Bias Assessment of Observational Studies

THOMAS G. STEWART, PHD

Note: These slides will be posted on my faculty page within days of the conference (after correcting errors and incorporating suggestions made by you).

See tgstewart.xyz

Who am I?

- Assistant Professor of Biostatistics at Vanderbilt University Medical Center
- Collaborative experience with a number of national and international registries, most notably
  - HCV-TARGET
  - AHSQC
- Collaborative experience with state and national claims and administrative data

Why observational studies?

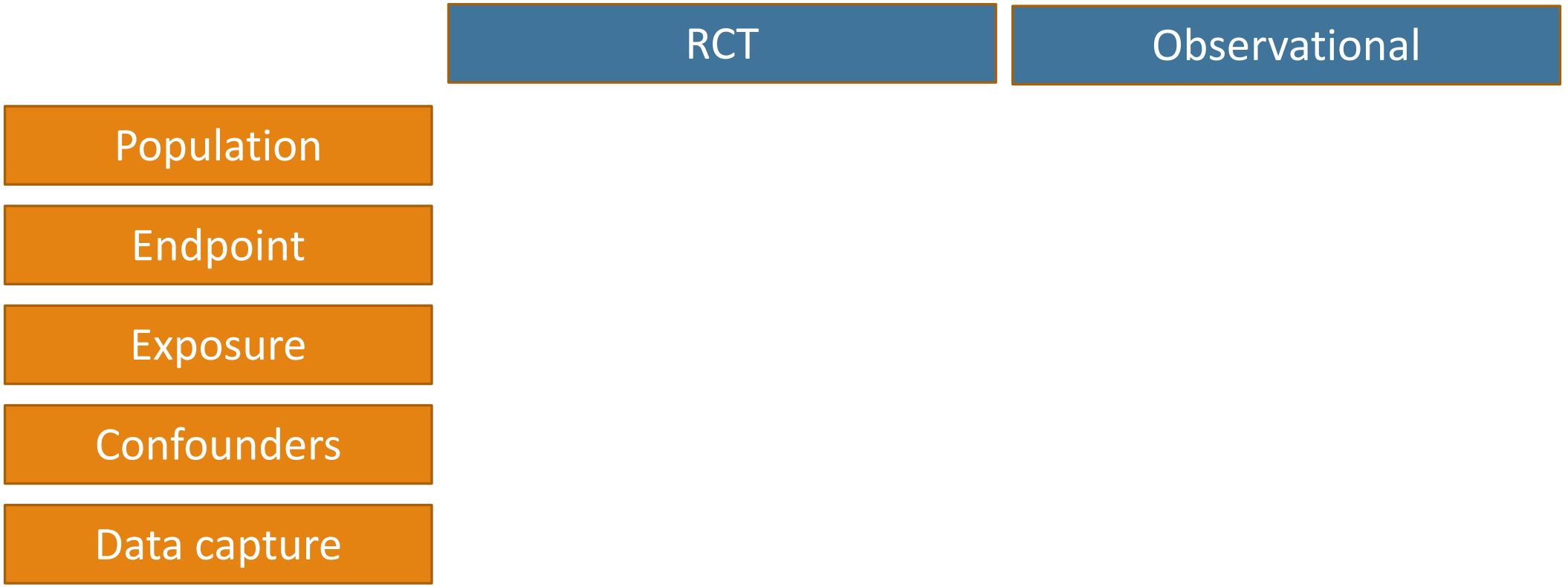
In the past

- Ethical considerations:
  - Can’t randomize individuals to smoke/not-smoke

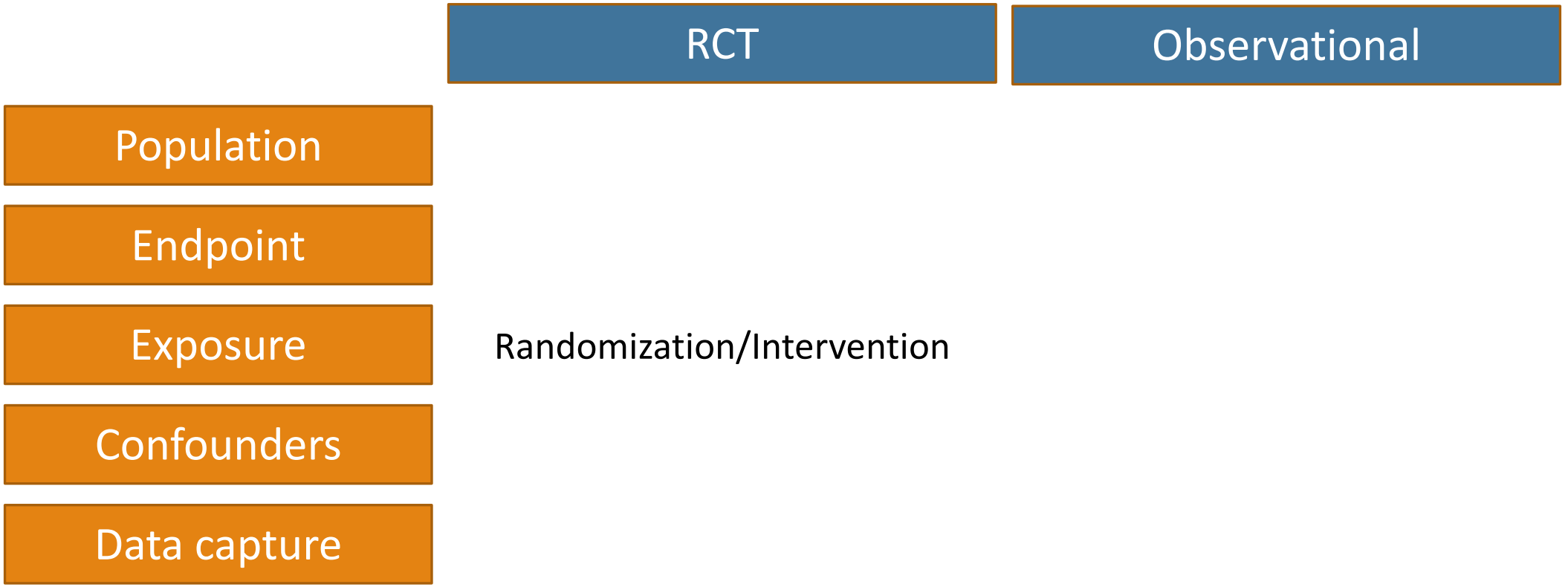
More recently

- Economic considerations:
  - The data is there … Why not use it?

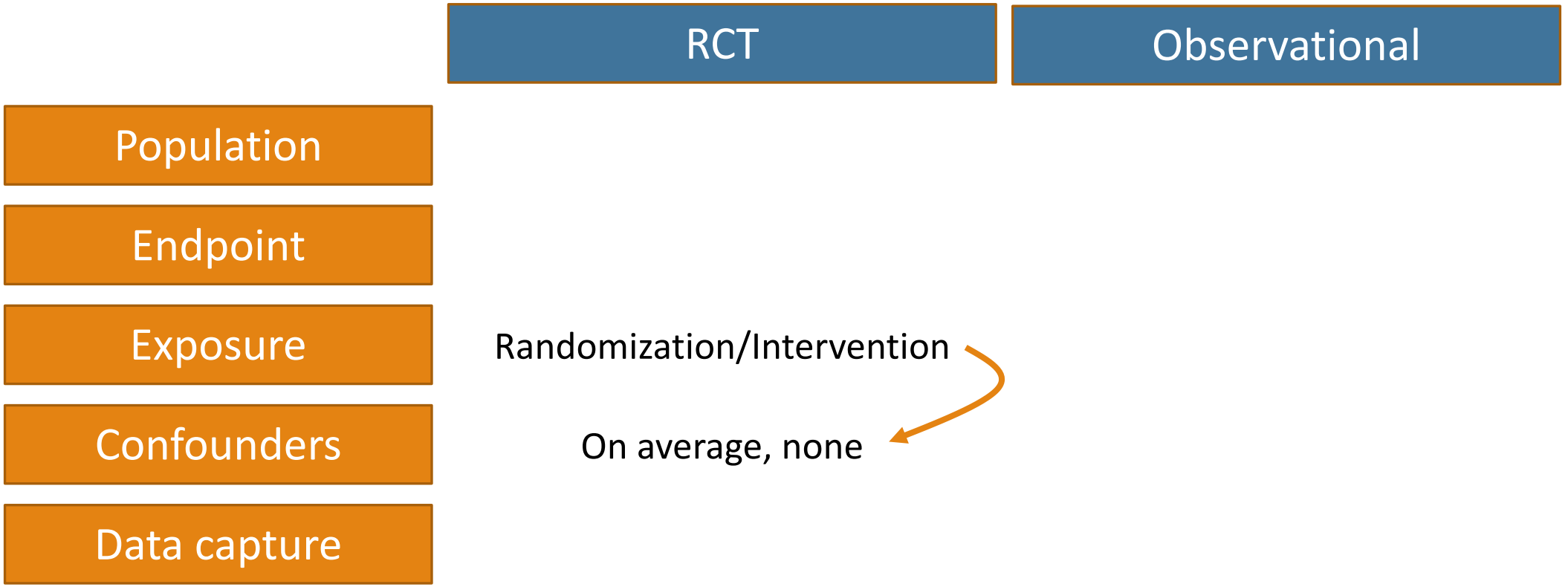
What can go wrong in an observational study?



What can go wrong in an observational study?



