Targeted register analyses

PhD short course

An introduction to the Causal Roadmap

Andrew Mertens
University of California, Berkeley, Division of Biostatistics

Presentation outline

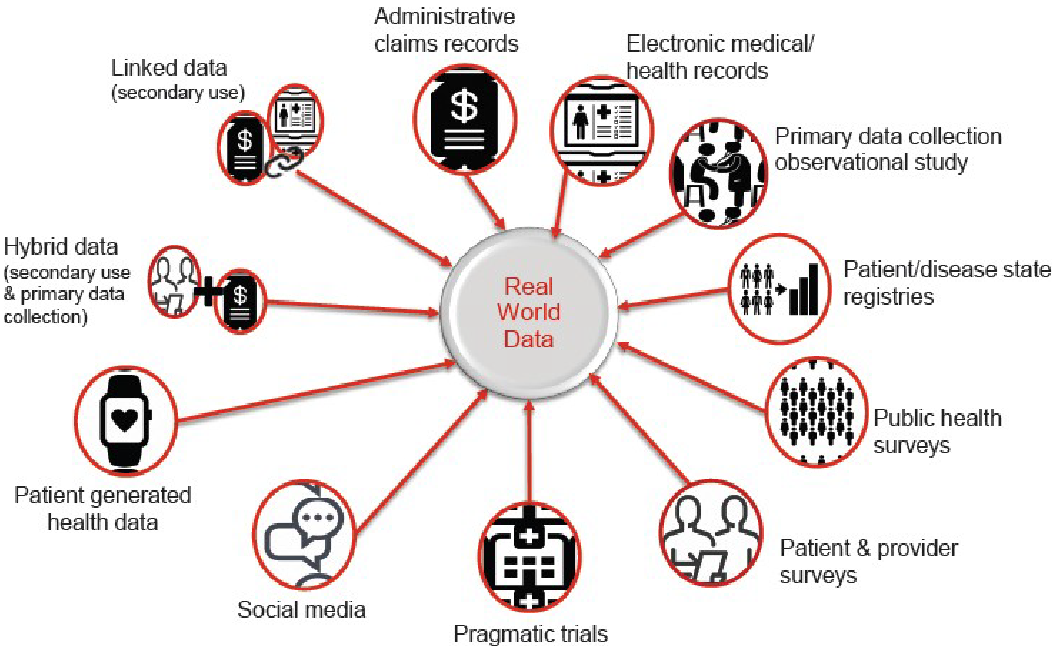
1. Background
2. Causal Roadmap:
    - I. Target causal parameter
    - II. Observed Data
    - III. Causal Model and Identification.
    - IV. Estimation
    - V. Interpretation
3. Conclusion

Presentation outline

1. Background
2. Causal Roadmap:
   - I. Target causal parameter
   - II. Observed Data
   - III. Causal Model and Identification.
   - IV. Estimation
   - V. Interpretation
3. Conclusion

A landscape of opportunities from electronic health registries and beyond

- Today’s data ecosystem: Rich and diverse data sources
  - Registries
  - Electronic health records
  - Clinical trials
- Ability to link and combine
- Powerful new analytic tools
  - Machine learning
  - Computing power

