Causal Modeling with Applications in Clinical & Policy Studies

Sharon-Lise Normand

Blavatnik Institute, Department of Health Care Policy, Harvard Medical School
Department of Biostatistics, Harvard Chan School of Public Health
Thanks: R01-GM11339

April 10, 2019

Outline

- Introduction
  - Regulatory Science
  - What is Evidence?
  - Examples
- Common statistical issues
  - Observational data - risk adjustment
  - Clustering
  - Multiple related outcomes* or multiple competing treatments
- Concluding remarks

Regulatory Science

- Medical product entry
  - **Food & Drug Administration**: evaluates medical product safety, effectiveness, and quality
  - Post-market **safety** assessments
- Insurance coverage for medical products/services
  - **Medicare Evidence Development and Coverage Advisory Committee**: evaluates medical literature, technology assessments, etc., on benefits, harms, and appropriateness of medical items and services to make health care coverage recommendations
    - Must **extrapolate** treatment benefits to their population
- Transportation
  - **Federal Motor Carrier Safety Administration**: evaluates the safety of large trucks, buses, and commercial vehicles
    - Use data on safety-based regulations collected from **roadside** inspections and crash reports

Sometimes, must rely on **observational** data

Scientific Evidence for Medical Decisions

- **Accumulation** of information to support or refute a theory or hypothesis
- Replication important
- Underlying mechanism important

Commentary

Statistics in Medicine

Received 18 January 2010, Accepted 3 March 2010 Published online in Wiley Interscience

(www.interscience.wiley.com) DOI: 10.1002/sim.3933

What is evidence?

Sharon-Lise T. Normand[a,b*†] **and Barbara J. McNeil**[a,c]

The assumption that comparative effectiveness research will provide *timely, relevant evidence* rests on changing the current framework for assembling evidence. In this commentary, we provide the background of how coverage decisions for new medical technologies are currently made in the United States. We focus on the statistical issues regarding how to use the ensemble of information for inferring comparative effectiveness. It is clear a paradigm shift in how clinical information is integrated in real-world settings to establish effectiveness is required. Copyright © 2010 John Wiley & Sons, Ltd.

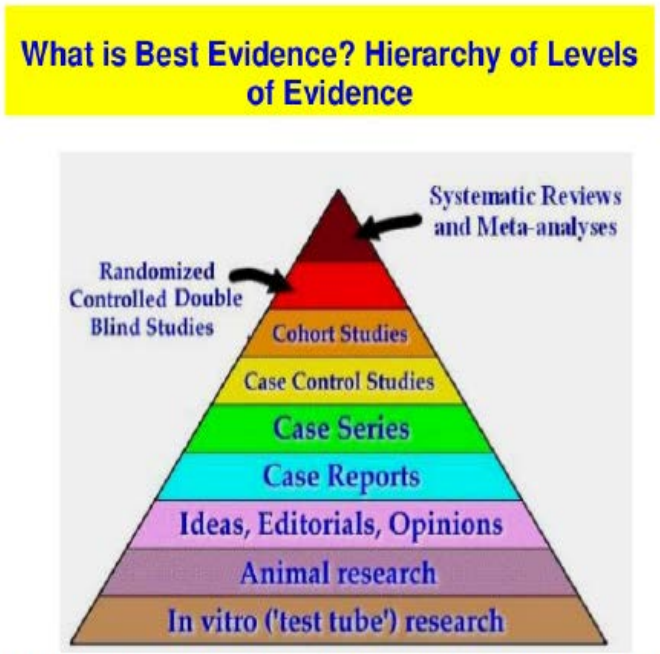
Keywords: evidentiary evaluation; multiple outcomes and comparisons; hierarchical Bayesian models; Bayes factors; posterior predictive probability

1. Introduction

The current paradigm for integrating clinical information in real-world settings to establish whether benefits outweigh risks is out-dated. Divergence from this paradigm involves the recognition that randomized controlled trials that often serve as the basis for new technology approval are small and short-term, and post-market studies are often voluntary and difficult to implement. These problems have become increasingly important over the last decade because technology is changing at a rapid pace, therapies are utilized outside their intended populations, and more representative groups of patients are likely to have differential responses to the same therapy.

