Leveraging missing data and measurement error approaches in propensity score-based analyses of real-world data

Rebecca Hubbard

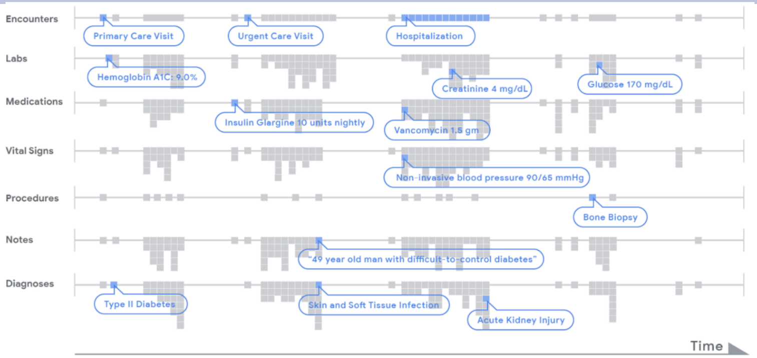
Kylie Getz, Dan Vader, Kristin Linn, Ronac Mamtani

DEPARTMENT of BIOSTATISTICS EPIDEMIOLOGY & INFORMATICS

Perelman School of Medicine
UNIVERSITY of PENNSYLVANIA

August 21, 2023
Sentinel Innovation and Methods Seminar

Using EHR and claims data for research



- Longitudinal observational study with no specification of timing of visits, data elements to be collected at each visit, or definition of data elements
- Unlike data from a designed study, the data capture process in EHR-based studies is entirely outside the control of the researcher
- The visit process often violates assumptions of standard statistical approaches

Missing data in EHR

- Because EHR data are not collected according to a research protocol they will often be missing variables of interest
- While missing data are virtually ubiquitous in EHR-based studies, a critical first step to dealing with missingness is consideration of what constitutes a “complete” record
- Unlike a designed observational study, there is no prior specification of which data elements should be collected for a patient or when they should be collected

Moving beyond clinical trials for comparative effectiveness

- Clinical trials are considered gold standard for estimating treatment efficacy
- But may have limited external validity, especially when some patient populations are underrepresented
- For instance in oncology trials, racial/ethnic minorities and poor prognosis patients are substantially underrepresented
- However, once treatments are approved they are prescribed broadly
- EHR data from patients receiving these treatments in routine care can help to bridge the gap between the observed **efficacy** of the treatment in the trial and its **effectiveness** in routine practice
- Confounder control is key to this endeavour but in practice many confounders will be sporadically captured

Sources of bias in EHR-based research

Information bias

- Measurement error and misclassification are commonplace
- Data elements lack harmonized, common definitions
- Usage of clinical terminology may not coincide with research usage
- Codes may be coarse/non-specific

Confounding

- Limited information on behavioral risk factors and social determinants of health
  - May be available in narrative text notes but difficult to extract and inconsistently collected across patients
- Information on symptoms, family history inconsistently available
- Information on severity of disease lacking

CER requires that we address both information bias and confounding (as well as information bias affecting confounders)

Objectives

Compare and contrast standard and new(er) approaches to handling missing data in EHR-based comparative effectiveness analyses

1. Motivating example
2. Multiple imputation using machine learning-based imputation
3. Multiple imputation vs propensity-score calibration

Overview

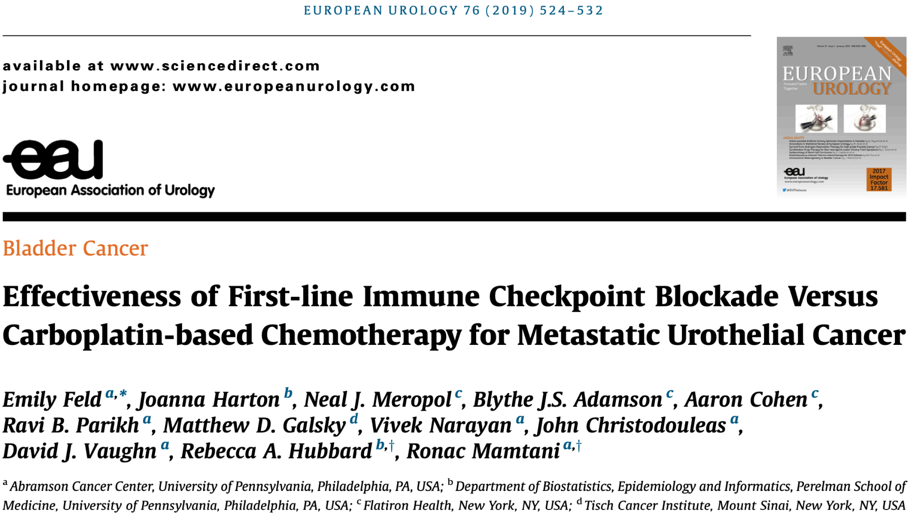
Motivating example: Immunotherapy for treatment of advanced urothelial cancer

Multiple imputation using machine learning-based imputation

Multiple imputation vs propensity-score calibration

Conclusions

Clinical question: Can immune checkpoint inhibitors improve survival in patients with metastatic urothelial cancer (mUC)?