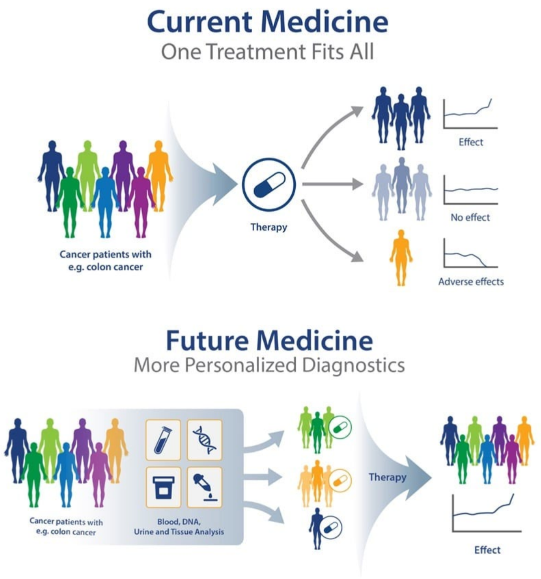

Administrative claims records
Electronic medical/ health records
Linked data (secondary use)
Primary data collection observational study
Hybrid data (secondary use & primary data collection)
Real World Data
Patient/disease state registries
Public health surveys
Patient generated health data
Social media
Pragmatic trials
Patient & provider surveys

https://www.nap.edu/read/25352/chapter/7#73

The Promise:

Big data and statistical advancements can provide novel insights for how best to treat patients and deliver care

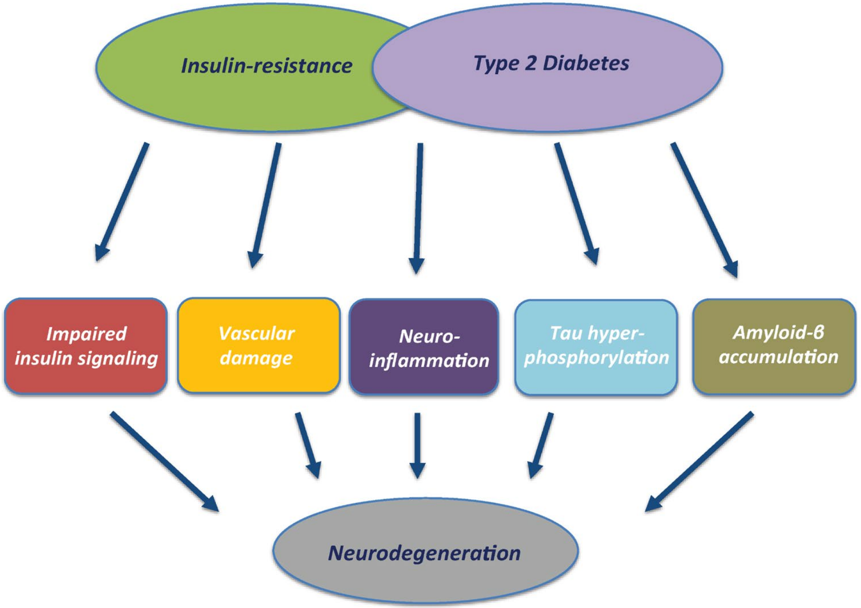
- Real-world comparative effectiveness
  - Long-term cumulative effects
  - Unexpected benefits and risks
- Transport of effects
  - From trials to new populations
- Personalized medicine
  - Which types of patients respond best to which treatments
  - When to initiate/modify/intensify therapies


Current Medicine
One Treatment Fits All
Cancer patients with e.g. colon cancer
Therapy
Effect
No effect
Adverse effects
Future Medicine
More Personalized Diagnostics
Cancer patients with e.g. colon cancer
Blood, DNA, Urine and Tissue Analysis
Therapy
Effect

https://blog.crownbio.com/pdx-personalized-medicine

Case Study Background

- Increasing evidence has linked Type 2 diabetes mellitus (T2DM) to dementia
- Diabetes medication may decrease dementia risk
- GLP-1 receptor agonists (GLP-1RA), a second-line treatment, may be particularly neuroprotective
- Studies to date have been limited
    - Not designed to answer clearly interpretable causal questions
    - Fail to fully adjust for measured confounders, including time-varying confounders
    - Use of statistical methods that require unrealistic assumptions



Tumminia et. al 2018

Objective: Illustrate application of the causal roadmap

- **Specific scientific question:** What is the effect of cumulative exposure to GLP-1RAs vs. active comparators (other second line drugs like SLGT2i or DPP4) on dementia risk?
- Today, we will step through the process of going from this scientific question to:
  - A well-defined causal target parameter
  - A well-defined statistical target parameter
  - A clear understanding of the assumptions needed for statistical parameter to have causal interpretation
  - A fully pre-specified statistical estimator
- Example application: analysis of **Danish Registry Data**

We need a roadmap to help drive this process!