In their paper, Tunis and colleagues identify several areas of improvement to provide timely and better evidence of

Hierarchy of Evidence

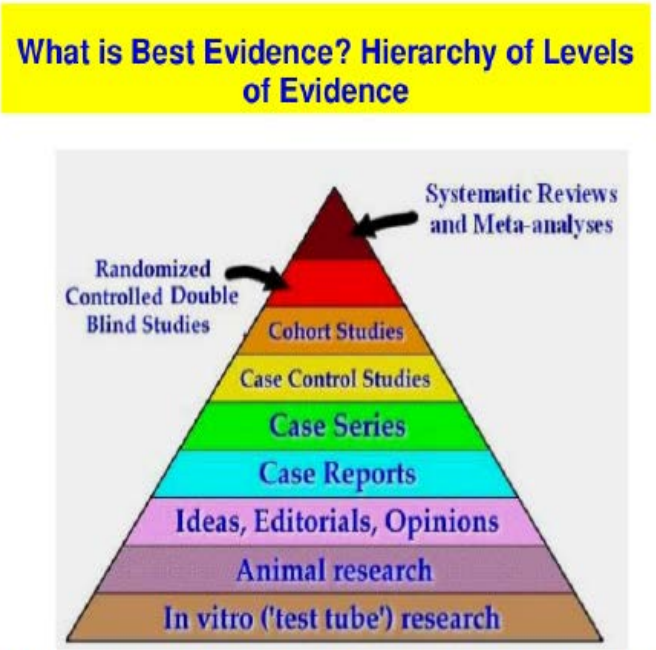


Source: quoteimg.com Research Evidence Hierarchy Pyramid

Therefore, **many designs** contribute to evidence base

Hierarchy of Evidence

Bradford Hill: 1965



Source: quoteimg.com Research Evidence Hierarchy Pyramid

1. Strength of association
2. Consistency
3. Specificity
4. Temporality
5. Biological gradient
6. Plausibility
7. Coherence
8. Experiment
9. Analogy

Hill AB. The environment and disease: association or causation? Proceedings of the Royal Society of Medicine, 1965;58:295-300.

Contemporary Setting

With data acquisition technologies, biggest challenge is **data integration**

- molecular biology
- toxicology
- genotoxicology
- imaging
- functional MRI
- electronic health records
- mobile applications

DuMouchel and Harris. Bayes methods for combining the results of cancer studies in humans and other species. JASA 1983;78(382):293-308.

Table 1. Two-Way Table of Epidemiological Studies and Laboratory Experiments on Nine Environmental Mixtures[a]

	Roofing Tar Emissions	Coke Oven Emissions	Diesel Engine Emissions A	B	C	D	Gasoline Engine Emissions	Benzo(a) pyrene	Cigarette Smoke[f]
Lung Cancer (Humans)[b]	1.64 1.41 .50	4.40 .34 1.48							.03 .15 −3.46
Skin Tumor Initiation (Sencar Mice)[c]	.54 .04 −.63	2.10 .04 .74	.53 .04 −.64	.16 .22 −1.86		.01 .82 −4.51	.03 .26 −3.61	85.28 .03 4.45	.00 1.30 −5.88
Enhancement of Viral Transform. (SHE cells)[d]	2.07 .18 .73	.86 .10 −.15	.65 .15 −.44	.07 .33 −2.70	.13 .18 −2.06	.04 .59 −3.24	.20 .12 −1.59	540.00 .04 6.29	.58 .08 −.54
Mutagenesis −MA (Mouse 5178Y Lymphoma Cells)[e]	.31 .39 −1.17	.73 .21 −.32	1.66 .31 .51	.27 .43 −1.31	2.55 .16 .93	.16 .24 −1.86	.35 .11 −1.06		.59 .23 −.53
Mutagenesis +MA (Mouse 5178Y Lymphoma Cells)[e]	9.56 .16 2.26	9.96 .07 2.30	1.87 .26 .63	.76 .14 −.27	1.01 .20 .01	.05 .43 −3.02	.99 .10 −.01		.45 .13 −.79