EUROPEAN UROLOGY 76 (2019) 524–532

available at www.sciencedirect.com
journal homepage: www.europeanurology.com

European Association of Urology

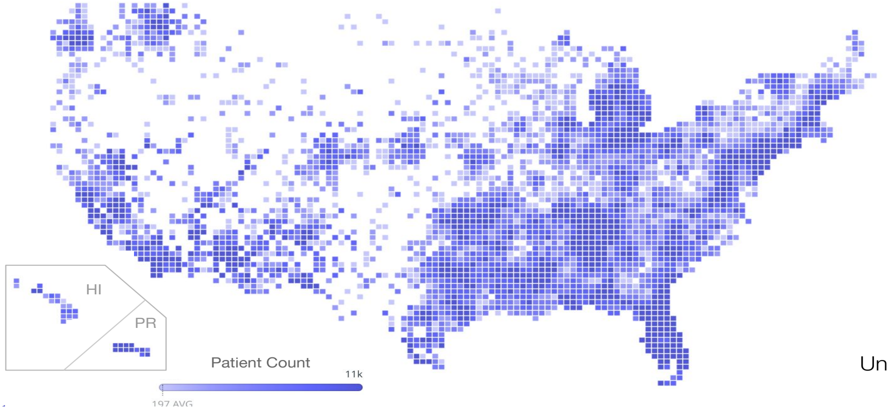
Bladder Cancer

Effectiveness of First-line Immune Checkpoint Blockade Versus Carboplatin-based Chemotherapy for Metastatic Urothelial Cancer

Emily Feld [a,*], *Joanna Harton* [b], *Neal J. Meropol* [c], *Blythe J.S. Adamson* [c], *Aaron Cohen* [c], *Ravi B. Parikh* [a], *Matthew D. Galsky* [d], *Vivek Narayan* [a], *John Christodouleas* [a], *David J. Vaughn* [a], *Rebecca A. Hubbard* [b,†], *Ronac Mamtani* [a,†]

[a] *Abramson Cancer Center, University of Pennsylvania, Philadelphia, PA, USA;* [b] *Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA;* [c] *Flatiron Health, New York, NY, USA;* [d] *Tisch Cancer Institute, Mount Sinai, New York, NY, USA*

Flatiron health EHR-based network



2M+
Active Patients

2,500
Clinicians

265
Cancer Clinics

800
Unique Sites of Care

Immune checkpoint inhibitors for treatment of mUC

- Stage IV urothelial cancer (mUC) patients treated at an oncology center contributing data to the Flatiron health oncology EHR database
- Survival in this population is poor with one-year survival of about 40%
- Immune checkpoint inhibitors offer potential for survival benefit but had not been evaluated head to head with chemotherapy in real-world settings

Results

	Immunotherapy	Chemotherapy
	(N=487)	(N=1530)
	Estimate, 95% CI	Estimate, 95% CI
12-mo OS	40% (34–45%)	46% (43–49%)
36-month OS	28% (22–35%)	13% (11–16%)
Hazard ratio $\leq$12 mo	1.37 (1.15–1.62)	1.00 (reference)
Hazard ratio $>$12 mo	0.50 (0.30–0.85)	1.00 (reference)

- Immunotherapy associated with poorer early but better late outcomes
- Used IPTW to account for confounding
- Combined with MI via MICE to address missing data

- But methodological concerns persist
- **A key confounder, ECOG performance status (PS) missing for 35% of patients**
- Clinical investigators raised concern about possible MNAR missingness in PS

Methods questions opened by this study

- Are there better approaches to MI in the context of EHR data?
    - Appropriate for high levels of missingness?
    - Able to accommodate complex patterns of missingness in confounders?
    - Able to reduce bias even under MNAR missingness?
- Are there alternative approaches that outperform MI in the context of EHR-based CER analyses using IPTW?
    - Capitalize on dimension reduction of the propensity score?
    - Computationally efficient in large EHR samples?

Overview

Motivating example: Immunotherapy for treatment of advanced urothelial cancer

Multiple imputation using machine learning-based imputation

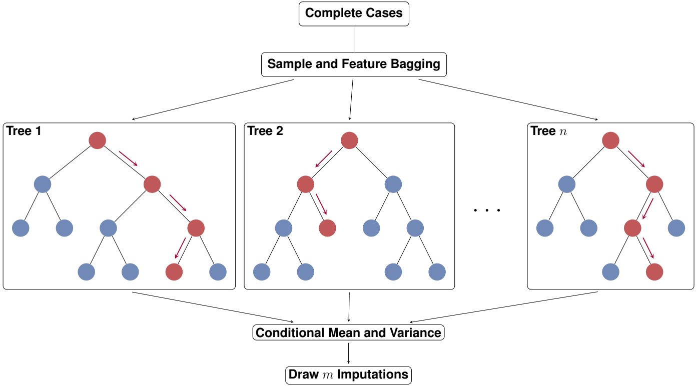
Multiple imputation vs propensity-score calibration