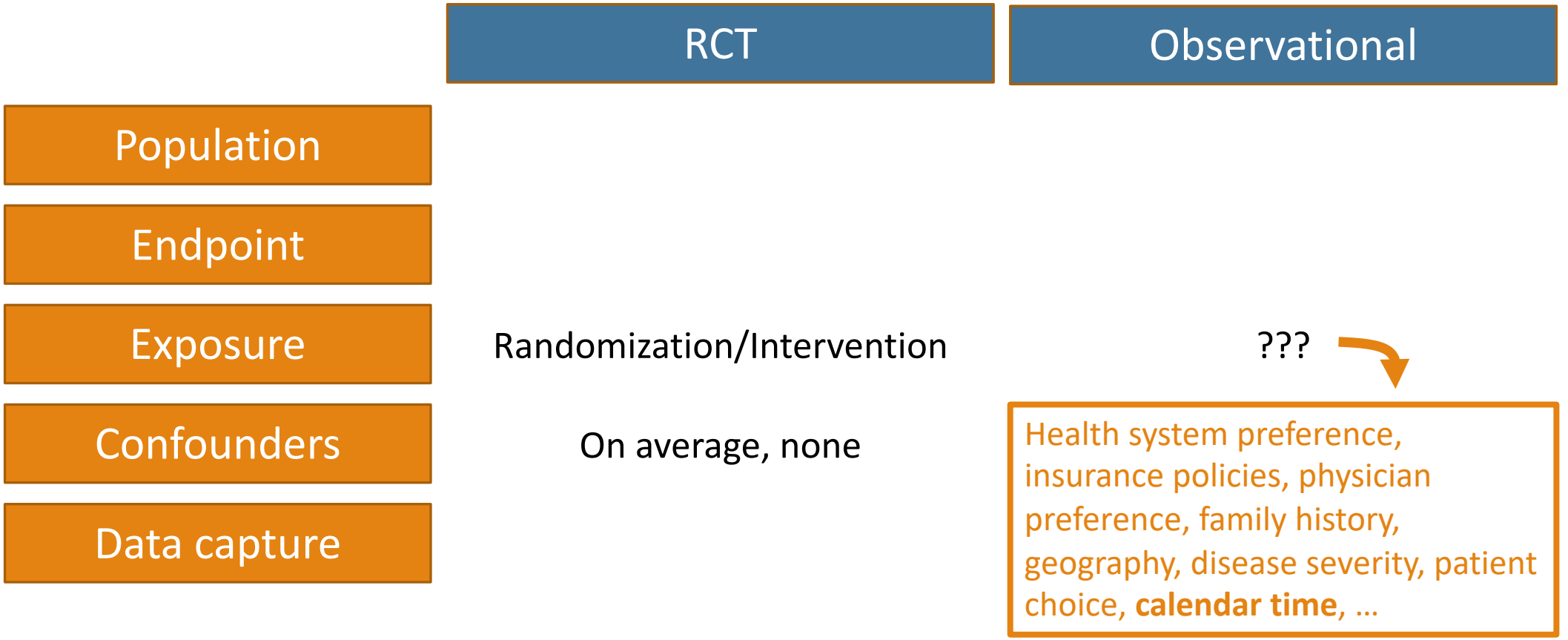
What can go wrong in an observational study?



What can go wrong in an observational study?

	RCT	Observational
Population		
Endpoint		
Exposure	Randomization/Intervention	???
Confounders	On average, none	
Data capture		

What can go wrong in an observational study?



Why care about confounders?

Take it to the extreme

If treatment A was only given to patients with severe disease and treatment B was only given to patients with mild disease,

What would one learn from a comparison of outcomes between treatment groups?

Why care about confounders?

Less extreme

If treatment A was **mostly** given to patients with severe disease and treatment B was **mostly** given to patients with mild disease,

What would one learn from a comparison of outcomes between treatment groups?

Why care about confounders?

Balance

One of the reasons randomization is so important is that it creates, **on average**, comparison groups that are balanced in terms of disease severity and other confounders.

What would one learn from a comparison of outcomes between treatment groups created by randomization?

What can go wrong in an observational study?

	RCT	Observational
Population		
Endpoint		
Exposure	Randomization/Intervention	???
Confounders	On average, none	**Failure to address confounding**
Data capture		

Disclaimer about examples:

I am a statistician, not a clinician.

While I can speak abstractly about the concept of “confounding” and “treatment by indication bias”, I am hesitant to editorialize about specific clinical examples outside the areas of my collaborative experience.

As such, the examples in this workshop of studies that may be biased are drawn from published criticism.

