Statistically Speaking Lecture Series

Sponsored by the Biostatistics Collaboration Center

Considerations when Leveraging Electronic Health Records for Causal Inference
A "Create-Your-Own-Data" Adventure

Lucia C. Petito, PhD

Assistant Professor

March 3, 2022

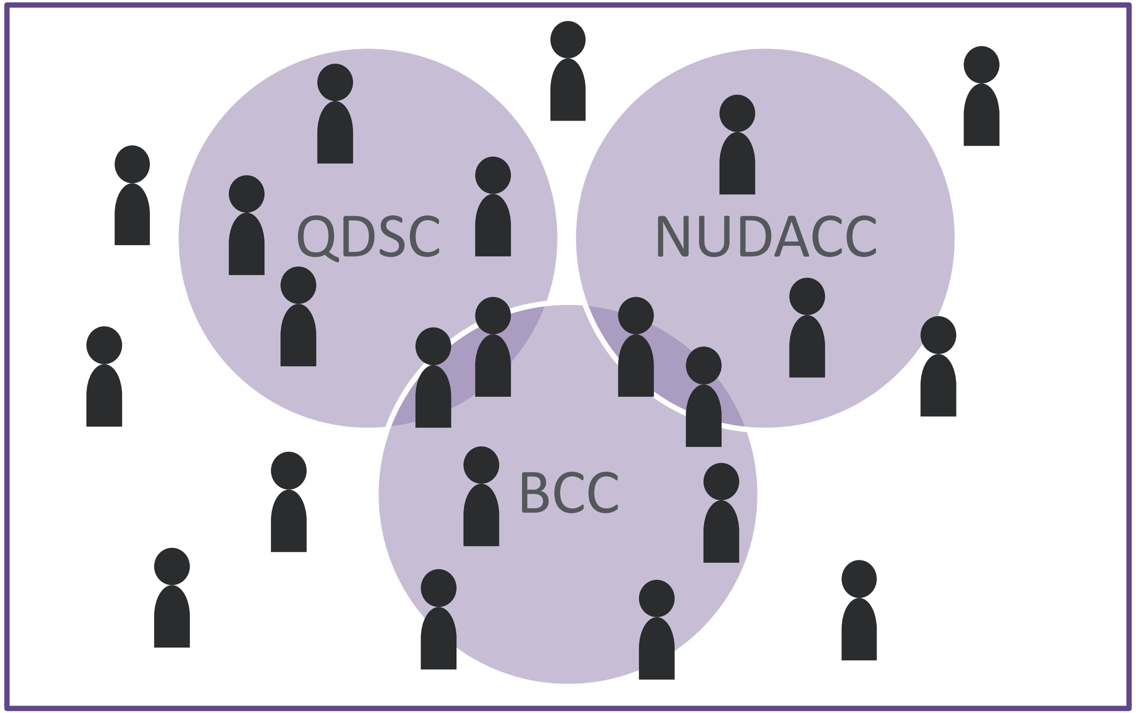
Via Zoom

Northwestern Medicine®

Biostatistics at NU

Overview

Division of Biostatistics (Chief: Denise Scholtens),
Department of Preventive Medicine (Chair: Donald Lloyd-Jones)



Northwestern Medicine®

Biostatistics Centers and Cores

Northwestern Medicine | NUCATS Clinical and Translational Sciences Institute

Robert H. Lurie Comprehensive Cancer Center of Northwestern University

Stanley Manne Children's Research Institute

Ann & Robert H. Lurie Children's Hospital of Chicago

Shirley Ryan Abilitylab

Biostatistics Collaboration Center (BCC)

- Supports **non-cancer** research at NU
- Initial 1-2 hour consultation subsidized by FSM Research Office
- Grant, Hourly
- https://www.feinberg.northwestern.edu/sites/bcc/

Quantitative Data Sciences Core (QDSC)

- Supports **cancer-related** research at NU
- Free to Lurie Cancer Center (LCC) members
- Grant
- https://www.cancer.northwestern.edu/research/shared-resources/quantitative-data-sciences.html

Northwestern University Data Analysis and Coordinating Center (NUDACC)

- Prospective, large **multicenter research**
- Comprehensive support (e.g., clinical monitoring, data analysis, project management)
- Grant
- https://www.feinberg.northwestern.edu/sites/nudacc/

Brief Overview

1. Overview of **causal inference** in observational data

2. How **target trials** can improve comparative effectiveness studies done in Electronic Health Record (EHR) data

3. **An example**: Effectiveness of DMARD as 2nd line treatment after methotrexate in rheumatoid arthritis patients