Conclusions

Multiple imputation via chained equations (MICE)

- Common MI approach in which a separate model is specified for each variable with missing observations
- Missing data in each variable are sequentially filled in and subsequently used in regression models for other variables; process iterated until convergence
- Convenient for use with EHR data because regression models for each variable can allow for different variable types and can include different predictors
- Limitations
    - The process of model specification can be quite laborious, especially if derived variables and interactions are involved
    - Parametric models may provide a poor fit to complex relationships in the data
    - Computationally intensive for large EHR samples

Multiple imputation with random forests

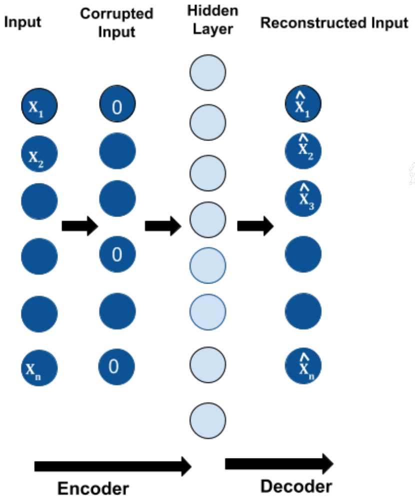


- MI using random forests (RF) proposed to relax parametric assumptions of traditional MICE implementations
- RF fits trees to bootstrap samples of complete cases
- Imputation based on conditional mean based on fitted RF

- Allows for arbitrary interactions and non-linearity
- Previous research found RF MICE reduced bias relative to parametric MICE when parametric model failed to capture interactions (Shah et al 2014)

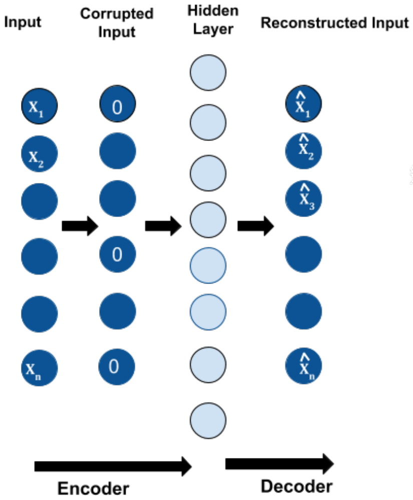
Imputation via denoising autoencoders (DAE)

- Denosing Autoencoder (DAE): neural network that learns an encoded representation of input data by attempting to predict the input data from a corrupted version of itself
- Encode the input data into an equal or higher-dimensional representation (overcomplete)
- Inputs corrupted to prevent learning identity function
- Hidden layers $\underline{h}=g(\mathbf{W}\underline{x_i}+\underline{b})$, where $\underline{x_i}$ = input data, $\mathbf{W}$= weight matrix, $\underline{b}$= bias term, g = nonlinear activation function
- Parameters of model are estimated based on minimizing MSE



Multiple imputation with DAE

- Past research (Beaulieu-Jones and Moore 2017; Gondara and Wang 2018) found that DAE outperformed other imputation approaches in terms of imputation accuracy
- Limited evaluation in epidemiologic settings (small *N* and *p*), MI, and performance in terms of bias and efficiency



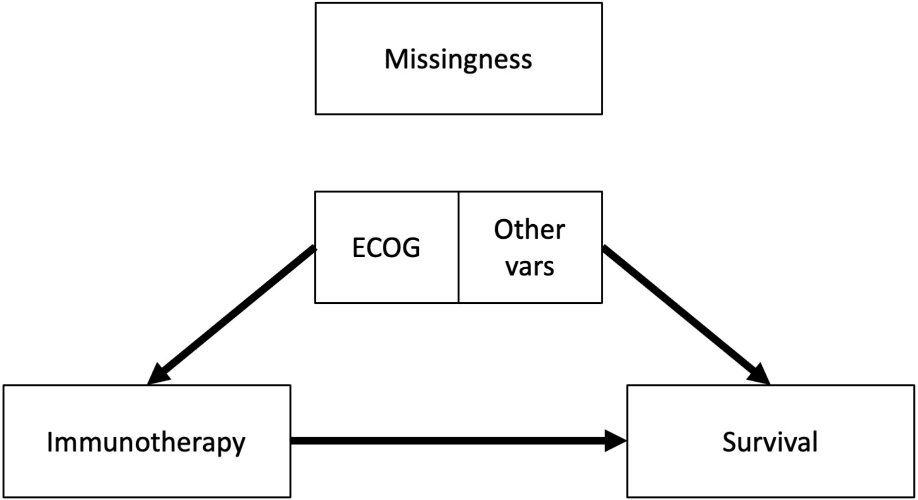
Plasmode simulation

“Plasmodes are data sets that are generated by natural biologic processes, under experimental conditions that allow some aspect of the truth to be known.” (Vaughan et al. 2009)