  - “estimand” = target causal question/parameter

Potential in EHR data, but easy to get lost…

- Example: In real world practice, which 2nd diabetes drug (GLP1 or SGLT2) is better for reducing dementia risk?

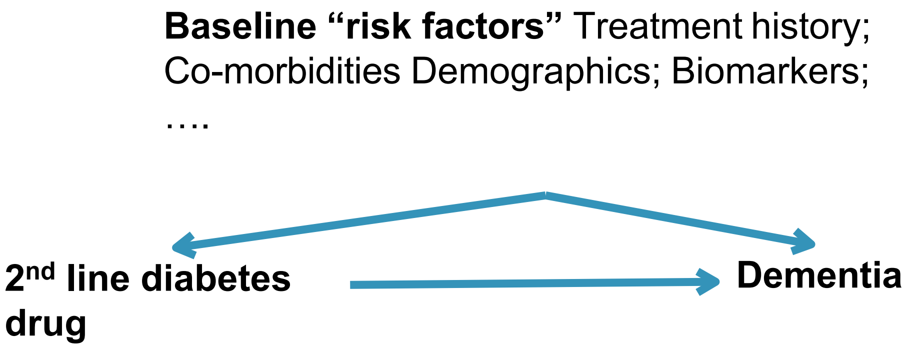
- **Defining the question**
  1. Target population (inclusion criteria)
  2. Baseline timepoint
  3. Definition of outcome
     - Ex. Dementia cumulative incidence over 5 years

An apparently simple causal question: "Average Treatment Effect"

How would dementia risk at 5 years differ if **all** were treated with GLP1 vs. **all** were treated with other 2nd line drugs?

The challenges of “real-world” data

- Classic confounding: Patients who received GLP1 differ from those who didn’t received GLP1
  - In ways that may affect (or be correlated with) risk of dementia



Baseline “risk factors” Treatment history; Co-morbidities Demographics; Biomarkers; ….

2nd line diabetes drug

Dementia

- **Which variables** should we adjust for?
  - A huge potential adjustment set

Randomized Trials face similiar challenges

- Intention to treat analyses: Still many choices!
  - Adjustment can improve precision **but** must be prespecified
  - Rigid pre-specified approaches can fail to perform

- And much we cannot control… (much about trials is observational!)
  - Adherence, Patient drop out, Treatment modifications, Protocol deviations
  - **Intent to Treat Analyses may not be interpretable or informative**

We can **and should** use our trial data to go after more relevant questions

- Ex. Per-protocol analyses
- Ex. Treatment effect heterogeneity

How to adjust for differences in risk?

We have a (large) set of adjustment variables… **now which estimator**?

Parametric outcome regression?

- Model specification?
  - Binary outcome? Maybe a logistic regression?
  - Time-to-event outcome with some right censoring? Maybe a Cox model?
  - Main terms? Interactions?
  - Polynomials/nonlinearities?....

Propensity scores?

  - Probability of receiving GLP1 (vs. SGLT2) given adjustment variables
- Model specification?
  - Main terms? Interactions? Etc…?
- Estimator?
  - Matching? How?
  - Inverse weighting?

Many options? How to choose?