[a] Each nonempty cell contains the observed dose-response slope, its coefficient of variation, and the natural logarithm of the dose-response slope.
[b] Units of measurement (for all agents except cigarette smoke) are the increment in the relative risk per 10^8 $\mu g/m^3$ of organic extractables per year.
[c] Units of measurement are papillomas per mouse per mg of organic extractables at 27 weeks.
[d] Units of measurement are transformations per 2×10^6 surviving cells per μg/ml of organic extractables; "SHE" = Syrian hamster embryo.
[e] Units of measurement are average mutant colonies per 10^6 survivors per μg/ml of organic extractables; "−MA" = without metabolic activator; "+MA" = with metabolic activator.
[f] Units of measurement for cigarette smoke refer to whole smoke condensate rather than organic extractables.

Combining **heterogenous** data sources not new

Schizophrenia

- High clinical need & difficulty navigating health care system
- Low quality & black/white disparities
- Medicaid is largest payer
  - Antipsychotics are costliest therapeutic category for state programs

Horvitz-Lennon et al., Health Services Research, 2014

Schizophrenia

- High clinical need & difficulty navigating health care system
- Low quality & black/white disparities
- Medicaid is largest payer
  - Antipsychotics are costliest therapeutic category for state programs

- Medicaid policies/benefits vary across states but within-state:
  - Covered by a single payment system with identical policies/benefit structure
  - Variation in access to federal, state, and local (county) resources
  - Extant evidence suggests geography plays a role

Horvitz-Lennon et al., Health Services Research, 2014

Schizophrenia

1. Does **quality of mental health care** differ among black, white, and Latino Medicaid beneficiaries?
2. Do quality and disparities **change** over **time**?
3. Do quality and disparities **vary** across **counties**?

Drug Reformulations

- Manufacturer reformulate existing products to extend product life cycle (1984 Hatch-Waxman Act)
- Shift demand for original formulation (soon lose patent protection) to the reformulation
- Reformulations involve less frequent dosing, gradual release of active ingredient, or easier to administer
- Antidepressant reformulations common (**original** vs reformulation):
  - **Celexa** vs Lexapro (single isomer); **Paxil** vs Paxil CR (controlled release); **Remeron** vs Remeron Soltab (disolvable tablet)
- Clinical trial evidence is sparse; mixed at best
- Huskamp et al., Health Affairs 2009

Drug Reformulations

Do **anti-depressant reformulations** decrease medication discontinuation rates compared to **original formulations**?

Specific Drug Eluting Stents

- Rapid proliferation of drug eluting stents (DES)
- U.S. has 2nd highest number of overall stent insertions per capita
- Multiple competing versions supported by a few manufacturers
- Differences include polymer coating, specific drug, platform type, and delivery system
- Study **21,000+** adults, **10** model-specific DES, **3** manufacturers
- Rose & Normand; Biometrics 2018

Specific Drug Eluting Stents

Do particular **model-specific DES** cause fewer adverse cardiac events compared to other **model-specific DES**?

Common Themes

- Observational data
- Clustered data
  - Adults living within counties
  - Patients nested within hospitals
- Multiple competing treatments: 10 different drug eluting stents
- Multiple outcomes: quality indicators

Lack of Randomization

- Causal inference
  - Special case of **predictive** inference among subjects who could receive any of the different treatment options
  - Renewed interest
- Why increased interest?
  - Increasing availability of data
  - Increasing availability of different data types (e.g, text, images, etc.)
  - Ignorable treatment assignment more plausible by conditioning on more data
- Notation
  - $T = 1$ new and $T = 0$ standard treatment (assume binary)
  - Y observed outcome
  - Y_1, Y_0 **potential outcomes** under $T = 1$ and $T = 0$
- Assumption that potential outcome exists is fundamental

Causal Assumptions

1. Sample is representative of **target** population
2. Outcomes for one subjects is independent of treatment assignment of other subjects **and** treatments are well-defined & the same for all subjects (SUTVA)
3. Within the subpopulations defined by the confounders, treatments are **randomly assigned**
   - Untestable assumption (sensitivity analysis, multiple comparison groups, control outcomes)
4. There are subjects from all treatment groups at every combination of observed confounders
   - **Structural** violations
   - Practical violations due to **finite** sample size
   - Statistically **testable**
5. Constant (vs Non-constant) treatment effect