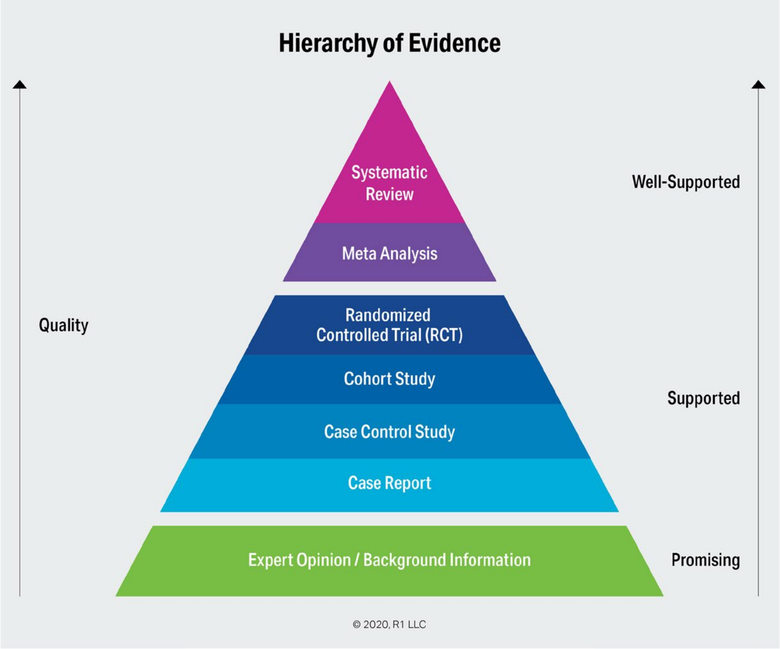
Causal Inference

Not all questions in medicine are causal, but many are.

- Examples of **causal** questions
  - Does liver transplant surgery increase the life expectancy of individuals with cirrhosis?
  - Does receiving the SARS-COV-2 vaccine reduce the incidence of covid-related hospitalization in adults?

- Examples of **non-causal** questions
  - How many people in the U.S. have early-onset dementia?
  - Does obesity in adulthood cause mental health problems in teenage years?

Causal Inference



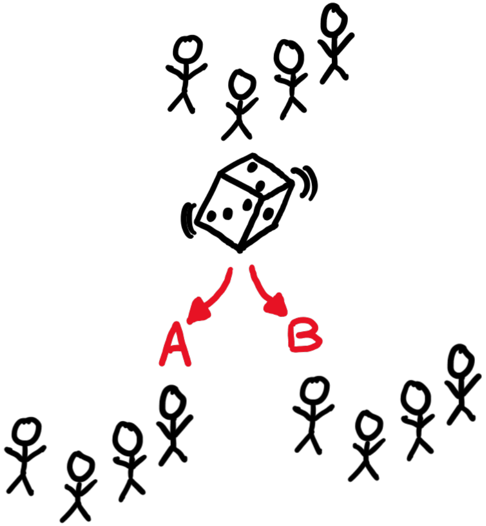
© 2020, R1 LLC

Many of these questions can be answered using a well-designed randomized controlled trial (RCT)!

Why are RCTs so great?

- **An imaginary perfect (vaccine) RCT**
  - Recruit n participants; randomize 1:1 to vaccine or placebo
  - Follow for a set period of time (e.g. 1 year); record outcome
  - Analyze according to the *Intention-to-Treat Principle*
    - Participants are analyzed according to the treatment they were assigned to

- **Causal inference relies on three assumptions:**
  - Exchangeability
  - Positivity
  - Consistency

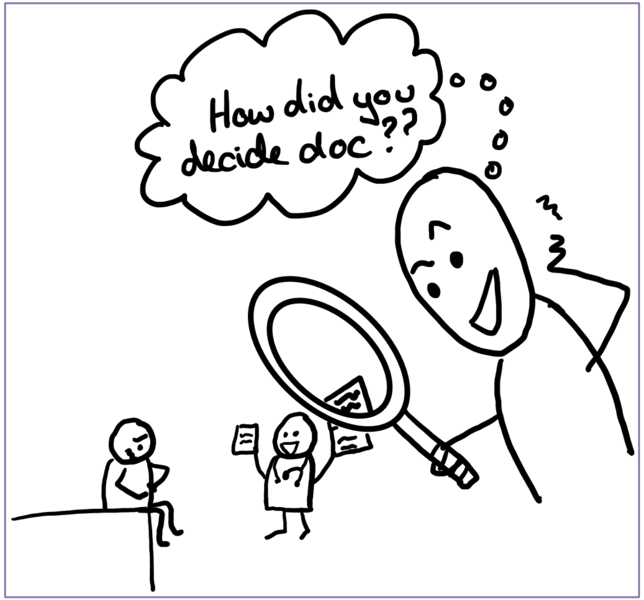


@EpiEllie

What is Exchangeability?

