Modern Statistical Learning Methods for Observational Biomedical Data

Chapter 1: Introduction to causal inference

David Benkeser
Emory Univ.

Marco Carone
Univ. of Washington

Larry Kessler
Univ. of Washington

MODULE 4

Summer Institute in Statistics for Clinical and Epidemiological Research

July 2019

Contents of this chapter

Motivation: the BOLD Study

BACK PAIN OUTCOMES USING LONGITUDINAL DATA (BOLD) STUDY

PI: Jeffrey (Jerry) G. Jarvik, MD MPH
University of Washington
Director, Comparative Effectiveness, Cost and Outcomes Research Center

Overview:

- Funded in 2010 by the Agency for Healthcare Research and Quality (AHRQ)
- Multicenter, multidisciplinary project
- Three primary aims:
  1. Establish BOLD cohort
  2. Conduct observational cohort of early imaging
  3. Conduct LESS Trial, an RCT of epidural steroid injections plus local anesthetic versus local anesthetic alone

Motivation: the BOLD Study

University of Washington Data Coordinating Center

Jerry Jarvik, MD MPH	Patrick Heagerty, PhD	Mahsid Lotfi, PhD
Zoya Bauer, MD PhD	Katie James, PA-C MPH	Chris Nefcy
Brian Bresnahan, PhD	Larry Kessler, PhD	Sean Rundell, PhD
Bryan Comstock, MS	Sarah Lawrence, MA	Sean Sullivan, PhD
Janna Friedly, MD	Danielle Lavallee, PharmD PhD	Judy Turner, PhD

Study sites

Andy Avins, MD	Kaiser Permanente Northern CA
David Nerenz, PhD	Henry Ford Health System, Detroit
Srdjan Nedeljkovic, MD	Harvard Pilgrim / Harvard Vanguard, Boston
Rick Deyo, MD MPH	OHSU, Portland
Venu Akuthota	University of Colorado, Denver
Felix Diehn	Mayo Clinic, Rochester
DJ Kennedy + Matt Smuck	Stanford University
Thiru Annaswamy	Dallas VA
David Sibell	OHSU, Portland
Chris Standaert	University of Washington, Seattle
Michael DePalma	Virginia iSpine, Richmond

Motivation: the BOLD Study

The BOLD registry

- 5239 patients of age $\geqslant$ 65 years with new primary care visit for back pain
- three integrated systems:
  - Kaiser Permanente (Northern CA)
  - Henry Ford Health System (Detroit, MI)
  - Harvard Vanguard / Harvard Pilgrim (Boston, MA)
- patients identified through Health Care Information Systems
- contacted at 3, 6, 12 and 24 months
- main outcomes: pain, disability (Roland Morris Disability Questionnaire – RMDQ), depression and anxiety, healthcare utilization

BOLD/BOLDER status:

- 24 month follow-up completed at all sites
- overall follow-up rates:

  3 months (89%), 6 months (86%), 12 months (85%) and 24 months (78%)

Motivation: the BOLD Study

Early imaging cohort study

- Primary objective is to compare effectiveness of early imaging compared with no early imaging in seniors with a new visit for back pain.
- Outcomes: disability (RMDQ), pain, and subsequent use of healthcare resources.
- Original analysis:
  - Treatment group according to existing guidelines: patients who underwent lumbar spinal imaging within 6 weeks of their index visit were said to have received early imaging.
  - Each treated patient was propensity-matched with a BOLD cohort patient who did not have any spine imaging within 6 weeks of the index visit.
  - Propensity score was estimated using logistic regression.
  - Linear mixed models used to estimate differences between treatment groups adjusted for many variables, including:
    - age, education and race;
    - total relative value units (RVU);
    - spine-specific RVUs (subdivided as physical therapy, injection therapy, imaging, and surgery);
    - patient-reported outcomes measures at 3, 6 and 12 months;
    - and reimbursement estimates.

Motivation: the BOLD Study

Research

Original Investigation

Association of Early Imaging for Back Pain With Clinical Outcomes in Older Adults

Jeffrey G. Jarvik, MD, MPH; Laura S. Gold, PhD; Bryan A. Comstock, MS; Patrick J. Heagerty, PhD; Sean D. Rundell, DPT, PhD; Judith A. Turner, PhD; Andrew L. Avins, MD, MPH; Zoya Bauer, MD, PhD; Brian W. Bresnahan, PhD; Janna L. Friedly, MD; Kathryn James, PA-C, MPH; Larry Kessler, ScD; Srdjan S. Nedeljkovic, MD; David R. Nerenz, PhD; Xu Shi, BS; Sean D. Sullivan, PhD; Leighton Chan, MD, MPH; Jason M. Schwalb, MD; Richard A. Deyo, MD, MPH

JAMA. 2015;313(11):1143-1153. doi:10.1001/jama.2015.1871
Corrected on March 25, 2015.