- **Be flexible, use our knowledge, look at the data as we go?**
  - In practice -> try a bunch of models and estimators
  - Choose the approach with results that “make the most sense”

- **Perils:**
  - Misleading (under)estimate of uncertainty
    - Not accounting for model selection
    - P-hacking
  - Bias
    - Humans are good at creating narratives
    - Tend to confirm what we expect to find
    - Separating causation from association in observational data

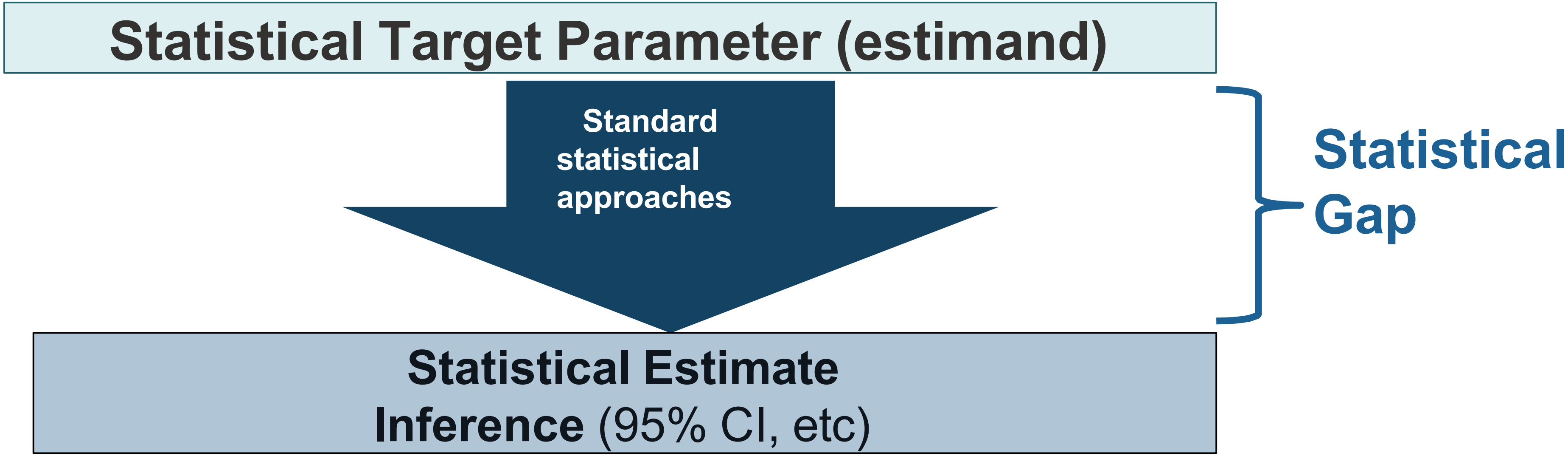
Many options? How to find our way?

- **Fully pre-specify our analysis?**
  - Before looking at the data, pre-specify our choices
    - Adjustment variables, estimator, model specification, etc…
  - Protects against “researcher degrees of freedom”
    - The pack of p-values

Perils:

- Relationships in the real world are complex
  - We don’t understand them fully
- A pre-specified parametric model may fit the data terribly
  - Again! Bias and misleading inference

Mind the Statistical Gap!



Statistical Target Parameter (estimand)

Standard statistical approaches

Statistical Gap

Statistical Estimate

Inference (95% CI, etc)

- Pre-specified parametric models: Model misspecification and bias
- Exploratory “common sense”: Underestimate uncertainty; Human bias

Forget about causal inference- **we don’t even have reliable statistical inference**!

Back to the beginning:

Did I ask the right causal question?

- What is the ideal *hypothetical* experiment (or protocol)?
  - "**Ideal**" – *what do we want to learn?*
- **Ex. What about changes in treatment over time?**
  - Patients may switch treatments, interrupt, or discontinue entirely

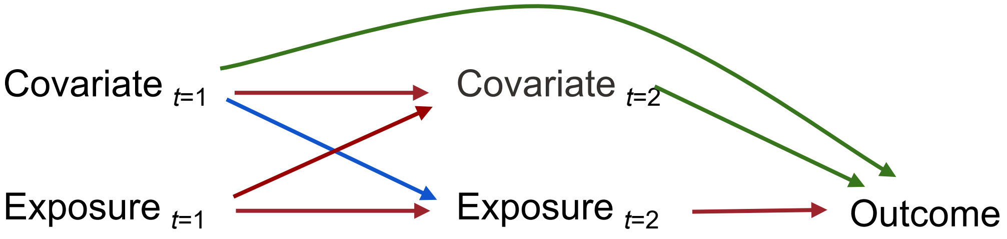
Example of ideal protocol ("longitudinal treatment regime")

- Start a specific diabetes drug (eg GLP1 or SGLT2)
- Remain on initial drug for full follow up period
- Ensure specific background/rescue therapies used (or not!)
- Prevent treatment interruptions (or not!)