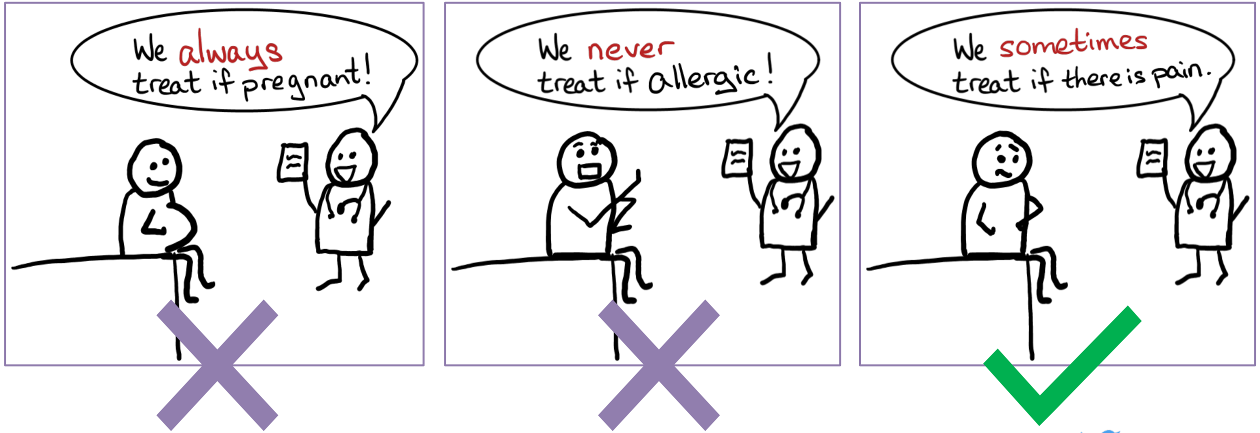
- No **unmeasured confounding**
  - All common causes of the treatment and outcome are known and measured in the data

- No **selection bias**
  - We have not conditioned or restricted on a variable that is a common effect of the exposure and outcome (or outcome cause)



@EpiEllie

What is Positivity?

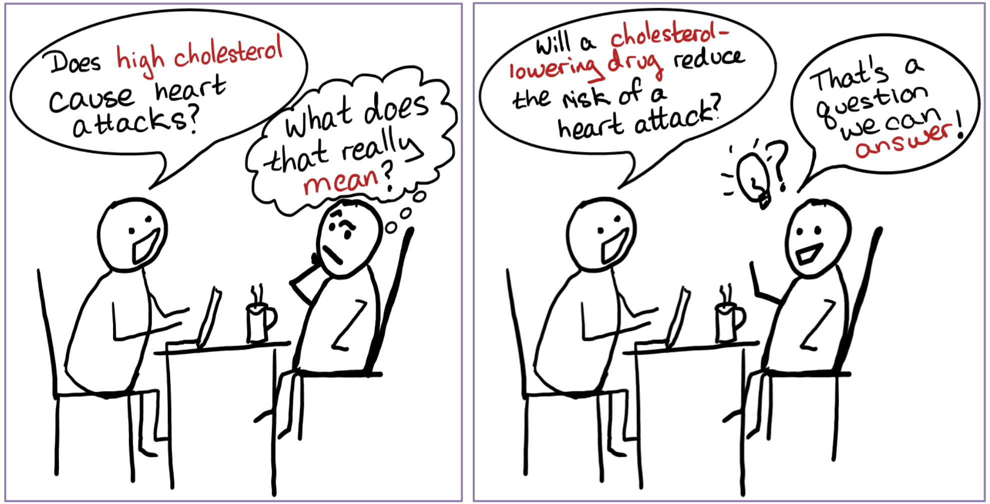

We always treat if pregnant!
We never treat if allergic!
We sometimes treat if there is pain.

@EpiEllie

- Non-zero probability of all levels of treatment for all types of individuals in our population

What is Consistency?

- Clear specification of treatment levels, corresponding to:
  - Well-defined interventions
  - Well-defined causal questions


Does high cholesterol cause heart attacks?
What does that really mean?
Will a cholesterol-lowering drug reduce the risk of a heart attack?
That's a question we can answer!

@EpiEllie

Northwestern Medicine®

Why are RCTs so great for causal inference?

