Causal Inference in a Big Data World: The Roadmap

Laura Balzer
Causal Inference in a Big Data World: The Roadmap

Laura B. Balzer, PhD

Department of Biostatistics & Epidemiology
School of Public Health & Health Sciences
University of Massachusetts, Amherst

X Congresso Brasileiro de Epidemiologia
October 2017
We often ask causal questions

- Not all questions are causal:
  - What was the average age of participants in our study?
    - Descriptive
    - e.g. Table 1 in our journal papers
  - What risk factors were “significantly” associated with the outcome?
    - Prediction

- But many are causal:
  - Would adherence be higher if all patients had 3-month vs. 1-month prescription refills?
Causal questions

- Asking about the **world under changed conditions**
  - Inference about a distribution that we do not (fully) observe

- How would adherence differ if all patients had 3-month refills VS. if the same patients, over the same time-frame, and under the same conditions had 1-month refills?

  Sadly, we are not time travelers →

  Answering causal questions can be hard
Seemingly non-causal with causal elements

- Missing data can be framed as a causal question

- What proportion of the HIV+ population have suppressed viral replication (<500 copies/ml)?
  - Did not measure viral loads on all HIV+ persons
  - Hypothetical intervention to ensure measurement of viral loads

- Longitudinal settings: hypothetical interventions to prevent censoring (e.g. due to death or outmigration).

Sadly, we are not wizards →

Answering causal questions can be hard
Causal Frameworks to the Rescue

- Provide a systematic process: scientific question → estimation and inference
  - Builds off the scientific method (Bacon’s *Novum Organum*, 1620)
- Many frameworks available
  - Focus on framework for *Targeted Learning* (van der Laan and Rose, 2011)
    - Integrates causal modeling with modern machine learning algorithms and statistical inference
Roadmap for Causal Inference

1. Scientific question
2. Causal model
3. Counterfactuals & causal parameter
4. Linking causal to observed
5. Identify: Knowledge + data sufficient?
6. Estimation
7. Interpretation

(van der Laan and Rose, 2011; Petersen and van der Laan, 2014; Balzer et al., 2016)
1. Specify the scientific question

- What system (including the target population) do we want to study?
- What do we want to learn?
- e.g. What is the effect 3-month vs. 1-month prescription refills on adherence?
  - Which patients? Definition of adherence? Time-frame? Scale?
- **Big Data:** many potential questions \(\rightarrow\) even more important and challenging
2. Causal model representing knowledge & uncertainties

- Causal modeling provides a **formal language** for expressing our knowledge and uncertainties
  - Which variables may affect each other
  - The potential role of unmeasured/background factors
  - The functional form of the relationships

- Many causal models
  - Focus on the directed acyclic graphs
One possibility for the example

- **Reflects knowledge**: what variables may affect another
- **Reflects limits of our knowledge**:
  - No functional form restrictions
  - No assumptions on distribution of unmeasured factors $U$

(Pearl, 2000)
Many sources of association
(a) direct effects
(b) indirect effects
(c) measured confounding
(d) unmeasured confounding
(e) selection bias
(f) all
- Many others not listed

Big Data does not undo “correlation is not causation”
Where are we?

1. Scientific question ✓
2. Causal model ✓
3. Counterfactuals & causal parameter
4. Linking causal model to observed
5. Identify: Knowledge + data sufficient?
6. Estimation
7. Interpretation
Generate counterfactuals (potential outcomes) by intervening on the causal model

- Forces us to be clear about our exposures of interest

For our example:

- $Y_1$: adherence if, possibly contrary-to-fact, the patient had a 3-month refill
- $Y_0$: adherence if, possibly contrary-to-fact, the patient had a 1-month refill

Many more options

- Missing data as the intervention variable
- Different types of interventions: dynamic (personalized), longitudinal
Use counterfactuals to translate the scientific question into a target causal parameter

e.g. difference in the expected adherence if all patients had 3-mo prescription refills vs. if all patients had 1-mo prescription refills

Average treatment effect: $\mathbb{E}[Y_1] - \mathbb{E}[Y_0]$

Causal risk difference: $\mathbb{P}(Y_1 = 1) - \mathbb{P}(Y_0 = 1)$

Many more options
- e.g. causal relative risks, marginal structural models, effects of treatment among treated
4. Linking causal to observed

- Causal model: what we know
- Observed data: what we measure
- Assume the causal model provides a description of our study under
  - Existing conditions (i.e. the real world)
  - Specific interventions (i.e. the counterfactual world)
- Our statistical model should represent real knowledge and uncertainty
  → reflected in our causal model

Linking two worlds
Where are we?

1. Scientific question ✓
2. Causal model ✓
3. Counterfactuals & causal parameter ✓
4. Linking causal model to observed ✓
5. Identify: Knowledge + data sufficient?
6. Estimation
7. Interpretation
5. Identify: Knowledge + data sufficient?

- Currently the parameter of interest is expressed in terms of counterfactuals (e.g. $\mathbb{E}[Y_1] - \mathbb{E}[Y_0]$)
- **Identifiability**: what assumptions are needed to write the causal parameter as a function of the observed data distribution?

We link our day-job (estimation based on the observed data) to our superhero-job (answering causal questions)
5. Identify: Knowledge + data sufficient?