Assumptions: Randomized vs Not

Feature	Clinical Trial	Observational Setting Study	Comment
Treatment, T*	Well-defined and does not vary across subjects		
P(T)	Known	Unknown	**Estimate**
Positivity $0 < P(T) < 1$	Yes by design	Not necessarily	Real & finite sample sample violations
Comparability	Almost always	Sometimes	Assess balance on **observables**
Effect Estimate	Intention to Treat (ITT)	Adjusted as-treated	Sensitivity to measured and **hidden** confounders**
Postrandomization Bias	Possible	Possible	
	Loss to follow-up, time-varying risks, competing risks		

*Assume treatment received by subject A does not affect the outcome of subject B; **Mitigation strategies: falsification outcome, multiple control groups; Kunz, Rose, Spiegelman, Normand (Chapter 1); Hernán, Robins (Chapter 3); Methods in Comparative Effectiveness Research, 2017

Approaches (T Treatment, Y Outcomes, $\mathbf{X}$ Covariates)

Joint Distribution

$$\begin{aligned} \mathrm{P}(\mathrm{Y}, \mathsf{T}, \mathbf{X}) &= \mathrm{P}(\mathrm{Y} \mid \mathsf{T}, \mathbf{X}) \times \mathrm{P}(\mathsf{T} \mid \mathbf{X}) \times \mathrm{P}(\mathbf{X}) \\ &= \mathrm{Q}_{\mathrm{Y}} \times \Pi_{\mathsf{T}} \times \mathrm{Q}_{\mathrm{X}} \end{aligned}$$

1. Treatment effect depends **only** on Q_{Y} **and** Q_{X}

$$\mathrm{E}_{\mathrm{X}}\left(\mathrm{E}(\mathrm{Y} \mid \mathsf{T}=1, \mathbf{X})\right) - \mathrm{E}_{\mathrm{X}}\left(\mathrm{E}(\mathrm{Y} \mid \mathsf{T}=0, \mathbf{X})\right)$$

2. Π_{T} is the **propensity score** (nuisance)

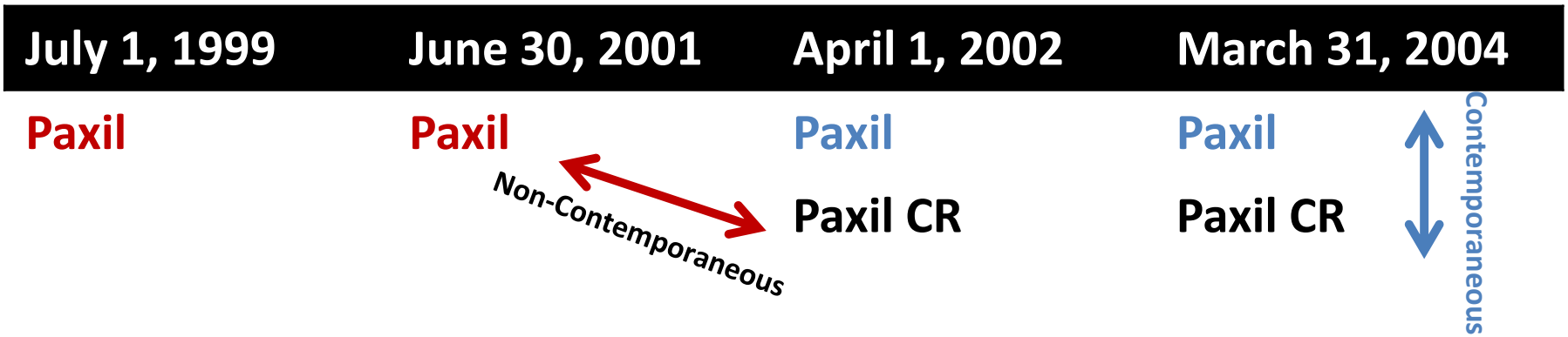
$$\Pi_{\mathsf{T}} = \mathrm{P}(\mathsf{T}=1 \mid \mathbf{X})$$