- **When we estimate intention-to-treat effects in RCTs, randomization becomes our “action”**
  - Randomization ensures no confounding at baseline (**Exchangeability**) ✓
  - Randomization also ensures **Positivity** ✓
  - Randomization is a well-defined intervention (**Consistency**) ✓
  - ITT analyses (usually) give unbiased estimates of ITT effects!

…And yet, RCTs can have issues

- **There can be practical barriers to conducting RCTs**
  - Expense
  - Time-constraints
  - Ethics

- **And sometimes, an ITT analysis does not approximate the per protocol effect**
  - Consider a drug trial that depends on participant adherence to a protocol
    - Recruit n participants; randomize 1:1 to taking drug A <u>for 3 months</u> or placebo
    - Follow for a set period of time (e.g. 1 year); record outcome
    - Analyze according to ITT
  - <u>Post-randomization events</u> are not guaranteed to be unconfounded!

Instead, use observational data to answer questions

- Two categories:
  - Classic epidemiologic studies: cohort studies, case-control studies
  - **“Found”** data: electronic medical records, administrative claims databases, national registers

- Big picture: We want to conceptualize observational studies designed in found data as **conditionally randomized experiments**

Commonly identified sources of bias in causal inference studies using “found” data

- **Confounding**. The bias that arises when we make causal inferences based on comparing non-comparable groups
  - Unmeasured confounders can pose critical issues

- **Selection bias**. This can occur:
  - At baseline (e.g. including prevalent users of a medical treatment)
  - During follow-up (e.g. loss to follow-up of study participants)

- **Measurement error**. This may occur in the:
  - Outcome variable
  - Treatment/exposure variable
  - Confounders

Assessing extent of (unmeasured) confounding

- Observational studies will always have some degree of unmeasured confounding

- Negative controls can provide some reassurance that you aren't missing something monumental

- Instrumental variable analysis

- That said, **unmeasured confounding is often not the biggest issue with observational studies**
  - Selection bias!
  - Assignment of baseline time for analysis

Other issues for causal inference

- Estimates may not be **transportable** to other populations
  - No external validity

- Even if the estimate is unbiased and transportable, it may be too **unstable**
  - Because the effective sample size is too small
  - Use statistical methods to quantify the role of chance

- The model may be **misspecified**.
  - The choice of parametric model to represent the confounding may impact study results

Dr. Miguel Hernán's 2-step Algorithm for Causal Inference

Hernán & Robins. (2016) *American Journal of Epidemiology* doi:10.1093/aje/kwv254

Dr. Miguel Hernán’s 2-step Algorithm for Causal Inference

Step 1. **Ask** a causal question.

Step 2. **Answer** that causal question.

Hernán & Robins. (2016) *American Journal of Epidemiology* doi:10.1093/aje/kwv254

Dr. Miguel Hernán's 2-step Algorithm for Causal Inference

Step 1. **Ask** a causal question.

Step 2. **Answer** that causal question.

Thought experiment: Imagine a *hypothetical* randomized trial that we would prefer to conduct and analyze: the **target trial**

Hernán & Robins. (2016) *American Journal of Epidemiology* doi:10.1093/aje/kwv254

Dr. Miguel Hernán’s 2-step Algorithm for Causal Inference

Step 1. **Ask** a causal question.

Step 2. **Answer** that causal question.

Thought experiment: Imagine a *hypothetical* randomized trial that we would prefer to conduct and analyze: the **target trial**

Then we have a choice:

1. Go into the world and secure funding to conduct the target trial
2. Analyze “found” data as an attempt to emulate the target trial

Hernán & Robins. (2016) *American Journal of Epidemiology* doi:10.1093/aje/kwv254

Dr. Miguel Hernán’s 2-step Algorithm for Causal Inference

Step 1. **Ask** a causal question.

Step 2. **Answer** that causal question.

Thought experiment: Imagine a *hypothetical* randomized trial that we would prefer to conduct and analyze: the **target trial**

Then we have a choice:

1. Go into the world and secure funding to conduct the target trial
2. Analyze “found” data as an attempt to emulate the target trial

Hernán & Robins. (2016) *American Journal of Epidemiology* doi:10.1093/aje/kwv254

Components of the Target Trial

Target trial (hypothetical)	
Eligibility criteria	
Treatment strategies	
Assignment procedures	
Follow-up Period	
Outcome	
Causal contrast(s) of interest	
Analysis plan	

Northwestern Medicine®

Hernán & Robins. (2016) *American Journal of Epidemiology* doi:10.1093/aje/kwv254

