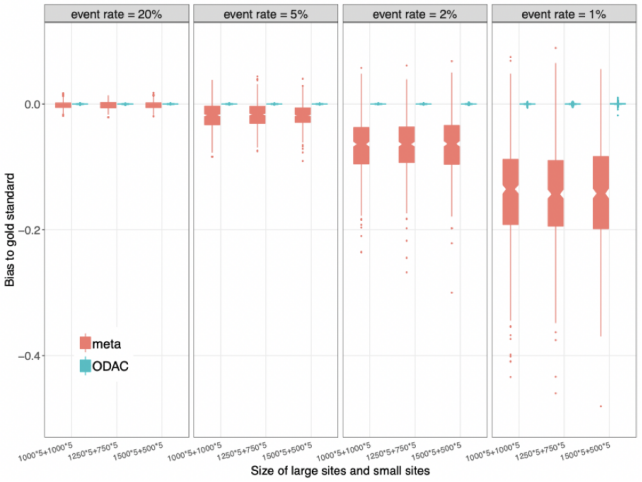
$$L(\beta) = \frac{1}{nK}\sum_{j=1}^{K}\sum_{i=1}^{n}\delta_{ij}\log\frac{\exp(x_{ij}^T\beta)}{\sum_{(l,m)\in R(y_{ij})}\exp(x_{lm}^T\beta)}$$

- Combined likelihood function cannot be written as sum of individual terms
- Denominator involves data from all sites

ODAC - step by step illustration



Benefit in studying rare events



- Meta-analysis has increasing bias when event is rarer.
- ODAC provides estimates close to the pooled analysis.

Case study

- Four claims datasets.
- Population: pharmacologically-treated major depressive disorder
- Outcome: acute myocardial infarction (AMI)
- Cox regression model with 8 risk factors.

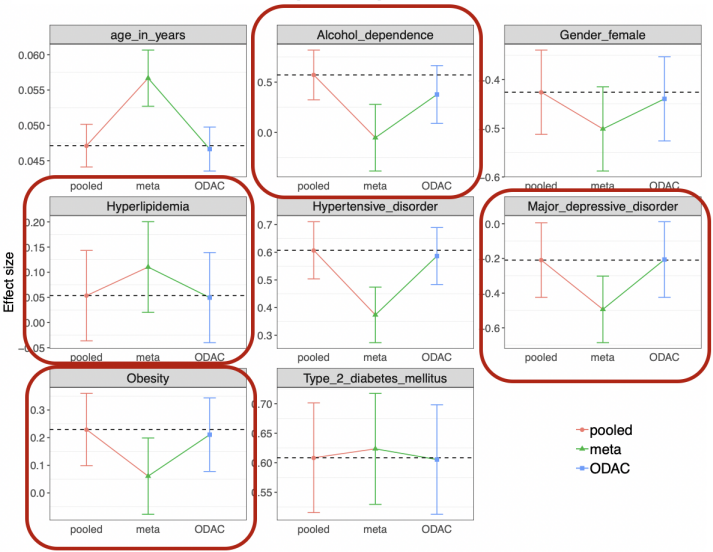
Table 1. Characteristics of the four claims datasets at OHDSI

Dataset	CCAE	MDCD	MDCR	Optum
Number of subjects	64,222	59,861	69,164	62,348
Median Age	43	35	71	47
% of Female	69.21	73.82	68.08	69.68
% of Congestive heart failure	0.70	3.06	7.58	2.61
% of Hypertensive disorder	20.81	31.80	57.70	32.96
% of Ischemic heart disease	1.70	3.82	10.27	4.10
% of Type 2 diabetes mellitus	7.49	14.63	21.83	12.71
% of Coronary arteriosclerosis	2.39	4.92	18.43	5.75
% of Renal failure syndrome	0.69	2.67	2.31	2.49
% of Transient cerebral ischemia	0.41	0.64	2.32	0.71
% of Hyperlipidemia	20.96	22.00	43.21	33.85
% of Obesity	7.15	16.54	6.71	9.62
% of Alcohol dependence	7.15	16.54	6.71	9.62
% of Major depressive disorder	4.17	3.55	3.16	3.34
% of Acute myocardial infarction	0.26	0.75	2.03	0.51
% of Stroke	0.24	0.73	1.75	0.58

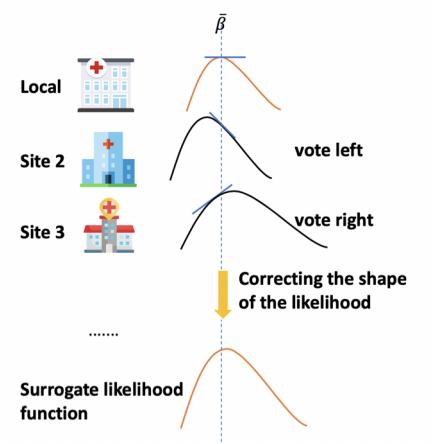
*The four claims datasets are CCAE (IBM MarketScan® Commercial), MDCD (IBM MarketScan® Medicaid), MDCR (IBM MarketScan® Medicare) and Optum (Optum© De-Identified Clinformatics).