$\mathbf{X}$ could be **very high-dimensional**

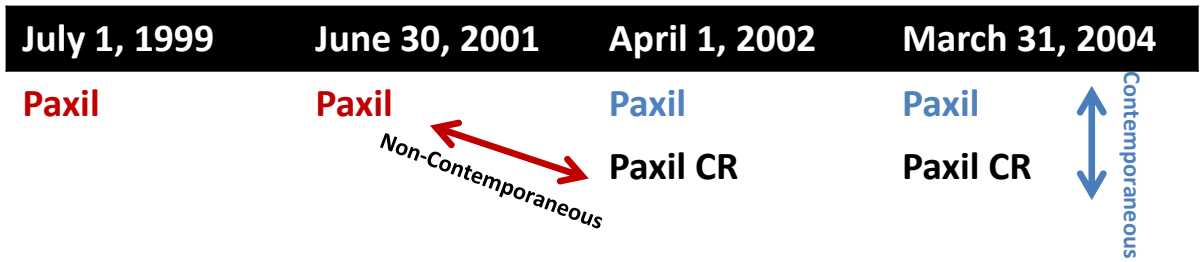
Approaches: 3 Types (considered today)

1. Model **only** the treatment assignment mechanism via regression
   - Propensity score, $\Pi_T = P(T \mid \mathbf{X})$
   - Weight or match using $\hat{\Pi}_T$
2. Model **only** the outcome via regression
   - Multiple regression modeling
   - Parametric g-computation
     - Step 1: Estimate regression model for outcome and treatment association
     - Step 2: Plug predictions from Step 1 into parameter mapping for causal parameter
3. Model **both** the treatment assignment mechanism and outcome
   - Augmented Inverse Probability of Treatment Weights
   - Target Maximum Likelihood

Reformulation Example



Reformulation Example



Rx Dates	Number of Subjects
4/1/02-3/31/04	Paxil CR = 24224
7/1/99-6/30/01	Paxil = 36811

Non-Contemporaneous (Matched)

Rx Dates	Number of Subjects
4/1/02-3/31/04	Paxil CR = **23713**
7/1/99-6/30/01	Paxil = 23713

Rx Dates
4/1/02-3/31/04
Number of Subjects
Paxil = 19413

Contemporaneous (Matched)

Rx Dates	Number of Subjects
4/1/02-3/31/04	Paxil CR = 14307
4/1/02-3/31/04	Paxil = 14307

Reformulation Example

Non-Contemporaneous Matched Samples[‡]

Reformulation versus Original Formulation

Comparison	Days to Discontinue	RR[†] (95% CI)	# Pairs
Lexapro	91	0.83 (0.80, 0.85)	18,045
Celexa	64	1.00	
Paxil CR	64	0.87 (0.85,0.89)	23,713
Paxil	61	1.00	
Remeron Soltab	65	1.04 (1.00,1.08)	10,820
Remeron	66	1.00	

[†]Kaplan-Meier Analysis of Risk of Antidepressant Discontinuation

[‡]No change with **contemporaneous** matched sample except for Remeron Soltab; RR $= 0.94(0.90, 0.98)$

Target Maximum Likelihood

- **Key idea**: no need to maximize entire likelihood $(Y, T, \mathbf{X})$ because causal parameter only depends on Q_Y and Q_X
- Step 1: Estimate outcome model and treatment model
- Step 2: Plug predictions from Step 1 into parameter mapping for causal parameter

$$\text{Step 1}: \quad E^*(Y \mid T = t, \mathbf{X}) = E^0(Y \mid T = t, \mathbf{X}) + \epsilon_t H^*(T, \mathbf{X})$$

$$H^*(T, \mathbf{X}) = \frac{T}{\Pi_T} - \frac{1 - T}{1 - \Pi_T}$$

$$\text{Step 2}: \quad \frac{1}{N} \sum_{i=1}^{N} \left(E^*(Y \mid T = 1, \mathbf{X_i}) - E^*(Y \mid T = 0, \mathbf{X_i}) \right)$$

- E^*: **targeted** estimate of regression of Y on $(T, \mathbf{X})$ obtained by moving the initial estimate E^0 by fluctuations defined by $\epsilon_t H^*(T, \mathbf{X})$