The challenge of time-dependent confounding

- A time-dependent confounder is a variable that:
  - Is affected by <u>prior</u> exposure
  - Predicts <u>subsequent</u> exposure
  - Associated with / causes the outcome

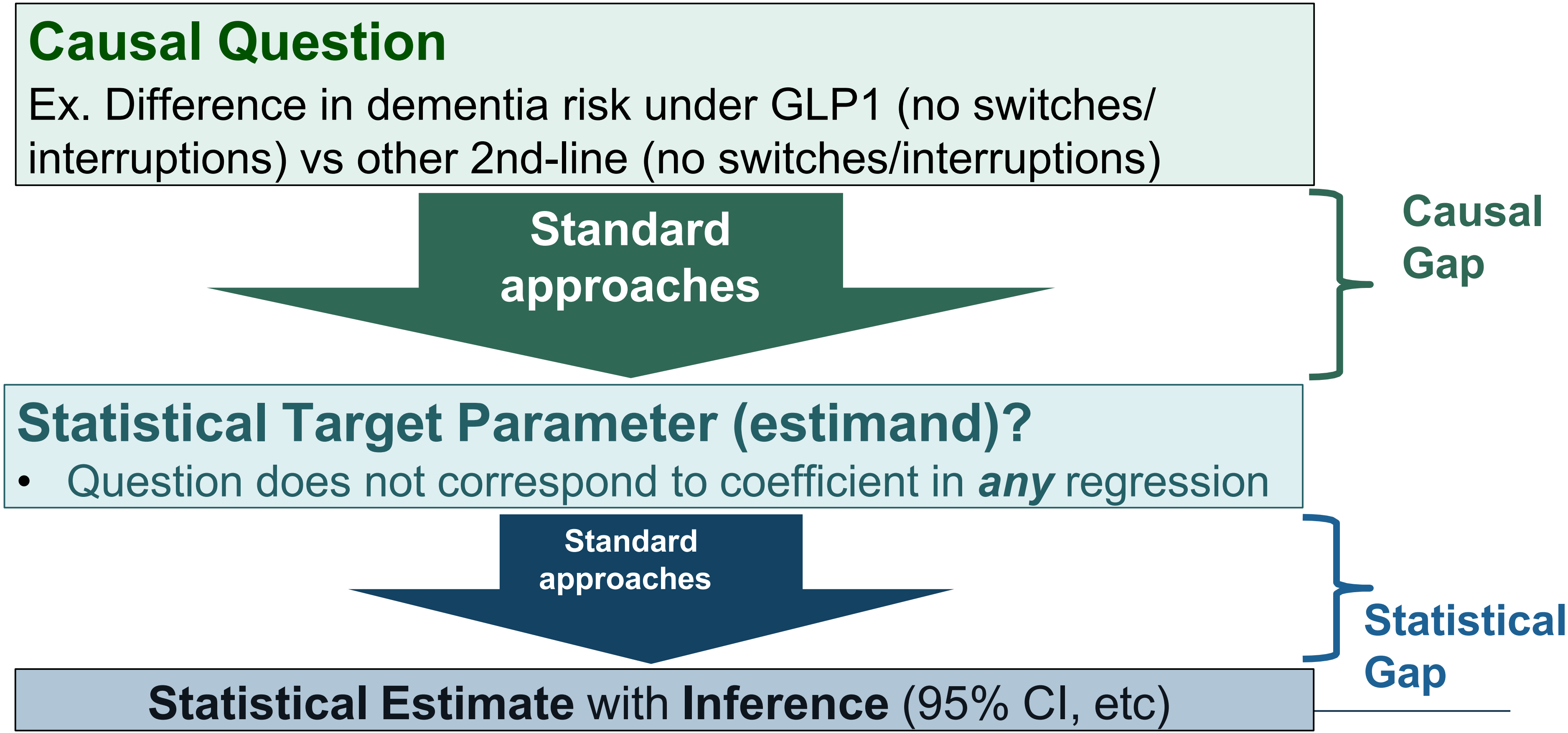
- Classical outcome regressions cannot handle this problem