- Proposed for simulation studies of EHR data to preserve the complex relationship among variables (Franklin et al. 2014)
- Using complete data, estimate baseline hazard for overall survival and censoring, and covariate effects on overall survival
- Sample with replacement from complete data
- Simulate outcome and censoring data using inverse transform method based on observed baseline hazards and confounder effects, with treatment effect fixed at desired value
- Introduce missing data according to missingness mechanism of interest

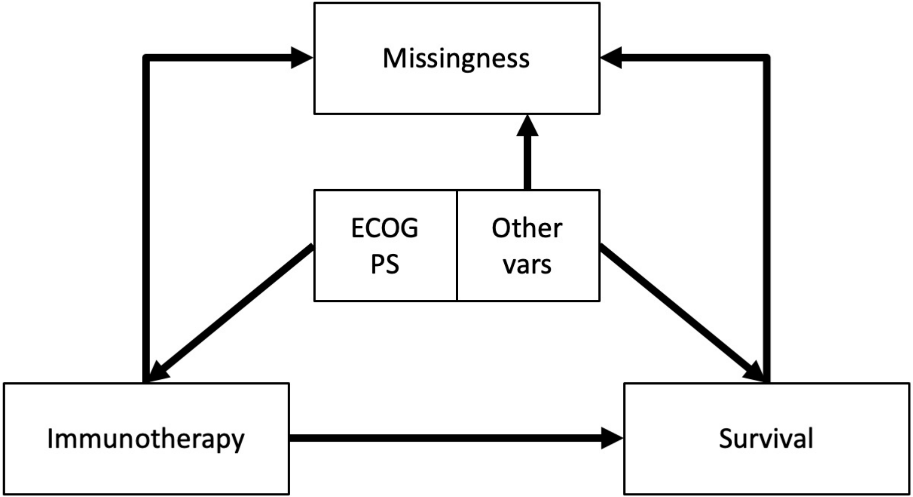
Missingness mechanisms

MCAR

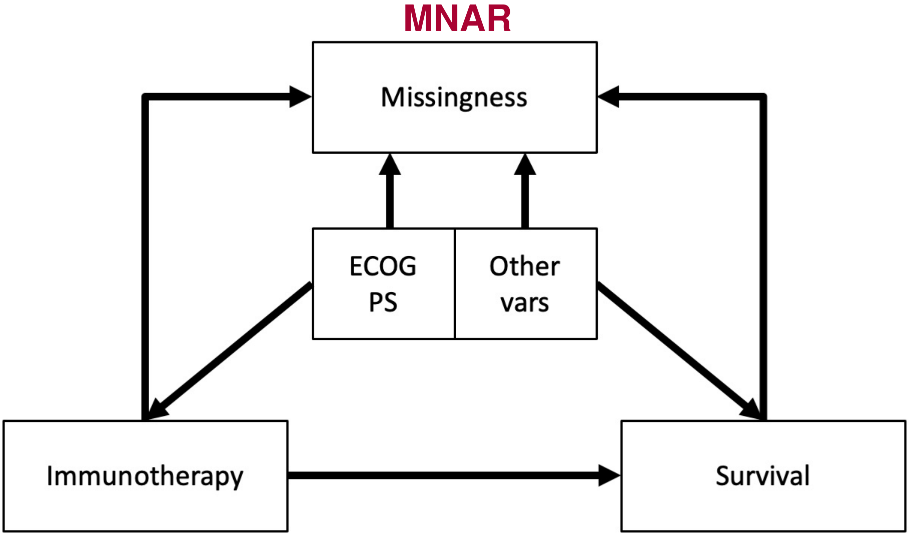


Missingness mechanisms

MAR



Missingness mechanisms



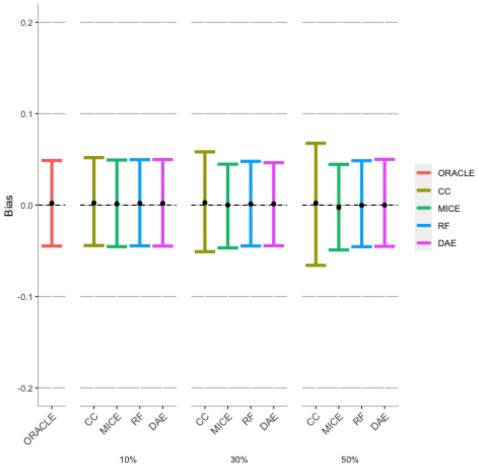
Overview of simulation study design

Objective: Estimate adjusted hazard ratio describing association between immunotherapy and overall survival using alternative imputation approaches to address missingness in confounder variables