Example:

JAMA | Original Investigation | CARING FOR THE CRITICALLY ILL PATIENT

Association Between Tracheal Intubation During Pediatric In-Hospital Cardiac Arrest and Survival

Lars W. Andersen, MD, MPH; Tia T. Raymond, MD; Robert A. Berg, MD; Vinay M. Nadkarni, MD; Anne V. Grossestreuer, PhD; Tobias Kurth, MD, ScD; Michael W. Donnino, MD; for the American Heart Association's Get With The Guidelines–Resuscitation Investigators

JAMA. 2016;316(17):1786-1797. doi:10.1001/jama.2016.14486

Criticism:

JAMA Guide to Statistics and Methods

Confounding by Indication in Clinical Research

Demetrios N. Kyriacou, MD, PhD; Roger J. Lewis, MD, PhD

JAMA November 1, 2016 Volume 316, Number 17

Criticism:

JAMA Guide to Statistics and Methods

Confounding by Indication in Clinical Research

Demetrios N. Kyriacou, MD, PhD; Roger J. Lewis, MD, PhD

JAMA November 1, 2016 Volume 316, Number 17

The nonrandomized assessment of tracheal intubation vs bag-valve-mask ventilation for pediatric cardiopulmonary arrest reported by Andersen et al[2] in the November 1, 2016, issue of *JAMA* is likely to be complicated by confounding by indication. Clinical conditions (eg, asthma, cystic fibrosis, and upper airway obstruction) existing before and during a patient's cardiopulmonary resuscitation will both affect the patient's outcome and influence the type of airway management.[2] In other words, it is likely that children with more severe disease and worse overall prognosis for survival had a greater probability to be intubated.[2] This pos-

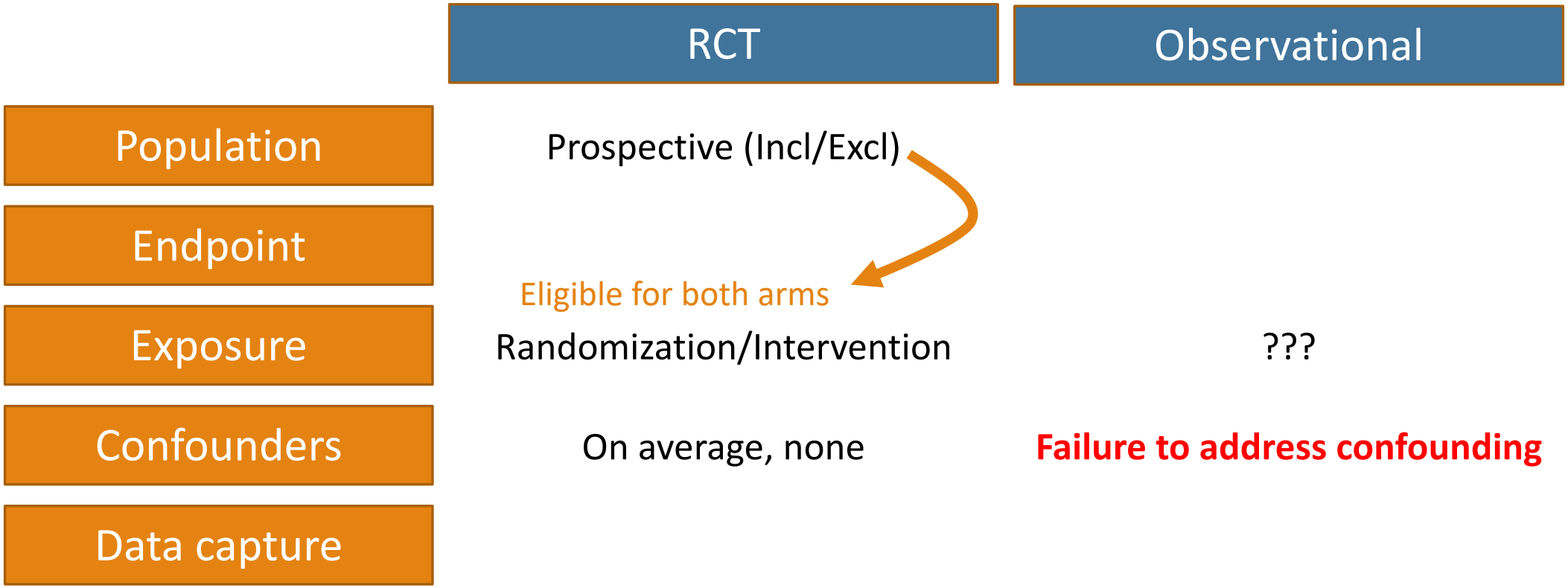
What can go wrong in an observational study?

	RCT	Observational
Population		
Endpoint		
Exposure	Randomization/Intervention	???
Confounders	On average, none	**Failure to address confounding**
Data capture		

What can go wrong in an observational study?