Author Affiliations: Author affiliations are listed at the end of this article.

Corresponding Author: Jeffrey G. Jarvik, MD, MPH, Comparative Effectiveness, Cost and Outcomes Research Center (CECORC), Department of Radiology, University of Washington, PO Box 359728, 325 Ninth Ave, Seattle, WA 98104-2499 (jarvikj@uw.edu).

Motivation: the BOLD Study

Conclusions (?) regarding early imaging:

- Older adults undergoing early imaging for back pain do not experience improved outcomes but incur greater healthcare resource utilization.
- Early imaging does not result in a higher rate of cancer detection.
- These findings should help to inform imaging guidelines for older adults.

What do alternative analyses of these data reveal?

Motivation: the BOLD Study

Updates on projects and publications can be found at **www.backpainproject.com**

What is causal analysis?

Standard statistical analysis summarizes features of an observed population.

- Average disability score in patients observed to undergo early imaging.
- Average disability score in patients observed not to undergo early imaging.

We compare these features to infer associations between variables.

- Average disability score in patients with versus without early imaging.
- "Association" between early imaging and disability score.

Causal analysis summarizes features of a population under an intervention of interest.

- Average disability score if all patients underwent early imaging.
- Average disability score if no patient underwent early imaging.

We compare these features under different interventions to infer treatment effects.

- Average disability score if all patients underwent early imaging versus if none did.
- Causal relationship between early imaging and disability score.

What is causal analysis?

The data alone cannot tell us how things would change under different interventions.

- This must be inferred based on scientific understanding.
- What do we know biologically about potential causes of disability?
- What do we know clinically about who doctors refer to early imaging?

Randomized trials provide some but not all experimental control.

- Participants may be randomized to treatment arm, but at times drop out of trials and do not comply with treatment.
- What would the effect be if no one dropped out and everyone complied?

Consequently, causal analysis drawn from all observational studies and many clinical trials will rely on untestable assumptions.

What is causal analysis?

To move from associative analysis to causal analysis, we require new tools.

1. How do we codify our scientific knowledge about observed data?
2. How do we describe the way we would like to intervene?
3. How do we express the data we would have seen under intervention?
4. When can we use observed data to infer what would happen under intervention?
5. When we can infer from observed data, how can we do so efficiently?

A formal framework is useful to move towards causal analysis (Petersen & van der Laan, 2014).

A road map for causal analysis

1. Specify a causal model representing scientific knowledge.
2. Specify causal question and causal parameter.
3. Specify observed data and link to causal model.
4. Assess identifiability of quantity of interest.
5. Specify a statistical parameter and statistical model.
6. Estimate the chosen parameter.
7. Interpret results.

A road map for causal analysis

Causal models describe scientific information underlying our study.

- What affects a physician's decision to prescribe early imaging?
- What are major determinants of back-related disability?