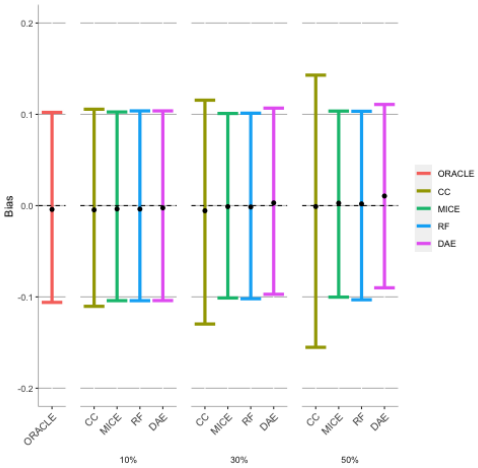
- Generate complete data sets using plasmode simulation applied to Flatiron mUC cohort
- Introduce missingness varying proportion missing and missingness mechanism
- Missingness in ECOG PS varied across MCAR, MAR, MNAR
- Missingness in other confounders assumed MCAR
- Estimate association using complete case, MICE, MI RF, or DAE
- Compute bias, SE, CI coverage probability based on 1000 simulation iterations

Results: MCAR

ECOG PS

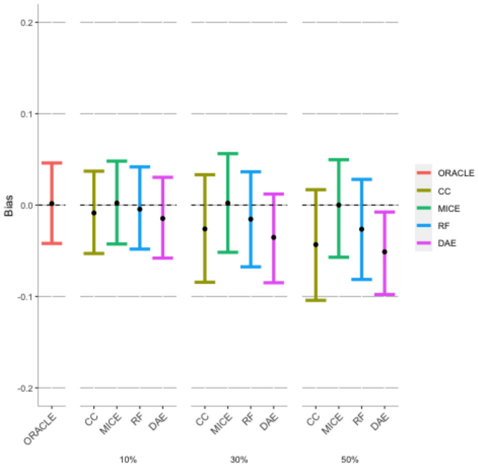


Immunotherapy

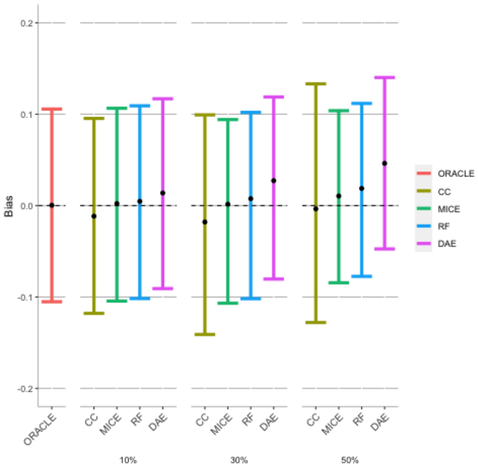


Results: MAR

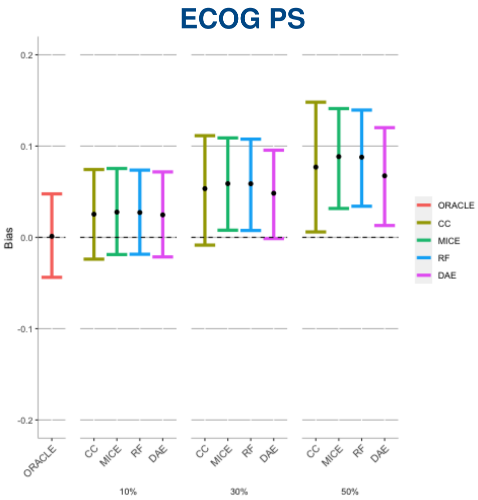
ECOG PS

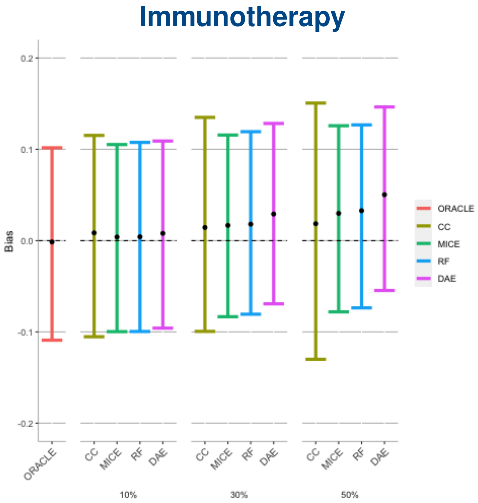


Immunotherapy



Results: MNAR





Lessons learned

- No advantage of machine learning methods in the setting of an EHR-based CER study
- RF and DAE may overfit the data leading to poor confounder control
- Use of more flexible imputation approaches does not mitigate bias induced by MNAR missingness
- Caveats
  - Simulation-based results depend on details of the simulation
  - There are infinitely many kinds of MNAR missingness, we have evaluated only one
  - Results in other contexts may differ
- Important to evaluate missing data methods in terms of performance of parameter estimates of interest (not imputation accuracy)

Overview

Motivating example: Immunotherapy for treatment of advanced urothelial cancer

Multiple imputation using machine learning-based imputation

Multiple imputation vs propensity-score calibration

Conclusions

Missing data or measurement error?