	RCT	Observational
Population	Prospective (Incl/Excl)	
Endpoint		
Exposure	Randomization/Intervention	???
Confounders	On average, none	**Failure to address confounding**
Data capture		

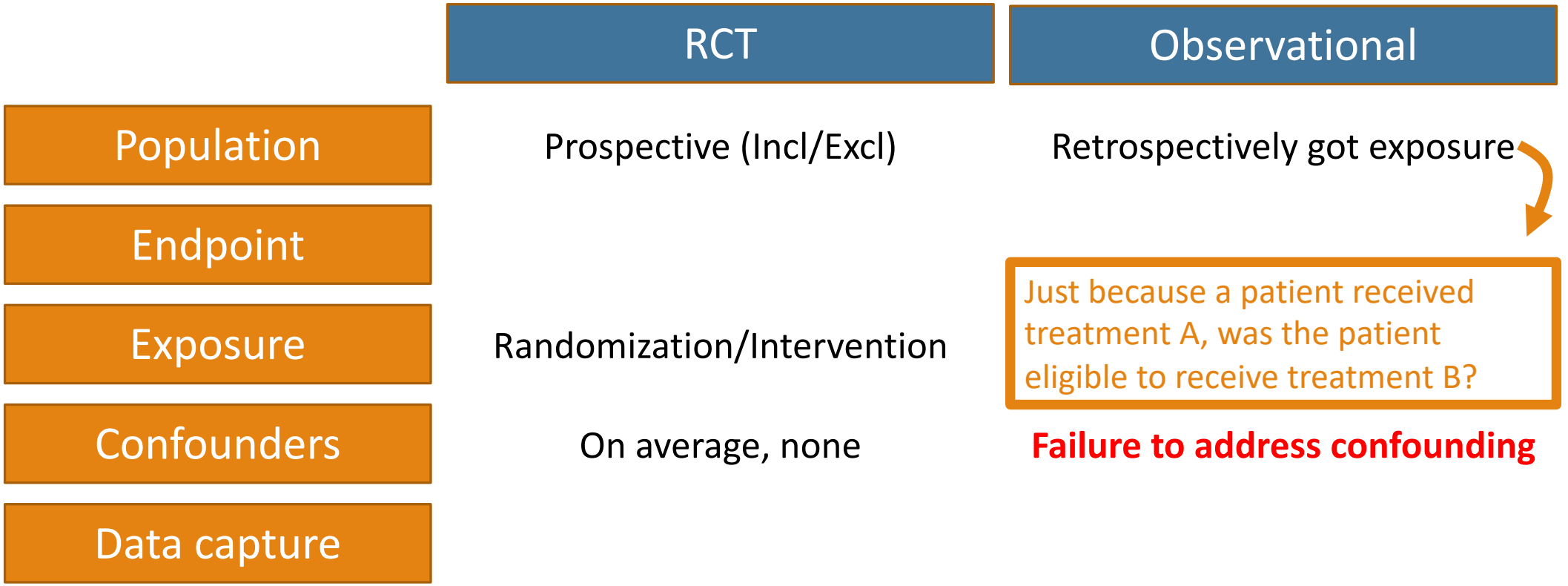
What can go wrong in an observational study?



What can go wrong in an observational study?

	RCT	Observational
Population	Prospective (Incl/Excl)	Retrospectively got exposure
Endpoint		
Exposure	Randomization/Intervention	???
Confounders	On average, none	**Failure to address confounding**
Data capture		

What can go wrong in an observational study?



Why care about eligibility?

Take it to the extreme

If patients in treatment group A were only eligible to receive treatment A, and patients in treatment group B was only eligible to receive treatment,

What would one learn from a comparison of outcomes between treatment groups?

Why care about eligibility?

Equipoise

One of the reasons an interventional comparison is meaningful is that all patients could have potentially have received both treatments. There was some degree of equipoise of the treatments for the study subjects.

What would one learn from a comparison of outcomes between treatment groups created by randomization?

What can go wrong in an observational study?

	RCT	Observational
Population	Prospective (Incl/Excl)	Retrospectively got exposure
Endpoint		
Exposure	Randomization/Intervention	**Failure to consider equipoise**
Confounders	On average, none	**Failure to address confounding**
Data capture		

Example from collaborative experience:

Open repair vs Laparoscopic repair of ventral hernias

Among surgeons in a national registry of hernia repairs, certain complex hernias are only repaired using an open surgical approach.

To include patients with this specific type of hernia in a comparison of open and laparoscopic approaches would be a mistake because this group of patients is not eligible to receive both surgical approaches.

What can go wrong in an observational study?

	RCT	Observational
Population	Prospective (Incl/Excl)	Retrospectively got exposure
Endpoint	At risk for outcomes	
Exposure	Randomization/Intervention	???
Confounders	On average, none	**Failure to address confounding**
Data capture		

What can go wrong in an observational study?