- Adjusting for a time-dependent confounder changes the quantity that we estimate because it is on the causal path from the exposure to the outcome.

- If we don't adjust for a time-dependent confounder, our estimate will be confounded.

Mind the Causal Gap!



Causal Question

Ex. Difference in dementia risk under GLP1 (no switches/ interruptions) vs other 2nd-line (no switches/interruptions)

Standard approaches

Causal Gap

Statistical Target Parameter (estimand)?

- Question does not correspond to coefficient in ***any*** regression

Standard approaches

Statistical Gap

Statistical Estimate with **Inference** (95% CI, etc)

Beware of the estimator driving the question

Initial Question: **Average treatment effect**

Dementia risk if **all** were treated with GLP1 vs. **all** with other 2nd line drugs?

- With magical intervention, I pre-specify an outcome regression perfectly…
- What is my estimand….?
  - Standard practice: coefficient on treatment variable
  - Ex. Logistic regression-> **conditional odds ratio**
  - Ex. Cox model -> **conditional hazard ratio**
    - Over time- risk of event in non-comparable (selected) populations!

› ***Best*** case scenario: **answering a different (noninterpretable?) question**

Many more options for our questions…

- What is our ideal *hypothetical* experiment?
  - “Dynamic” Treatment protocols that respond to patient events

Examples of ideal protocols (“dynamic treatment regimes”)

- Strategies for initiating, intensifying, or switching therapies
  - Ex. Intensify or switch when HbA1c exceeds a certain threshold
  - Ex. Switch if experience adverse events
- Treatment assignment based on predicted patient response?
  - Using the same data to learn what type of patient will respond well…

- Again- these do not correspond to *any* classical statistical parameter

We need a roadmap!