Components of the Target Trial

Target trial (hypothetical)	**Analysis in “found” data**
Eligibility criteria	Eligibility criteria
Treatment strategies	Treatment strategies
Assignment procedures	Assignment procedures
Follow-up Period	Follow-up Period
Outcome	Outcome
Causal contrast(s) of interest	Causal contrast(s) of interest
Analysis plan	Analysis plan

Hernán & Robins. (2016) *American Journal of Epidemiology* doi:10.1093/aje/kwv254

Strengths of the Target Trial Approach

- Forces researchers to be explicit in their specification of the (hypothetical) experimental protocol

- Allows others to easily identify areas of study design that could introduce bias

- Use of familiar language facilitates discussions between more quantitative researchers and their clinical colleagues

Hernán & Robins. (2016) *American Journal of Epidemiology* doi:10.1093/aje/kwv254

Example: DMARDs and MACE in RA

- Rheumatoid arthritis (RA) is a chronic inflammatory disease

- People with RA have increased risk of cardiovascular disease

- Addition of disease modifying anti-rheumatic drugs (DMARDs) may provide protection against CVD in people with RA, but existing trials have limitations
  - Sample size
  - Length of follow-up

NM Enterprise Data Warehouse (NMEDW)

- Contains all EHR data from patients in the Northwestern Medicine system from 2000-present
  - Some historical data available as well

- Comprehensive outpatient and inpatient EHR data

- Access to structured data only
  - Vitals, lab results
  - Prescriptions
  - No physician notes for text mining

Example: DMARDs and MACE in RA

Our goal: To use EHR data from the NMEDW to emulate a target trial for the effect of addition of a DMARD to a methotrexate regimen on major adverse cardiovascular events (MACE).

Example: DMARDs and MACE in RA

Our goal: To use EHR data from the NMEDW to emulate a target trial for the effect of addition of a DMARD to a methotrexate regimen on major adverse cardiovascular events (MACE).

How? Pull EHR data from the NMEDW as though it were a prospective cohort study.

A *"Create-Your-Own-Data"* Adventure!

Target Trial: *Who/When?*

	Target trial (hypothetical)	**Emulation in NMEDW (“found” data)**
Eligibility criteria	• Diagnosis of rheumatoid arthritis between January 1, 2000 and December 31, 2020 • Management of RA symptoms via methotrexate monotherapy (12.5mg/wk) for at least 12 weeks prior to enrollment • Age 18-75 years	Same

Rivera et al. (2022) *Available upon request.*

Target Trial: *Who/When? (Cont.)*

	Target trial (hypothetical)	**Emulation in NMEDW ("found" data)**
Eligibility criteria	• Laboratory assessments, taken within 6 months prior to enrollment: • Platelet > 100,000/mm^3 • Estimated glomerular filtration rate > 60 mL/min • White blood cell count > 3,000/mm^3 • Absolute neutrophil count > 1200/mm^3 • Liver transaminases <1.5x upper limit of normal • Hemoglobin > 9 g/dL • Hematocrit > 30%	Same, plus: Expanded window for laboratory assessments to up to 3 months after enrollment

Northwestern Medicine®

Rivera et al. (2022) *Available upon request.*

Target Trial: *Who/When? (Cont.)*

	Target trial (hypothetical)	Emulation in NMEDW ("found" data)
Eligibility criteria	• Physician confirmation of no prior history of serious cardiovascular disease: • Myocardial infarction, heart failure, or coronary revascularization • Other autoimmune rheumatic disease; • Inflammatory bowel disease • Serious infection including hepatitis B, hepatitis C, or HIV • Evidence of active, latent or inadequately treated mycobacterial tuberculosis infection • Lymphoproliferative disorder • Cancer excluding nonmelanoma skin cancer	Same, plus: Comorbid diagnoses were identified using validated ICD-9 and ICD-10 definitions. "No prior history" was limited to the amount of available information in the EDW prior to enrollment

Rivera et al. (2022) *Available upon request.*

Target Trial: *What?*

	Target trial (hypothetical)	**Emulation in NMEDW (“found” data)**
Treatment Strategies	• **Strategy A**: Initiate additional DMARD therapy (any dose and type) within the **grace period**: 24 months of randomization • **Strategy B**: Do not initiate any DMARD within the grace period Under both strategies, leave decision to discontinue methotrexate or DMARD to physician and patient. Patients can receive any additional therapies.	Same, plus: Therapy initiation will be identified through prescription records in the NMEDW.
Outcome	4-point composite of non-fatal myocardial infarction, non-fatal stroke (including hemorrhagic stroke), incident heart failure (including first hospitalization and outpatient diagnosis), and cardiovascular death, certified by a clinician within 3 years of enrollment	Same, plus: Components will be identified using validated ICD-9 and ICD-10 codes