Summary

Approach	Strengths	Weaknesses
IPTW $\frac{1}{\Pi_T}$	Simple	Large variance estimates Weight trimming bias
Regression	Parametric Simple	Extrapolation if violate positivity Functional form
G-Comp	Parametric Simple	Extrapolation if violate positivity Functional form
A-IPTW	Double robust Asymptotic efficiency	Finite sample inefficient
TMLE	Double robust Asymptotic efficiency Finite sample efficiency	

Clustering

- Clustered data: when units are nested completely within other units
  - Longitudinal data: observations are clustered within subjects
  - Levels of clustering may be > 2
- Problem introduced: observations within a common unit are statistically **dependent**
- In practice: the between-unit variance may be a **nuisance** parameter or it may be of **interest**
- Marginal models (GEE) treat the between-unit variation as a nuisance parameter
  - Regression parameters represent association of patient-level covariates with changes in the **population mean** outcome
- Hierarchical models (mixed models, random effect models) introduce random effects
  - Regression parameters represent association of patient-level covariates with changes in the **patient's** outcome

Clustering: Other Considerations

- Interest in covariate effects at different levels of the hierarchy
- Cross-level interactions

$$\begin{aligned} Y_{ij} \mid \beta_i &= x_{ij}\beta_i + z_{ij}\alpha + \epsilon_{ij}; \epsilon_{ij} \sim N(0, \sigma_i^2) \\ \beta_i \mid \tau^2 &= w_i\gamma + u_i; u_i \sim N(0, \tau^2) \end{aligned}$$

- w_i is a county-level covariate
- x_{ij} is subject j living in country i covariate

Can always derive marginal model from conditional model

Multiple Outcomes

- Increasingly collected in clinical trials to measure effectiveness or efficacy
  - Label extensions
- Often measures are **non-commensurate**
  - Measured on different scales
- Common approaches:
  - Consider each outcomes separately using a univariate framework
  - Create a composite measure
- A multivariate approach would:
  - Use information contained in the correlation between outcomes
  - Permit better control over Type I error rates
  - Answer intrinsically multivariate questions

Schizophrenia Example: Quality Indicators

Established quality indicators endorsed by professional societies

- Cover pharmacological, psychosocial, and appropriateness services
- Majority operationalizable using billing data:
  - adequate clozapine dose and duration
  - avoidance of antipsychotic polypharmacy
  - use of psychosocial services
  - follow-up within 7-days of hospital discharge
  - few emergency department visits
- At least **15** established measures of quality of care

Measures are non-commensurate but thought to reflect a single concept (quality of care)

- Why clustered within county?
  - Counties are the administrative unit

Likelihood-based Approaches

Key Idea: avoid **direct** specification of the multivariate likelihood.

- Factorization (a few variables)
  - Cox and Wermuth, *Biometrika*, 1992; Fitzmaurice and Laird, *JASA*, 1995; Catalano and Ryan, *JASA*, 1992.
- Introduction of a latent variable to model correlation among the multiple outcomes
  - Sammel et al., *JRSSB*, 1997; Arminger and Kusters, *Latent Trait and Latent Class Models*, 1988; Dunson, *JRSSB*, 2000.
- Quasi-likelihood: use quadratic exponential model to develop joint estimating equations
  - Prentice and Zhao, *Biometrics*, 1991.