	RCT	Observational
Population	Prospective (Incl/Excl)	Retrospectively got exposure
Endpoint	Prospectively at risk	At risk for outcomes?
Exposure	Randomization/Intervention	???
Confounders	On average, none	**Failure to address confounding**
Data capture		

Why care about at-risk?

Take it to the extreme

Are males reasonable controls for studies of treatment for ovarian cancer?

What can go wrong in an observational study?

	RCT	Observational
Population	Prospective (Incl/Excl)	Retrospectively got exposure
Endpoint	Prospectively at risk	**Inappropriate controls**
Exposure	Randomization/Intervention	**Failure to consider equipoise**
Confounders	On average, none	**Failure to address confounding**
Data capture		

Editorial:

Selection of appropriate controls is more difficult than most than most of us (including me) realize.

Selection of appropriate controls is especially difficult when the control arm is passive, i.e., defined by NOT receiving an intervention.

Editorial:

Selection of appropriate controls is more difficult than most than most of us (including me) realize.

Selection of appropriate controls is especially difficult when the control arm is passive, i.e., defined by NOT receiving an intervention.

> Why? The decision to not actively intervene is generally not recorded in the medical record. It is also correlated with patient priorities and the desire for aggressive care.

Editorial:

Selection of appropriate controls is more difficult than most than most of us (including me) realize.

Selection of appropriate controls is especially difficult when the control arm is passive, i.e., defined by NOT receiving an intervention.

Why? The decision to not actively intervene is generally not recorded in the medical record. It is also correlated with patient priorities and the desire for aggressive care.

Are these passive controls really eligible to receive both treatment arms? What is the “entry date” for passive controls?

What can go wrong in an observational study?

	RCT	Observational
Population	Prospective (Incl/Excl)	Retrospectively got exposure
Endpoint	Prospectively at risk	**Inappropriate controls**
Exposure	Randomization/Intervention	**Failure to consider equipoise**
Confounders	On average, none	**Failure to address confounding**
Data capture		

What can go wrong in an observational study?

	RCT	Observational
Population	Prospective (Incl/Excl)	Retrospectively got exposure
Endpoint	Prospectively at risk	**Inappropriate controls**
Exposure	Randomization/Intervention	**Failure to consider equipoise**
Confounders	On average, none	**Failure to address confounding**
Data capture	Identical follow-up protocol	

What can go wrong in an observational study?

	RCT	Observational
Population	Prospective (Incl/Excl)	Retrospectively got exposure
Endpoint	Prospectively at risk	**Inappropriate controls**
Exposure	Randomization/Intervention	**Failure to consider equipoise**
Confounders	On average, none	**Failure to address confounding**
Data capture	Identical follow-up protocol	???, EHR, Registry,

Why care about follow-up?

Take it to the extreme

If patients in treatment group A interact with their health-care providers weekly, but patients in treatment group B interact with their care providers only as needed,

Are treatment groups equally likely to capture study endpoints?

What can go wrong in an observational study?

	RCT	Observational
Population	Prospective (Incl/Excl)	Retrospectively got exposure
Endpoint	Prospectively at risk	**Inappropriate controls**
Exposure	Randomization/Intervention	**Failure to consider equipoise**
Confounders	On average, none	**Failure to address confounding**
Data capture	Identical follow-up protocol	**Inconsistent follow-up**

Ways to address possible pit-falls in observational studies

	Pit-fall	Solution
Population	Retrospectively got exposure	
Endpoint	**Inappropriate controls**	
Exposure	**Failure to consider equipoise**	
Confounders	**Failure to address confounding**	
Data capture	**Inconsistent follow-up**	