Rivera et al. (2022) *Available upon request.*

Target Trial: *What?*

	Target trial (hypothetical)	**Emulation in NMEDW ("found" data)**
Treatment Assignment	Open-label (unblinded) randomization to **one** treatment strategy at baseline. Participants and clinicians *are aware* of the strategy they are assigned.	Randomization is assumed to be conditional on baseline covariates: • Demographics: age, gender, race and ethnicity, • Comorbid conditions: diabetes, hypertension, other (atrial fibrillation, atherosclerotic cardiovascular disease, chronic kidney disease, chronic obstructive pulmonary disease) • Laboratory Assessments: cholesterol level, estimated glomerular filtration rate

Northwestern Medicine®

Rivera et al. (2022) *Available upon request.*

Target Trial: *When?*

	Target trial (hypothetical)	**Emulation in NMEDW ("found" data)**
Follow-up Period	• Follow-up begins at **time zero**, when an individual meets all eligibility criteria • When an individual is randomly assigned to one of the treatment strategies • Occurs on date patient has been on methotrexate monotherapy for 12 weeks • Follow-up ends at the earliest of • Composite outcome • Administrative end of follow-up (5 years after time zero or 12/31/2020)	Same, plus: End of follow-up includes 2 years without a patient encounter at Northwestern Medicine

Rivera et al. (2022) *Available upon request.*

Target Trial: How?

	Target trial (hypothetical)	Emulation in NMEDW (“found” data)
Causal Contrast	• **Intention-to-treat effect**: effect of being randomized to the strategies at baseline, regardless of whether the individuals adhere to them during follow-up • **Per-protocol effect**: effect of adhering to the strategies (as defined in the protocol) during follow-up	Per-protocol effect only

Northwestern Medicine®

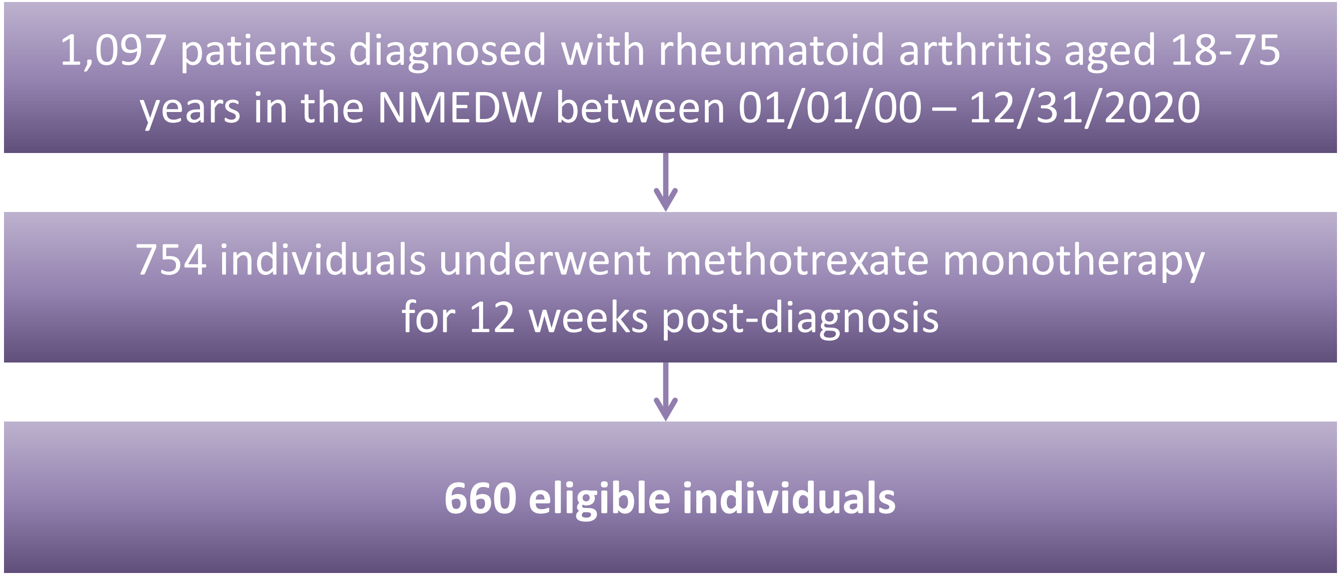
Rivera et al. (2022) *Available upon request.*

Target Trial: How?