Roadmap- Overview

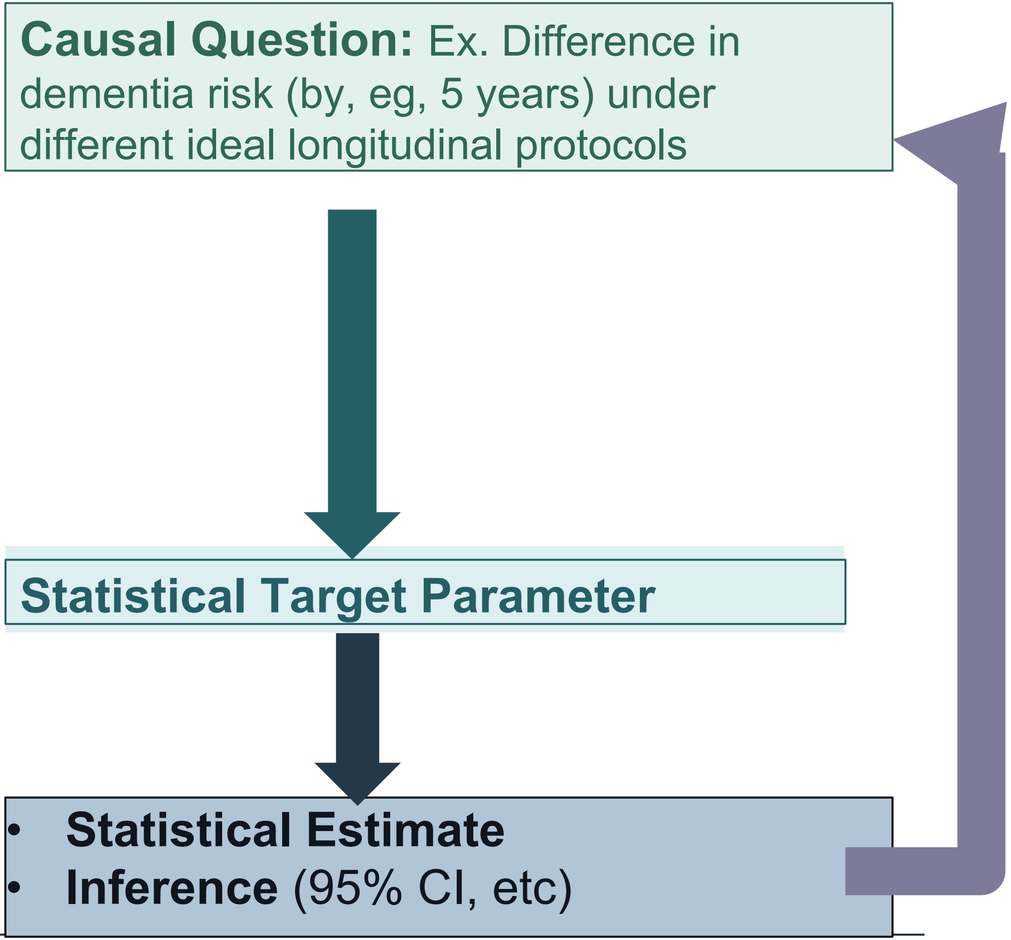
1. **Causal question**
   - Translate scientific question into causal parameter (defined in terms of counterfactual outcomes)

2. **Observed data & statistical model**
   - Model should reflect uncertainty

3. **Identify**
   - Translate causal parameter to statistical parameter under explicit causal assumptions

4. **Estimate**

5. **Interpret**



Causal Question: Ex. Difference in dementia risk (by, eg, 5 years) under different ideal longitudinal protocols

Statistical Target Parameter

- **Statistical Estimate**
- **Inference** (95% CI, etc)

Roadmap- Objectives

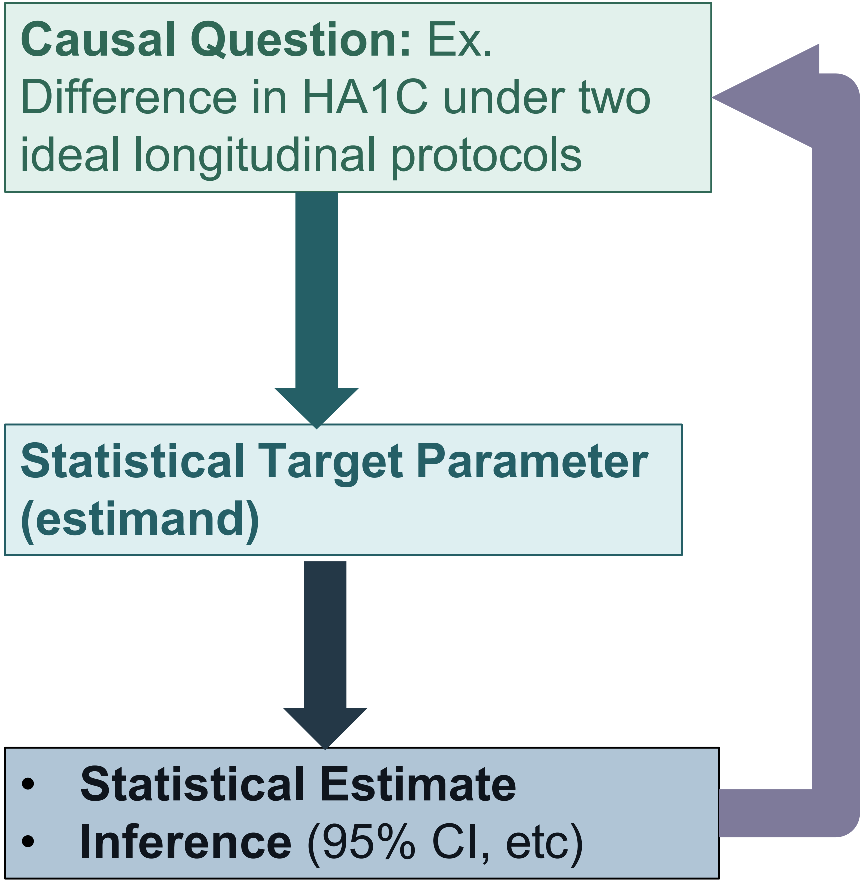
1. Better **questions-** more informative for patient care