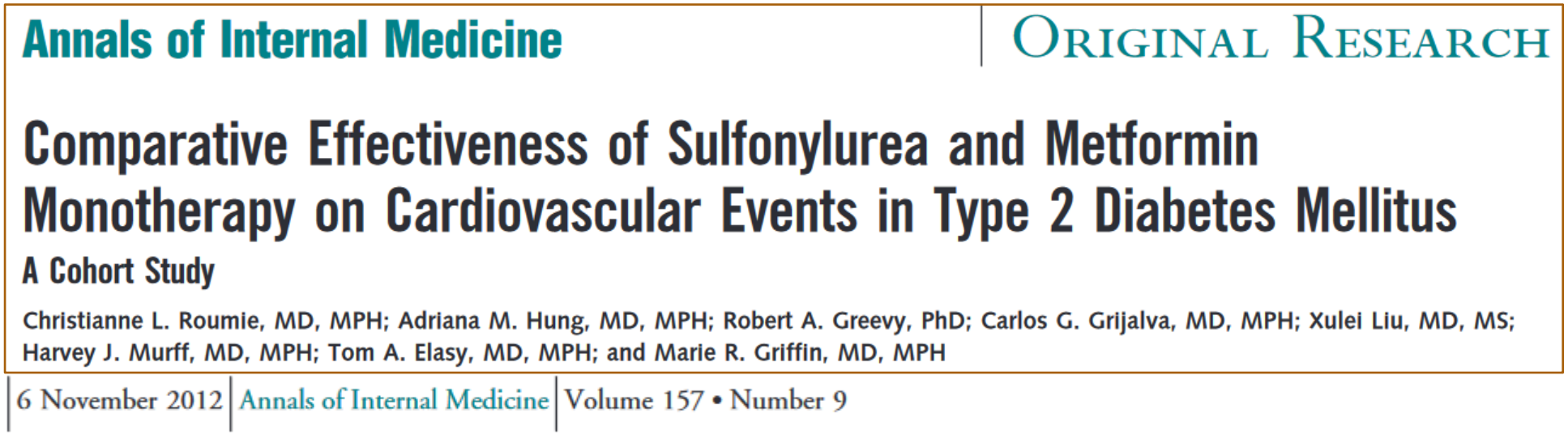
Design Choices vs Statistical Methods

Not every bias can be corrected with a statistical method. Some bias is only controlled with appropriate study design.

Ways to address possible pit-falls in observational studies

	Pit-fall	Solution
Population	Retrospectively got exposure	
Endpoint	**Inappropriate controls**	
Exposure	**Failure to consider equipoise**	
Confounders	**Failure to address confounding**	**Regression, propensity score methods**
Data capture	**Inconsistent follow-up**	

Matching (The beautiful table 1 approach)



Annals of Internal Medicine | ORIGINAL RESEARCH

Comparative Effectiveness of Sulfonylurea and Metformin Monotherapy on Cardiovascular Events in Type 2 Diabetes Mellitus

A Cohort Study

Christianne L. Roumie, MD, MPH; Adriana M. Hung, MD, MPH; Robert A. Greevy, PhD; Carlos G. Grijalva, MD, MPH; Xulei Liu, MD, MS; Harvey J. Murff, MD, MPH; Tom A. Elasy, MD, MPH; and Marie R. Griffin, MD, MPH

6 November 2012 | Annals of Internal Medicine | Volume 157 • Number 9

Table 1. **Patient Characteristics in Full and Propensity Score–Matched Cohorts, by New Exposure to Metformin or Sulfonylureas**

Characteristic	Full Cohort			Propensity Score–Matched Cohort		
	Metformin (n = 155 025)	Sulfonylureas (n = 98 665)	Standardized Difference*†	Metformin (n = 80 648)	Sulfonylureas (n = 80 648)	Standardized Difference*‡
Median age (IQR), y	62 (56–71)	67 (57–76)	0.33	65 (57–74)	64 (56–74)	0.03‡
Men, %	95	97	0.12	97	97	0.01
Race, %						
White	74	75	0.04	75	75	0.01
Black	12	13	0.04	13	13	0.00
Hispanic/other	6	6	0.03	6	6	0.00
Available§	91	95	0.13	94	94	0.01
HbA_{1c}						
Median level (IQR), %	7.0 (6.4–7.8)	7.3 (6.6–8.2)	0.17	7.2 (6.5–8.2)	7.2 (6.6–8.2)	0.02
Available§	67	61	0.14	63	63	0.01
LDL cholesterol						

Table 1. **Patient Characteristics in Full and Propensity Score–Matched Cohorts, by New Exposure to Metformin or Sulfonylureas**

Characteristic	Full Cohort			Propensity Score–Matched Cohort		
	Metformin (n = 155 025)	Sulfonylureas (n = 98 665)	Standardized Difference*†	Metformin (n = 80 648)	Sulfonylureas (n = 80 648)	Standardized Difference*‡
Median age (IQR), y	62 (56–71)	67 (57–76)	0.33	65 (57–74)	64 (56–74)	0.03‡
Men, %	95	97	0.12	97	97	0.01
Race, %						
White	74	75		75	75	0.01
Black	12					
Hispanic/other	6					
Available§	91					
HbA_{1c}						
Median level (IQR), %	7.0 (6.4–7.8)	7.3 (6.6–8.2)	0.17	7.2 (6.5–8.2)	7.2 (6.6–8.2)	0.02
Available§	67	61	0.14	63	63	0.01
LDL cholesterol						

Mean differences in observed covariates are minimized in the matched cohort.

Table 1. Patient Characteristics in Full and Propensity Score–Matched Cohorts, by New Exposure to Metformin or Sulfonylureas

Characteristic	Full Cohort			Propensity Score–Matched Cohort		
	Metformin (*n* = 155 025)	Sulfonylureas (*n* = 98 665)	Standardized Difference*†	Metformin (*n* = 80 648)	Sulfonylureas (*n* = 80 648)	Standardized Difference*‡
Median age (IQR), y	62 (56–71)	67 (57–76)	0.33	65 (57–74)	64 (56–74)	0.03‡
Men, %	95	97	0.12	97	97	0.01
Race, %						
White	74	75	0.01	75	75	0.01
Black	12					
Hispanic/other	6					
Available§	91					
HbA_{1c}						
Median level (IQR), %	7.0 (6.4–7.8)					
Available§	67					
LDL cholesterol						

Matching selects pairs (or groups) of subjects that are similar.