Results (AMI)

- **Risk factors for acute myocardial infarction (AMI):**
  gender
  age
  obesity
  alcohol dependence
  hypertensive
  major depressive disorder
  type 2 diabetes mellitus
  hyperlipidemia



Intuitions: a nice interpretation of surrogate likelihood



- Meta-analysis requires point estimate and standard error.
- In ODAL/ODAC, slopes and curvatures help to correct the shape of local likelihood function.

Recap: definition of surrogate likelihood

The surrogate likelihood (SL) approach

- For an initial value $\bar{\beta}$,

$$L(\beta) = L(\bar{\beta}) + \nabla L(\bar{\beta})^T(\beta - \bar{\beta}) + \sum_{t=2}^{\infty} \frac{1}{t!} \nabla^t L(\bar{\beta})(\beta - \bar{\beta})^{\otimes t}$$

$$L_1(\beta) = L_1(\bar{\beta}) + \nabla L_1(\bar{\beta})^T(\beta - \bar{\beta}) + \sum_{t=2}^{\infty} \frac{1}{t!} \nabla^t L_1(\bar{\beta})(\beta - \bar{\beta})^{\otimes t}$$

$$\sum_{t=2}^{\infty} \frac{1}{t!} \nabla^t L_1(\bar{\beta})(\beta - \bar{\beta})^{\otimes t} = L_1(\beta) - L_1(\bar{\beta}) - \nabla L_1(\bar{\beta})^T(\beta - \bar{\beta})$$

First-order SL function

$$\tilde{L}^1(\beta) = L_1(\beta) + \{\nabla L(\bar{\beta}) - \nabla L_1(\bar{\beta})\}^T \beta$$

$$\nabla L(\bar{\beta}) = \frac{1}{K} \sum \nabla L_j(\bar{\beta}); \quad \nabla L_j(\bar{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \{y_{ij} - \text{expit}(x_{ij}^T \bar{\beta})\} x_{ij}$$

Other exercises using surrogate likelihood - Poisson regression

ODAP

One-shot **D**istributed **A**lgorithm for Quasi-**P**oisson regression

ODAP: One-Shot Distributed Algorithm for Performing Quasi-Poisson Regression

Mackenzie J. Edmondson[a], Chongliang Luo[a], Zhaoyi Chen[b], Jiang Bian[b], Yong Chen[a]

a. Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA
b. University of Florida, Gainesville, FL, USA

Other exercises using surrogate likelihood - Hurdle regression

ODAH

One-shot **D**istributed **A**lgorithm for **H**urdle Regression

Distributed Learning from Electronic Health Records Across Multiple Sites for Zero-Inflated Count Outcomes

Mackenzie J. Edmondson[a], Chongliang Luo[a], Rui Duan[a], Mitchell Maltenfort[b], Justine Shults[a], Patrick B. Ryan[c], Christopher B. Forrest[b], Yong Chen[a]

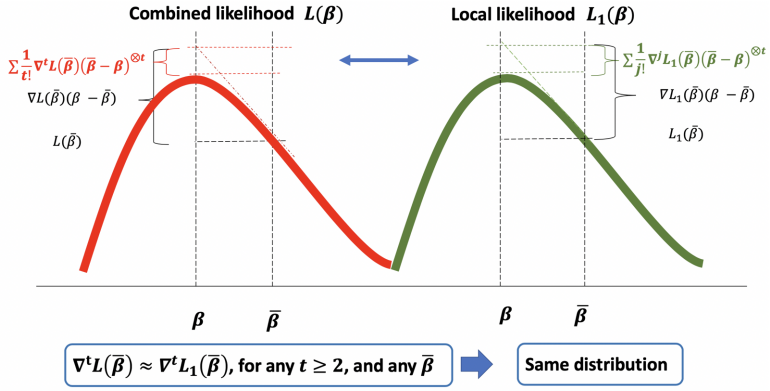
a. Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA
b. Department of Pediatrics, Children's Hospital of Philadelphia, Philadelphia, PA, USA
c. Janssen Research and Development, Titusville, NJ, USA