Structural causal models (SCM) (Pearl, 2000) are used to

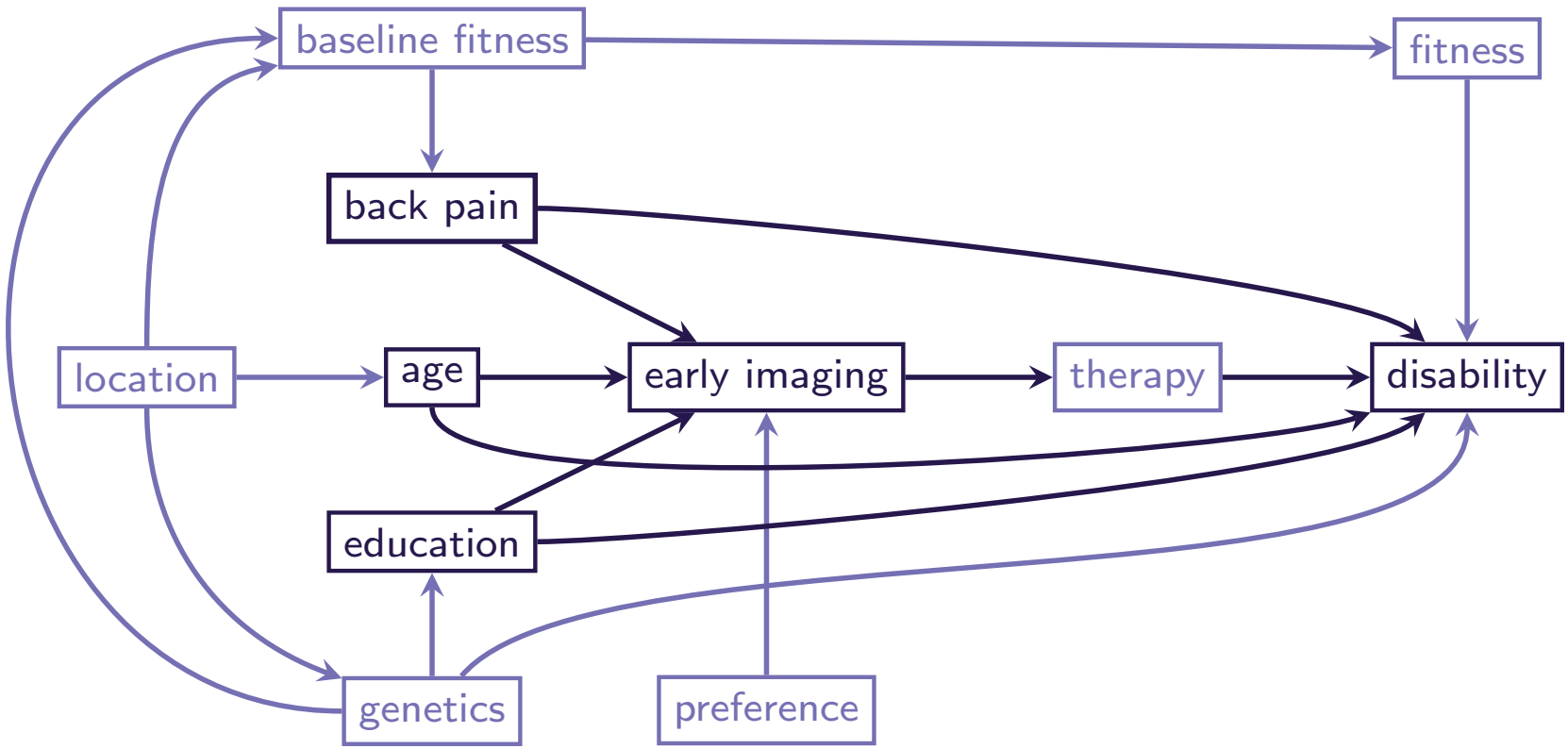
1. codify scientific knowledge underlying our observations;
2. define relevant causal questions;
3. evaluate assumptions needed to link causal effects to observed associations.

There are three main components to an SCM:

1. endogenous variables: measured factors meaningful to our scientific question;
2. exogenous variables: unmeasured factors that influence observations;
3. structural equations: how Nature uses these variables to assign observation values.

A road map for causal analysis

1. endogenous variables: measured factors meaningful to our scientific question;
2. exogenous variables: unmeasured factors that influence observations;
3. structural equations: how Nature uses these variables to assign observation values.



A road map for causal analysis

Nature generates experimental data as follows:

1. exogenous characteristics (e.g., genetics) determine patient's covariates;
2. based on these covariates and exogenous characteristics (e.g., clinician's preference), a clinician assigns early imaging (or not);
3. based on covariates, early imaging status, and exogenous characteristics (e.g., fitness), 12-month disability score is determined

A road map for causal analysis

Causal models encode knowledge and assumptions about how data are generated.

- What factors did Nature rely upon when determining patient's back pain?
- What factors did the physician consider when assigning treatment?

Two common forms of assumptions:

1. **Exclusion restrictions:**
   - Does not refer to criteria for excluding a subject from a study!
   - Leaving off arrows between endogenous variables.
   - Example: early imaging does not depend on covariates. When is this justified?
2. **Independence assumptions:**
   - Leaving off arrows between exogenous variables.
   - Do unmeasured factors that affect whether patient receives early imaging also affect subsequent therapy that patient might receive?

A road map for causal analysis

Causal models should represent real knowledge and often make us uncomfortably aware of how little we know and/or how little we actually measured.

Later, we will see that additional assumptions may be needed to make progress.

- However, best practice to keep this process separate.
- In Chapter 6, we discuss what to do when necessary assumptions are not met.

A road map for causal analysis

1. Specify a causal model representing scientific knowledge.
2. **Specify causal question and causal parameter.**
3. Specify observed data and link to causal model.
4. Assess identifiability of quantity of interest.
5. Specify a statistical parameter and statistical model.
6. Estimate the chosen parameter.
7. Interpret results.

A road map for causal analysis

Our SCM encoded what we know about our existing experiment.

- Likely different than the experiment we would like to have performed.
- We would like to see what happens if everyone (or no one) received early imaging.

We can use the SCM to define our ideal experimental conditions.

- What parts of the experiment do we wish to change?
- Rather than letting clinicians determine early imaging, we would make everyone (or no one) undergo early imaging.

A causal parameter is a summary of the post-intervention data defined by:

1. deciding what variable(s) to intervene on;
2. deciding how to intervene on them;
3. specifying the post-intervention outcome of interest;
4. specifying a summary of this outcome.