- In EHR research, lack of confounder data most often conceived of as a missing data problem
- But rich “proxy” data available in the form of diagnosis codes, prescriptions, etc
- Harton et al. (2021) compared alternative regression calibration approaches applied to the case of an error-prone propensity score
- However, propensity score adjustment in multivariable models is limited by need to correctly specify the propensity score/outcome relationship
- Also did not include head-to-head comparison of missing data and regression calibration approaches

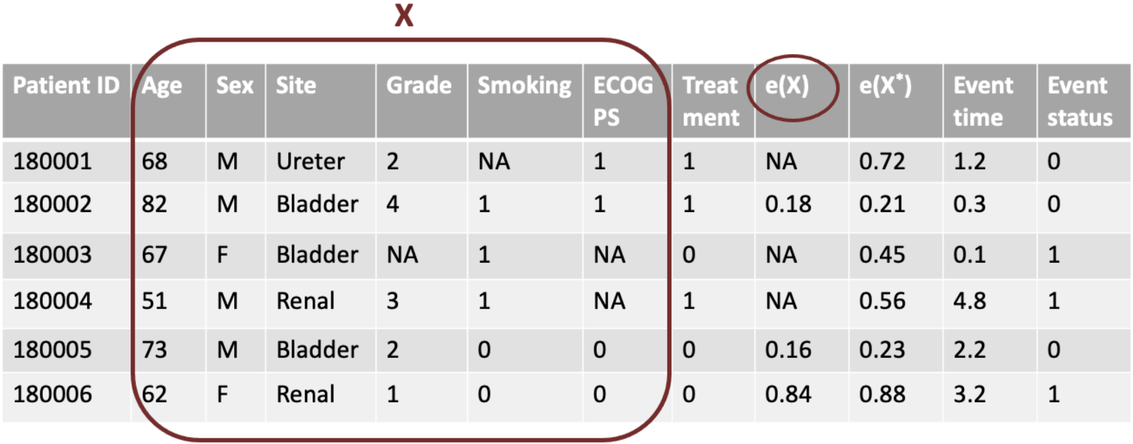
The propensity score

- The propensity score is a fundamental tool for confounder control, frequently used in CER (Rosenbaum and Rubin 1983)
- **Propensity score** $e(x)$ defined as the probability that a person with observed covariates $X = x$ is in exposure group $Z = 1$

$$e(x) = P(Z = 1 | X = x)$$

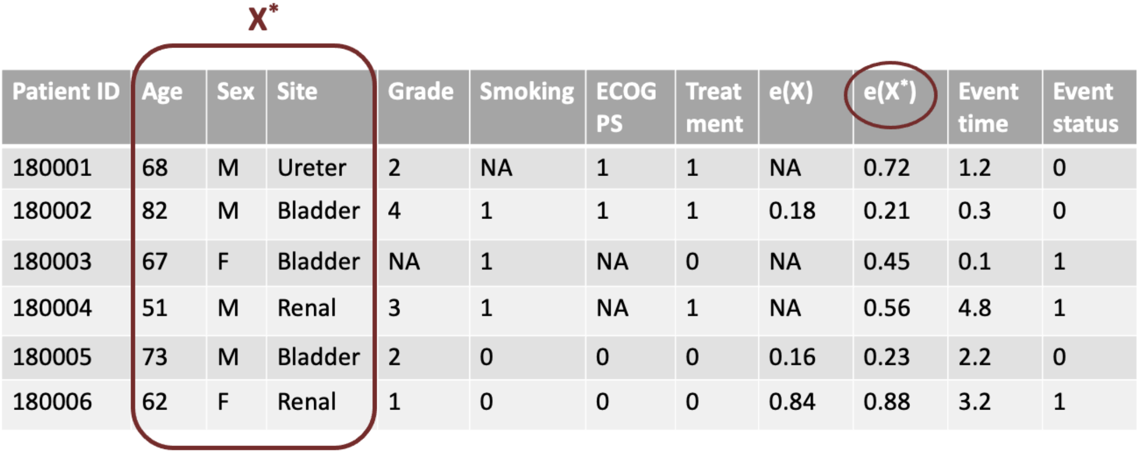
- Scalar function of X that summarizes information required to balance the covariate distribution between exposure groups
- Can be estimated using supervised learning approach of choice and incorporated in subsequent analyses via regression adjustment, matching, stratification, or weighting

Propensity scores and missing data



Patient ID	Age	Sex	Site	Grade	Smoking	ECOG PS	Treatment	$e(X)$	$e(X^*)$	Event time	Event status
180001	68	M	Ureter	2	NA	1	1	NA	0.72	1.2	0
180002	82	M	Bladder	4	1	1	1	0.18	0.21	0.3	0
180003	67	F	Bladder	NA	1	NA	0	NA	0.45	0.1	1
180004	51	M	Renal	3	1	NA	1	NA	0.56	4.8	1
180005	73	M	Bladder	2	0	0	0	0.16	0.23	2.2	0
180006	62	F	Renal	1	0	0	0	0.84	0.88	3.2	1