A challenge in real-world data

Heterogeneity in distributed research network

- Most of existing distributed algorithms ignored the intrinsic nature of between-site heterogeneity
- In real-world data, such heterogeneity cannot be ignored

Surrogate likelihood approach assumes homogeneous data



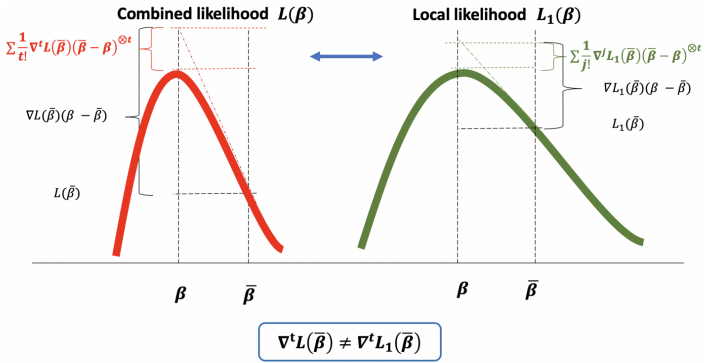
Heterogeneity assumption

- Data in the j-th site follows

$$Y_{ij} \sim f(y; \boldsymbol{\beta}, \boldsymbol{\gamma}_j)$$

- $\boldsymbol{\beta}$ is the ***parameter of interest***
- $\boldsymbol{\gamma}_j$ is the ***site-specific nuisance parameter***--- *allow site to be a covariate variable, allow interaction terms between site and other covariates.*

Challenges



- Local site cannot provide any information about nuisance parameters in other sites.

Heterogeneity-aware PDA algorithms

Biometrika (2015), **99**, 1, *pp.* 1–17
© 2015 Biometrika Trust
Printed in Great Britain

Advance Access publication on 31 July 2015

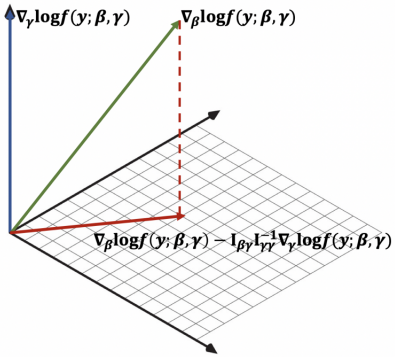
Heterogeneity-aware and communication-efficient distributed statistical inference

BY RUI DUAN
Department of Biostatistics, Harvard University, Boston, Massachusetts 02115, U.S.A.
rduan@hsph.harvard.edu

YANG NING
Department of Statistics and Data Science, Cornell University, Ithaca, New York, 14853, U.S.A.
yn265@cornell.edu

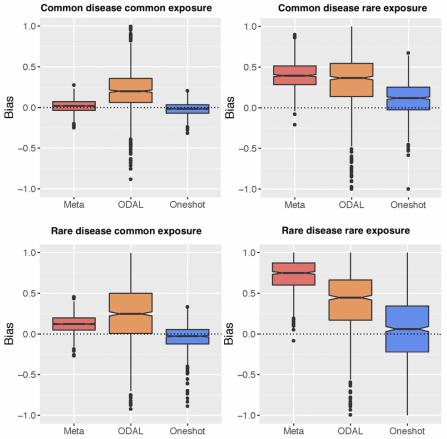
YONG CHEN
Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, Pennsylvania 19104, U.S.A.
ychen123@upenn.edu