	Target trial (hypothetical)	**Emulation in NMEDW (“found” data)**
Statistical Analysis	Per-protocol effect: **use randomization as IV** • **Censor** individuals when they deviate from their assigned protocol • Use a discrete hazards (pooled logistic) model to estimate **absolute risks** • Standardize to calculate an average **hazard ratio** adjusted for confounders • To adjust for potential selection bias, **inverse probability weight** the discrete hazards model to adjust for post-baseline prognostic factors associated with adherence to treatment strategy • **Non-parametric bootstrap** for 95% CIs	To avoid immortal time bias, two choices: 1. Randomly assign individuals who die or are censored in the grace period before fluorouracil initiation to treatment strategy 2. Clone all individuals, assign one clone to each strategy Then conduct analysis as for hypothetical target trial

Rivera et al. (2022) *Available upon request.*

Eligible Sample from NMEDW



1,097 patients diagnosed with rheumatoid arthritis aged 18-75 years in the NMEDW between 01/01/00 – 12/31/2020

754 individuals underwent methotrexate monotherapy for 12 weeks post-diagnosis

660 eligible individuals

Rivera et al. (2022) *Available upon request.*

Brief Sample Description

- 289 individuals (44%) initiated a DMARD within 24 months of baseline (27 months of starting methotrexate)
  - By the end of the grace period, 287 individuals remained in the DMARD arm; 352 in observation arm

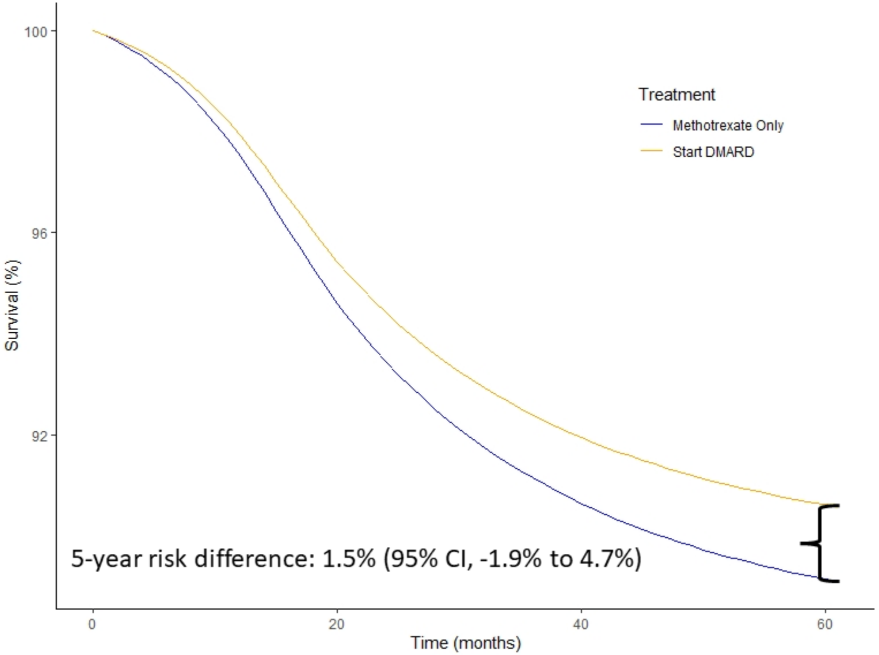
- DMARD initiation was **more likely** in:
  - Those with hypertension and diabetes
  - Those with slightly higher eGFR

- DMARD initiation was **less likely** in:
  - Men

Rivera et al. (2022) *Available upon request.*

Survival estimates for effect of addition of DMARD to Methotrexate on MACE in rheumatoid arthritis patients

Statistic	Estimate (95% CI)
RD	1.5% (-1.9 to 4.7%)
RMST	0.6m (-0.8 to 1.8m)
HR	0.9 (0.5, 1.8)
HR (baseline adjustment only)	0.8 (0.5, 1.4)



5-year risk difference: 1.5% (95% CI, -1.9% to 4.7%)

Northwestern Medicine®

Rivera et al. (2022) *Available upon request.*

Take-Away Messages

- The target trial framework is a useful tool to ensure that your observational analysis is appropriate to answer your scientific question
- Confounding, selection bias, and measurement error are all concepts to consider when designing causal studies in observational data
- **Big data is not always as big as we think it is!**