Propensity scores and missing data



X^*

Patient ID	Age	Sex	Site	Grade	Smoking	ECOG PS	Treatment	e(X)	$e(X^*)$	Event time	Event status
180001	68	M	Ureter	2	NA	1	1	NA	0.72	1.2	0
180002	82	M	Bladder	4	1	1	1	0.18	0.21	0.3	0
180003	67	F	Bladder	NA	1	NA	0	NA	0.45	0.1	1
180004	51	M	Renal	3	1	NA	1	NA	0.56	4.8	1
180005	73	M	Bladder	2	0	0	0	0.16	0.23	2.2	0
180006	62	F	Renal	1	0	0	0	0.84	0.88	3.2	1

Propensity score calibration (Stürmer et al 2007)

- Estimate gold standard propensity scores $\hat{e}(X)$
- Estimate error-prone propensity scores $\hat{e}(X^*)$
- Fit calibrated error-prone propensity score model

  $$E(e(X)|Z, e(X^*)) = \alpha + \beta Z + \gamma e(X^*)$$

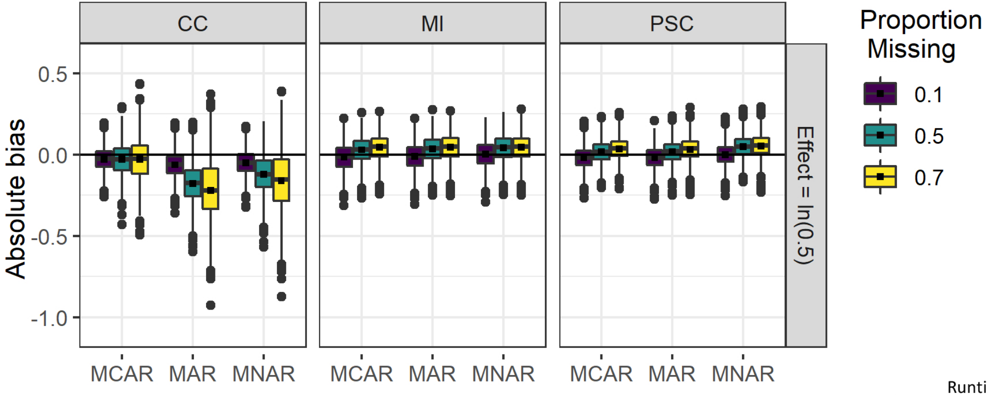
  by regressing gold-standard propensity scores on treatment and the error prone propensity scores
- Generate a single imputation of $\hat{e}(X)$ based on the calibration model
- Fit IPTW outcome model

Overview of simulation study design

Objective: Estimate IPTW hazard ratio describing association between immunotherapy and overall survival using PSC or MI

- Generate complete data sets using plasmode simulation applied to Flatiron mUC cohort
- Introduce missingness varying proportion missing and missingness mechanism
- Missingness in ECOG PS varied across MCAR, MAR, MNAR
- Missingness in other confounders assumed MCAR
- Three variables (gender, surgery, age) assumed complete across all patients
- Estimate association using complete case (CC), MI and PSC
- Compute bias, SE, CI coverage probability based on 1000 simulation iterations

Results



	Runtime (seconds)		
% Missing	CC	MI	PSC
0.1	11	604	14
0.5	3	562	13
0.7	1	527	13

Lessons learned

- Both MI and PSC performed well in terms of controlling bias
- PSC substantially more computationally efficient
- Performance of PSC degrades as more variables have missing data and must be excluded from the error-prone PS but works well when missingness is concentrated in a few variables

Overview

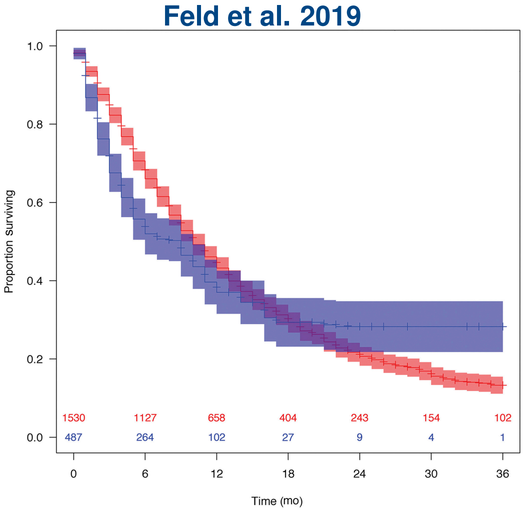
Motivating example: Immunotherapy for treatment of advanced urothelial cancer