arXiv:1912.09623 *Biometrika (in press)*

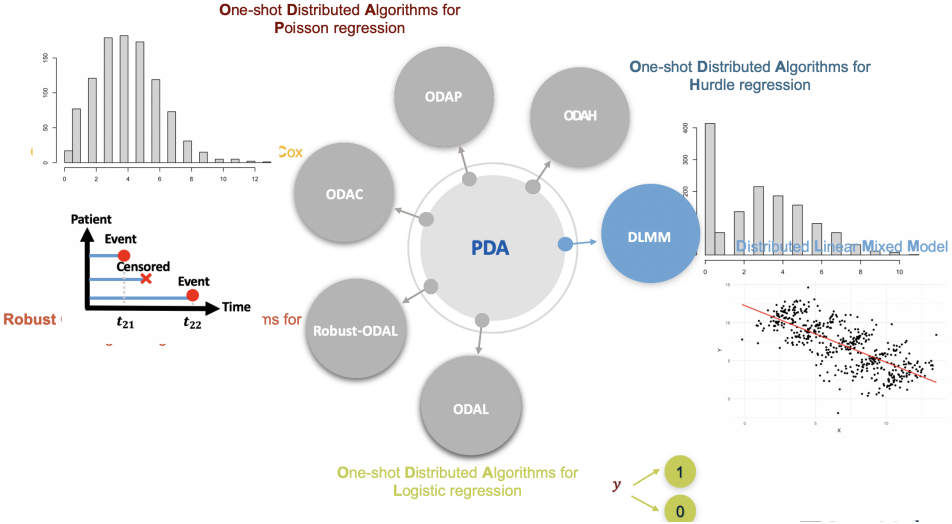


Conclusion

Desired Properties	Meta-analysis	Proposed estimator (T=1)	Proposed estimator (T=2)
Consistency	consistent	consistent	consistent
Distance to the Gold Standard Estimator	$\frac{C}{\sqrt{Kn}}$	$\preccurlyeq \frac{C}{n}$	$\preccurlyeq \frac{C}{n\sqrt{K}} + \frac{C}{n\sqrt{n}}$
Asymptotic Normality	asymptotic normal	asymptotic normal	asymptotic normal
Asymptotic Efficient	not efficient	efficient	efficient



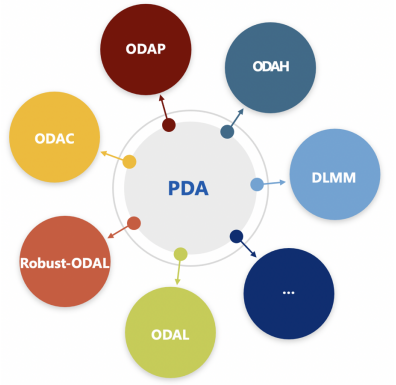
Duan, Rui, Yang Ning, and Yong Chen. "Heterogeneity-aware and communication-efficient distributed statistical inference." *arXiv preprint arXiv:1912.09623* (2019).


One-shot Distributed Algorithms for Poisson regression
One-shot Distributed Algorithms for Hurdle regression
ODAP
ODAH
ODAC
PDA
DLMM
Robust-ODAL
ODAL
Distributed Linear Mixed Model
Patient
Event
Censored
Event
Time
t_{21}
t_{22}
One-shot Distributed Algorithms for Logistic regression
y
1
0

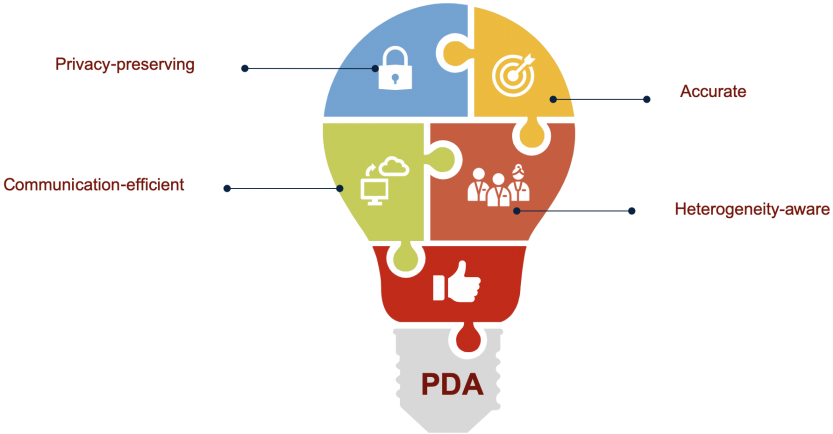
PDA as a collection of distributed algorithms

PDA : a solution for next generation data sharing

Privacy-preserving Distributed Algorithms



R package ‘pda’: four principles



Tutorial 3

Outline

Concluding thoughts

- Due to financial incentives and operational efficiencies, EHR will remain the dominant mode of clinical/administrative documentation of health encounters
- This creates a vast research resource but also requires knowledge of its complexities to use appropriately
- A key component of data science is expert knowledge about data sources
- To effectively use EHR data we (statisticians) must be willing to learn about where these data come from and how they are used clinically/administratively
- We wouldn't analyze observational data without reading the protocol!

Recommendations

- Engaging with clinicians, coders, informaticians allows us to
  - Understand data quality
  - Make smart choices about when and how EHR data can be used
  - Identify appropriate methods to mitigate limitations
  - Develop new statistical methods to fill gaps in available methodology
- EHR data can be messy but don't despair!
- Staying engaged in the research process from data extraction through analysis, interpretation, and reporting of results ensures higher quality research and gives us a seat at the table to help improve processes for the future

Acknowledgments

ENAR