- Intuition only gets you so far
- Causal frameworks can help identify the needed adjustment set(s)
  - Longitudinal effects with time-dependent confounding
  - Missing data and selection bias
  - Effect mediation
- Most statistical parameters are not equal to a regression coefficient

\[ \psi(P) = \sum_w \sum_{\bar{I}} \sum_y y Q(Y = y | \bar{C} = 0, \bar{a}, \bar{I}, w) \prod_{t=1}^{5} Q(I_t | \bar{C} = 0, \bar{a}_{t-1}, \bar{I}_{t-1}, w) Q(w) \]

\[ = \mathbb{E} \left[ \mathbb{E} \left[ \ldots \mathbb{E} \left[ Y | \bar{A}_5 = \bar{a}_5, \bar{C}_6 = 0, \bar{L}_5, W \right] | \bar{A}_4 = \bar{a}_4, \bar{C}_5 = 0, \bar{L}_4, W \right] \ldots | W \right] \]

- Complexity needed to get best possible answer to real world problems
Where are we?

1. Scientific question ✓
2. Causal model ✓
3. Counterfactuals & causal parameter ✓
4. Linking causal to observed ✓
5. Identify: Knowledge + data sufficient? ✓
6. Estimation
7. Interpretation
6. Estimation

- Non-parametric approaches (e.g. contingency tables) often break down due to sparse/empty cells
  - Even worse with Big Data
- But many new advances in machine learning
  - Help us learn complex relationships
  - Without introducing new assumptions
- Many algorithms available for estimating a statistical parameter
  - Corresponding to our causal effect (under the identifiability assumptions)
6. A few estimation algorithms

- **Parametric G-computation:**
  - Plug-in (substitution) estimator of the G-computation identifiability result (Robins, 1986)
  - e.g. Taubman et al. (2009); Young et al. (2011); Zhang et al. (2017)
6. A few estimation algorithms

- **Parametric G-computation:**
  - Plug-in (substitution) estimator of the G-computation identifiability result (Robins, 1986)
  - e.g. Taubman et al. (2009); Young et al. (2011); Zhang et al. (2017)

- **Inverse probability weighting (IPW):**
  - Controls for confounding (± missing data) with weighting
    - All the weight in the right places!
  - e.g. Robins et al. (2000); Bodnar et al. (2004); Hernán and Robins (2006)
6. A few estimation algorithms

Targeted maximum likelihood estimation (TMLE) with Super Learning

- **Super Learning** (van der Laan et al., 2007):
  - Ensemble algorithm using V-fold cross-validation to build the best weighted combination of a library of candidate estimators
  - Help us learn complex, real world relationships
6. A few estimation algorithms

Targeted maximum likelihood estimation (TMLE) with Super Learning

- **Super Learning** (van der Laan et al., 2007):
  - Ensemble algorithm using V-fold cross-validation to build the best weighted combination of a library of candidate estimators
  - Help us learn complex, real world relationships

- **TMLE** (van der Laan and Rose, 2011):
  - Targeting of initial Super Learner estimates to reduce bias/variance and maintain inference
  - Double robust, semiparametric, efficient substitution estimator
7. Interpretation

- **Final step** - consider whether and to what degree the identifiability assumptions have been met

- **Statistical:**
  - As close as we can get to causal effect given the limitations in the data
  - e.g. Estimate of the difference in average adherence associated with 3-month vs. 1-month refills, after adjusting for measured confounders

- **Causal:**
  - If our identifiability assumptions hold, an estimate of the average treatment effect
Where are we?

1. Scientific question ✓
2. Causal model ✓
3. Counterfactuals & causal parameter ✓
4. Linking causal to observed ✓
5. Identify: Knowledge + data sufficient? ✓
6. Estimation ✓
7. Interpretation ✓
Caution - Here be dragons!

- Over-simplification of causal graph
- Introducing new assumptions during estimation
- Machine learning in place of causal modeling
- Over-interpretation of results
- Complex notation and terminology obscuring common sense
Causal Frameworks as a Tool

1. Make uncertainty and limits of knowledge explicit
2. Frame better questions
3. Understand assumptions, and when assumptions are not met, provide guidance on future research
4. Help ensure that parameters estimated come as close as possible to answering the causal question posed
5. Interpret results appropriately
6. Widely applicable
Application - Does this matter?

- **Target population:** cohort of adult residents with an HIV diagnosis at or before baseline ($N = 7108$) in the SEARCH study
  - www.searchendaids.com
  - (NCT01864603)

- **Goal:** estimate the proportion that have suppressed viral replication ($<500$ cps/ml) at follow-up year 2
  - Potentially differential censoring due to death and out-migration
  - Potentially differential measurement of viral loads
  - May both depend on time-dependent characteristics

- **Causal:** Hypothetical intervention to prevent censoring and to ensure complete measurement
Failing to adjust for non-representative testing and differential censoring substantially inflated estimates (Petersen et al., 2017)
- Especially for the subgroup of newly diagnosed
Thank you & Acknowledgements

- Coverage follows “Introduction to Causal Inference” (Petersen & Balzer): www.ucbbiostat.com
- Supported by NIAID of the NIH (R01-AI074345, U01AI099959, 5UM1AI068636), PEPFAR, Bill and Melinda Gates Foundation, and Gilead Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official view of the NIH, PEPAR, Bill and Melinda Gates Foundation or Gilead. The SEARCH project gratefully acknowledges the Ministries of Health of Uganda and Kenya, our research team, collaborators and advisory boards, and especially all communities and participants involved.
A few references - not a complete bibliography