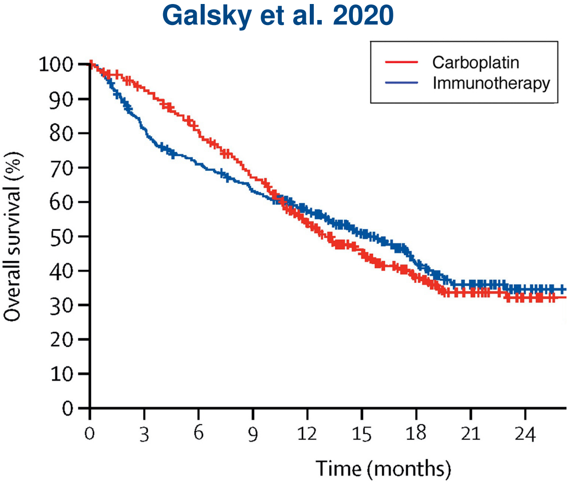
Multiple imputation using machine learning-based imputation

Multiple imputation vs propensity-score calibration

Conclusions

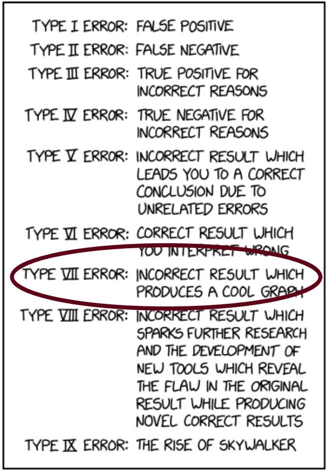
Real-world evidence can complement RCTs





Conclusions

- EHR data can facilitate treatment effectiveness evaluations not possible in trials
- Methods for CER have focused on issues arising due to confounding; information bias is also a major concern
- Novel approaches such as modern machine learning methods can be used to address these issues but should not be considered a panacea
- Practical methods investigations are needed to inform best research practices



TYPE I ERROR: FALSE POSITIVE

TYPE II ERROR: FALSE NEGATIVE

TYPE III ERROR: TRUE POSITIVE FOR INCORRECT REASONS

TYPE IV ERROR: TRUE NEGATIVE FOR INCORRECT REASONS

TYPE V ERROR: INCORRECT RESULT WHICH LEADS YOU TO A CORRECT CONCLUSION DUE TO UNRELATED ERRORS

TYPE VI ERROR: CORRECT RESULT WHICH YOU INTERPRET WRONG

TYPE VII ERROR: INCORRECT RESULT WHICH PRODUCES A COOL GRAPH

TYPE VIII ERROR: INCORRECT RESULT WHICH SPARKS FURTHER RESEARCH AND THE DEVELOPMENT OF NEW TOOLS WHICH REVEAL THE FLAW IN THE ORIGINAL RESULT WHILE PRODUCING NOVEL CORRECT RESULTS

TYPE IX ERROR: THE RISE OF SKYWALKER

`https://xkcd.com/2303/`

Acknowledgments

- Kylie Getz
- Daniel Vader
- Joanna Harton
- Kristin Linn
- Ronac Mamtani

Research reported in this presentation was supported by the National Institutes of Health under award number R21CA227613 and R21AG075574. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Beaulieu-Jones BK and Moore JH. Missing data imputation in the electronic health record using deeply learned autoencoders. Pacific Symposium on Biocomputing. 2017; 207-218.

Feld E et al. Effectiveness of first-line immune checkpoint blockade versus carboplatin-based chemotherapy for metastatic urothelial cancer. European Urology. 2019;76(4):524-32.

Franklin JM et al. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. Computational Statistics & Data Analysis. 2014;72:219-26.

Galsky MD et al. Atezolizumab with or without chemotherapy in metastatic urothelial cancer (IMvigor130): a multicentre, randomised, placebo-controlled phase 3 trial. The Lancet. 2020 ;395(10236):1547-57.

Getz K et al. Performance of multiple imputation using modern machine learning methods in electronic health records data. 2023; Epidemiology.34(2):206-215.

Gondara L and Wang K. MIDA: Multiple imputation using denoising autoencoders. Pacific-Asia Conference on Knowledge Discovery and Data Mining. 2018; 260-272.

Harton J et al. Bias reduction methods for propensity scores estimated from error-prone EHR-derived covariates. Health Services and Outcomes Research Methodology. 2021;21(2):169-87.

Shah AD et al. Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. American Journal of Epidemiology. 2014;179(6):764-74.

Stürmer T et al. Performance of propensity score calibration—a simulation study. American Journal of Epidemiology. 2007;165(10):1110-8.

Vader DY et al. Inverse probability of treatment weighting and missingness in confounder data in EHR-based analyses: a comparison of three missing data approaches using plasmode simulation. Epidemiology. 2023; In press.

Vaughan LK et al. The use of plasmodes as a supplement to simulations: a simple example evaluating individual admixture estimation methodologies. Computational statistics data analysis. 2009;53(5):1755-66.

Rebecca Hubbard

rhubb@upenn.edu
https://www.med.upenn.edu/ehr-stats/

DEPARTMENT of
BIOSTATISTICS
EPIDEMIOLOGY &
INFORMATICS

Perelman
School of Medicine
UNIVERSITY of PENNSYLVANIA