Acknowledgements

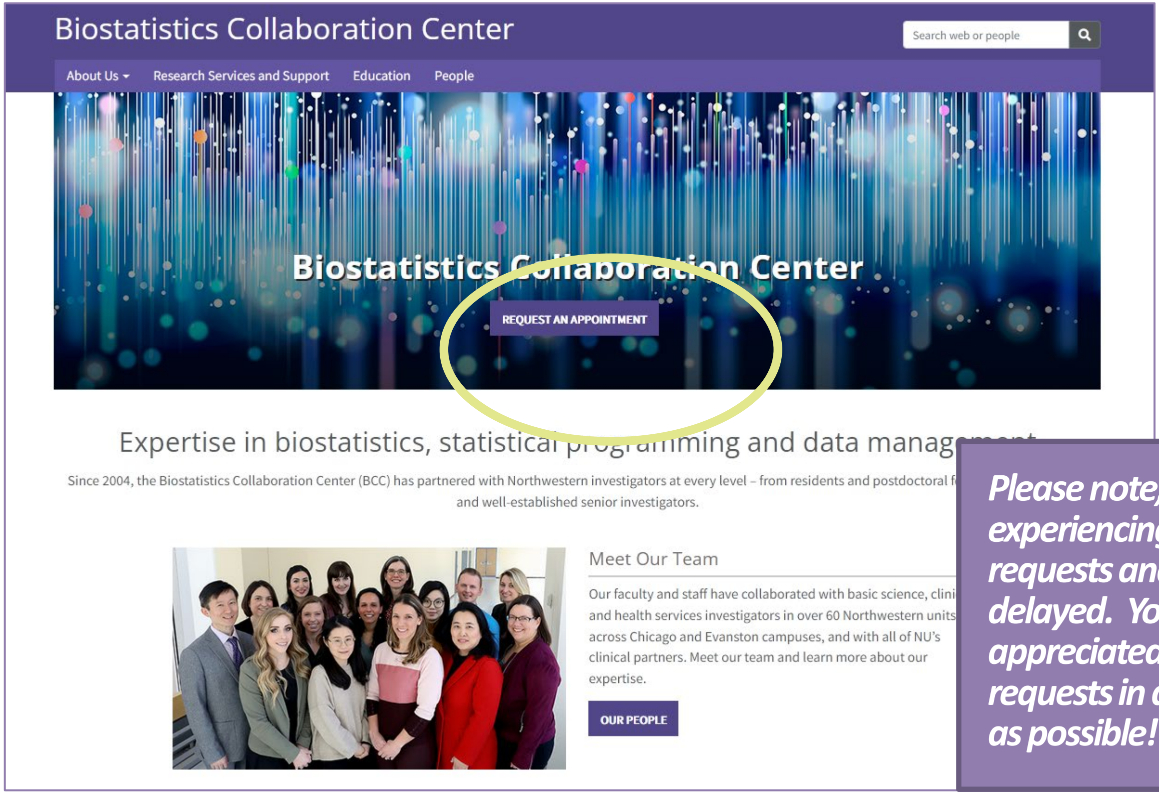
Thanks to my wonderful collaborators!

- Adovich Rivera, MD
- Jacob Pierce, MD
- Arjun Sinha, MD
- Anna Pawlowski
- Donald Lloyd-Jones, MD
- Yvonne Lee, MD
- Matthew Feinstein, MD

References

- Hernán, M.A. & Robins, J.M. (2020). *Causal Inference: What If?.* Boca Raton: Chapman & Hall/CRC.
- Hernán, M.A. and Robins, J.M. (2016). Using big data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology*, 183(8):758-764.
- Murray, E. J., Caniglia, E. C., & Petito, L. C. (2021). Causal survival analysis: A guide to estimating intention-to-treat and per-protocol effects from randomized clinical trials with non-adherence. *Research Methods in Medicine & Health Sciences*, 2(1), 39-49. https://github.com/eleanormurray/CausalSurvivalWorkshop_2019

Thank you!


Biostatistics Collaboration Center
Search web or people
About Us
Research Services and Support
Education
People
Biostatistics Collaboration Center
REQUEST AN APPOINTMENT
OUR PEOPLE

Expertise in biostatistics, statistical programming and data manag

Since 2004, the Biostatistics Collaboration Center (BCC) has partnered with Northwestern investigators at every level – from residents and postdoctoral f and well-established senior investigators.

Meet Our Team

Our faculty and staff have collaborated with basic science, clini and health services investigators in over 60 Northwestern units across Chicago and Evanston campuses, and with all of NU's clinical partners. Meet our team and learn more about our expertise.

Please note, we have been experiencing a high volume of requests and responses may be delayed. Your patience is appreciated as we work to field requests in as timely a manner as possible!

Northwestern Medicine®