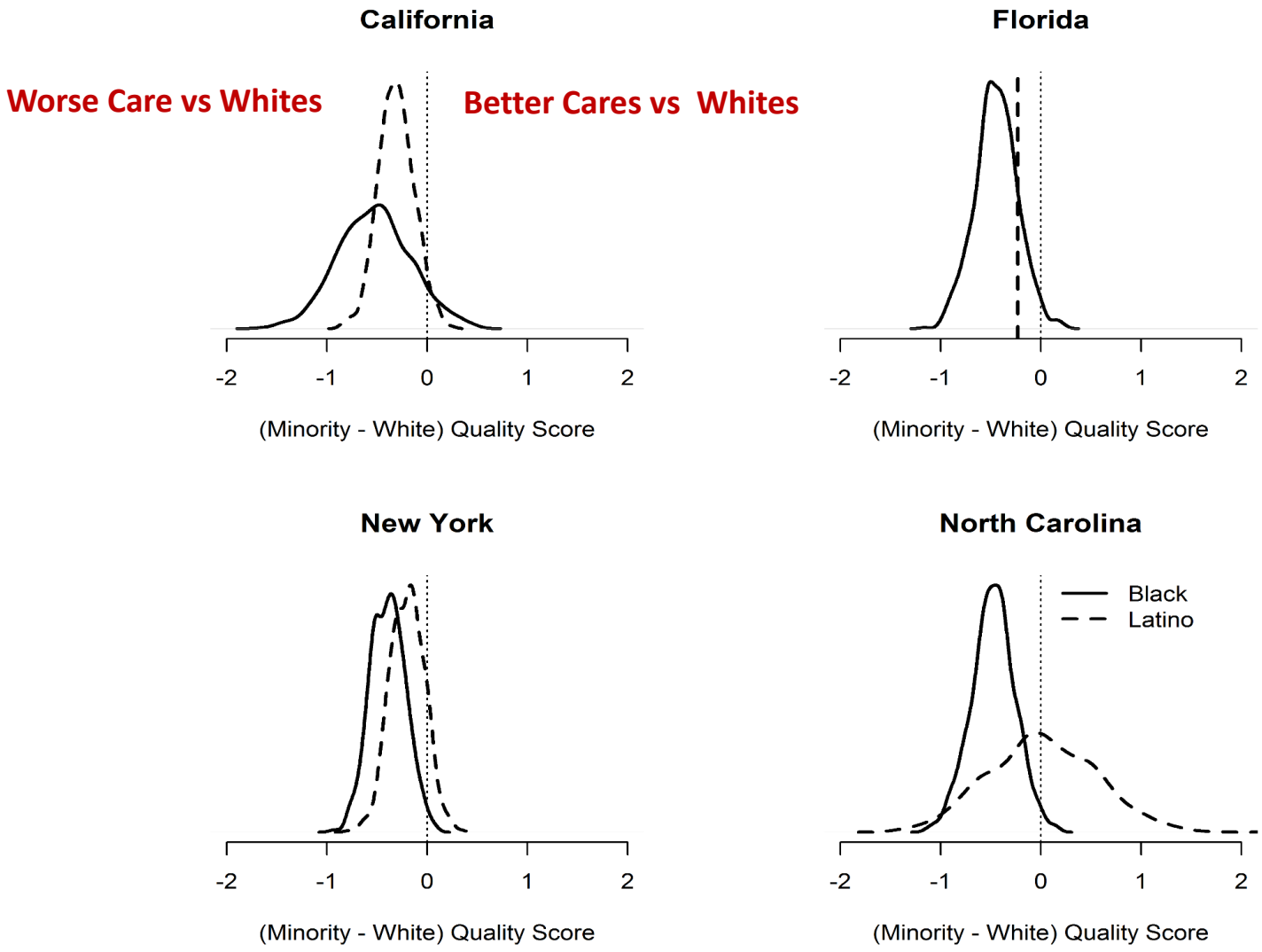
Remarks

- If strong correlation between outcomes, univariate approach results in less efficient estimates
- Higher efficiency gains realized when the two outcomes share different sets of covariates
- When missing data, can directly maximize the likelihood for the latent variable approach owing to conditional independence assumption
- Many situations with more than two outcomes
  - Latent variable approach is easily extended to several continuous and binary outcomes; not true for factorization approach.

Concluding Remarks

- Risk adjustment for causal inference
  - Must assess validity of assumptions
  - Must assess robustness of conclusions to reasonable departures from assumptions
  - High-dimensional data: parametric assumptions difficult to verify and more uncertainty
- Clustering
  - Understand question being posed
  - Is variability a nuisance parameter or is it of policy-interest?
- Multiple outcomes
  - Are outcomes manifest variables to inform about a latent variable
  - Do covariates affect outcomes differently

Schizophrenia Example: County Disparities



Key Principles

- Avoid **strong** parametric assumptions
  - In setting with many confounders, very likely to get the model really wrong
- Adhere to causal inference **assumptions**
  - Validate assumptions
  - Assess robustness to reasonable departures from assumptions
- Adopt a **design-based** approach
  - Separate treatment from outcome during the modeling process
- Reflect all **uncertainty** in estimates

Thank you