A road map for causal analysis

1. **Deciding what variable(s) to intervene on.**
   - What would we change in our SCM to achieve our ideal experiment?
   - We focus on interventions on a single variable, A.
   - This requires being very specific: e.g., "provide early imaging," "recommend early imaging," or something else?
2. **Deciding how to intervene on them.**
   - "Static" interventions set a fixed level for everyone.
   - Other interventions often interesting but beyond the scope of this short course.
3. **Specifying the outcome of interest.**
   - We focus on an outcome at a single time point.
   - We also assume no competing risks (e.g., death).
4. **Specifying a summary of this outcome.**
   - Do we care about what happens to the average outcome under intervention?
   - The median outcome? The 10th percentile?

A road map for causal analysis

Nature generates experimental data as follows:

1. exogenous characteristics (e.g., genetics) determine patient's covariates;
2. based on these covariates and exogenous characteristics (e.g., clinician's preference), a clinician assigns early imaging (or not);
3. based on covariates, early imaging status, and exogenous characteristics (e.g., genetics), 12-month disability score is determined

We would generate our ideal data as follows:

1. exogenous characteristics (e.g., genetics) determine patient's covariates;
2. **assign early imaging to all patients**;
3. based on covariates, early imaging status, and exogenous characteristics (e.g., genetics), a **counterfactual** 12-month disability score is determined;
4. **repeat, but assign no early imaging to patients**.

A road map for causal analysis

Our ideal data unit for each patient is thus $(W, Y(1), Y(0))$, where

- W are baseline covariates;
- $Y(1)$ is the counterfactual 12-month disability score if early imaging received;
- $Y(0)$ is the counterfactual 12-month disability score if no early imaging received.

The counterfactuals $Y(1)$ and $Y(0)$ have a distribution due to randomness in sampling patients. For example, differences in genetics lead to different levels of disability.

A causal parameter is a summary of the distribution of counterfactuals. A causal effect compares causal parameters under different interventions. Here, we will focus on the average treatment effect

$$ATE \;:=\; E[Y(1) - Y(0)] \;=\; E[Y(1)] - E[Y(0)] \;.$$

In some problems, we might prefer different contrasts of counterfactuals. For example, in the world of vaccine trials, we define vaccine efficacy as

$$VE \;:=\; 1 - \frac{E[Y(1)]}{E[Y(0)]} \;,$$

where outcome Y is infection status at a fixed time after vaccine administration.

A road map for causal analysis

There are two key assumptions underlying the definition of counterfactuals.

1. **No interference** (SUTVA):

   A patient's outcome does not depend on treatment assigned to other patients.

   - If all my friends get flu vaccines, I do not get flu regardless of my vaccine status.
   - If none of my friends get flu vaccines, I may not get the flu if I do get a vaccine, but get the flu if I do not.

2. **Autonomy of structural equations** (consistency):

   Our intervention does not change how the treatment works.

   - If we assign $A = a$, we see outcome $Y(a)$.
   - If Nature assigns $A = a$, we observe outcome $Y = Y(a)$.

Fundamental problem of causal analysis:

For each patient, we get to see $Y(0)$ or $Y(1)$ but never both.

A road map for causal analysis

1. Specify a causal model representing scientific knowledge.
2. Specify causal question and causal parameter.
3. **Specify observed data and link to causal model.**
4. **Assess identifiability of quantity of interest.**
5. Specify a statistical parameter and statistical model.
6. Estimate the chosen parameter.
7. Interpret results.

Key points of Chapter 1

- Associative analysis describes features of an observed data population. Causal analysis describes features of a population when an intervention is applied.
- The road map for causal analysis gives a framework for performing causal analysis.
- Structural causal models are a useful way to codify scientific knowledge and specify interventions that are of interest.
- Counterfactuals are outcomes that we would see under intervention.
- Causal parameters are summaries of the distribution of counterfactual outcomes.
- We never get to see both counterfactuals, so we need to understand whether and how we can link observed data to counterfactual data.

References and additional reading

References:

Jarvick J, Gold L, Comstock B, Heagerty P, et al. (2015). Association of Early Imaging for Back Pain With Clinical Outcomes in Older Adults. *JAMA*: 313(11): 1143–53. doi: 10.1001/jama.2015.1871.

Pearl J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York. doi: 10.1017/CBO9780511803161.

Petersen M, van der Laan M (2014). Causal models and learning from data: Integrating causal modeling and statistical estimation. *Epidemiology*; 25(3): 418426. doi: 10.1097/EDE.0000000000000078.

Additional reading:

Pearl J (2009). Causal inference in statistics: An overview. *Stat Surveys*; 3(1): 96–146. doi: 10.1214/09-SS057.

Rose S, van der Laan M (2011). Defining the Model and Parameter. *Targeted Learning: Causal Inference for Observation and Experimental Data*; Chapter 2: 21-42. Springer New York. doi: 10.1007/978-1-4419-9782-1.