Some subjects are discarded from the analysis.

Table 1. Patient Characteristics in Full and Propensity Score–Matched Cohorts, by New Exposure to Metformin or Sulfonylureas

Characteristic	Full Cohort			Propensity Score–Matched Cohort		
	Metformin (n = 155 025)	Sulfonylureas (n = 98 665)	Standardized Difference*†	Metformin (n = 80 648)	Sulfonylureas (n = 80 648)	Standardized Difference*‡
Median age (IQR), y	62 (56–71)	67 (57–76)	0.33	65 (57–74)	64 (56–74)	0.03‡
Men, %	95	97	0.12	97	97	0.01
Race, %						
White	74					.01
Black	12					.00
Hispanic/other	6					.00
Available§	91					.01
HbA_{1c}						
Median level (IQR), %	7.0 (6.4–					.02
Available§	67					.01
LDL cholesterol						

A common approach to identify similar subjects between each treatment group is to use the propensity score.

The propensity score is a summary measure of all important covariates. It is an estimate of the probability of receiving one of the treatments.

Regression (Use all the data approach)

$$E[\,\text{OUTCOME} \mid -] = \beta_0 + \beta_1\, TREATMENT\ A + \ \beta_2\, MALE + \beta_3\, AGE + \beta_4\, AGE^2 + \ \ldots$$

Regression (Use all the data approach)

$$E[\,\text{OUTCOME} \mid -] = \beta_0 + \underbrace{\beta_1\, TREATMENT\ A +}_{\text{Association of interest}}\ \underbrace{\beta_2\, MALE + \beta_3\, AGE + \beta_4\, AGE^2 +}_{\text{Potential confounding variables}}\ \ldots$$

Regression (Use all the data approach)

$$E[\,\text{OUTCOME} \mid -] = \beta_0 + \beta_1 \quad \textbf{METFORMIN} \quad + \; \beta_2\, MALE + \beta_3\, AGE + \beta_4\, AGE^2 + \; \ldots$$

Association of interest

Potential confounding variables

Regression and Propensity Score methods (Belt and suspenders approaches)

$$E[\,\text{OUTCOME} \mid -] = \beta_0 + \beta_1\,\text{TRT A} + \beta_2\,\text{MALE} + \beta_3\,\text{AGE} + \beta_4\,\text{AGE}^2 + \ldots + RCS(PS)$$

Association of interest

Potential confounding variables

Flexible association with propensity score

Regression and Propensity Score methods (Belt and suspenders approaches)

1. Regression with propensity score as flexible covariate
2. Propensity score matching then regression
3. Propensity score weighting with regression

Ways to address possible pit-falls in observational studies

	Pit-fall	Solution
Population	Retrospectively got exposure	
Endpoint	**Inappropriate controls**	
Exposure	**Failure to consider equipoise**	
Confounders	**Failure to address confounding**	**Regression, propensity score methods, OTHER**
Data capture	**Inconsistent follow-up**	

Ways to address possible pit-falls in observational studies

	Pit-fall	Solution
Population	Retrospectively got exposure	
Endpoint	**Inappropriate controls**	
Exposure	**Failure to consider equipoise**	**Covariate overlap plots**
Confounders	**Failure to address confounding**	**Regression, propensity score methods**
Data capture	**Inconsistent follow-up**	

Ways to address possible pit-falls in observational studies

	Pit-fall	Solution
Population	Retrospectively got exposure	
Endpoint	**Inappropriate controls**	**Reasonable incl/excl**
Exposure	**Failure to consider equipoise**	**Covariate overlap plots**
Confounders	**Failure to address confounding**	**Regression, propensity score methods**
Data capture	**Inconsistent follow-up**	

Editorial:

Selection of appropriate controls is more difficult than most than most of us (including me) realize.

Selection of appropriate controls is especially difficult when the control arm is passive, i.e., defined by NOT receiving an intervention.

One solution when the control arm is passive is to use time-varying covariate methods instead of awkward definitions for controls.

Ways to address possible pit-falls in observational studies

	Pit-fall	Solution
Population	Retrospectively got exposure	
Endpoint	**Inappropriate controls**	**Reasonable incl/excl**
Exposure	**Failure to consider equipoise**	**Covariate overlap plots**
Confounders	**Failure to address confounding**	**Regression, propensity score methods**
Data capture	**Inconsistent follow-up**	**Design**

Questions?