2. Better **statistical models**
- All models are <u>not</u> wrong
- Large enough to reflect uncertainty

3. Better **estimands-** closer to the causal question

4. Better **estimators**
- Less biased, Less variable
- Accurate quantification of uncertainty (inference)

5. **Interpretation-** more transparency



Causal Question: Ex. Difference in HA1C under two ideal longitudinal protocols

Statistical Target Parameter (estimand)

- **Statistical Estimate**
- **Inference** (95% CI, etc)

Presentation outline

1.Background

2.Causal Roadmap:

- I. Target causal parameter
- II. Observed Data
- III. Causal Model and Identification.
- IV. Estimation
- V. Interpretation

3.Conclusion

Roadmap step I: Target causal parameter

Scientific question: *What is the effect of cumulative exposure to GLP-1RAs vs. active comparator (other second line drugs; SLGT2i or DPP4) on dementia?*

- Translating this question into a formal causal parameter requires carefully defining an ideal hypothetical experiment:

1. **Target population**
2. **Treatment regimes of interest** (ideal protocols)
   - *What variables would you ideally intervene to control and how?*
     - Can include measurement frequency, follow-up
   - *What variables do you not intervene on?*
     - Ex. Competing risks, adherence (if part of the effect of interest)
3. **Outcome**
4. **Target Causal parameter:** Population-level summary measure used to contrast counterfactual outcomes under different treatment regimes

Roadmap step I: Target causal parameter

Scientific question: *What is the effect of cumulative exposure to GLP-1RAs vs. active comparator (other second line drugs; SLGT2i or DPP4) on dementia?*

- Translating this question into a formal causal parameter requires carefully defining an ideal hypothetical experiment:

1. **Target population**
2. **Treatment regimes of interest** (ideal protocols)
   - *What variables would you ideally intervene to control and how?*
     - Can include measurement frequency, follow-up
   - *What variables do you not intervene on?*
     - Ex. Competing risks, adherence (if part of the effect of interest)
3. **Outcome**
4. **Target Causal parameter:** Population-level summary measure used to contrast counterfactual outcomes under different treatment regimes

Roadmap step 1: Target causal parameter

Scientific question: *What is the effect of cumulative exposure to GLP-1RAs vs. active comparator (other second line drugs; SLGT2i or DPP4) on dementia?*

- **Target Population**
  - Who would be in the (hypothetical, but maybe not feasible) **ideal experiment**?
  - **Inclusion criteria**
    - Previously Diagnosed Type 2 diabetes mellitus
    - Age >50
    - At least 5 years of metformin (first-line treatment)
    - Not previously diagnosed with dementia
    - Initiating second-line treatment
    - Exclude if basic bolus insulin or a pump algorithm
- **Baseline time point**
  - Date at which initiate second line treatment on one of the drugs of interest (coming up)
  - Conceptually similar to a randomized control trial with rolling enrollment

Roadmap step I: Target causal parameter

Scientific question: *What is the effect of cumulative exposure to GLP-1RAs vs. active comparator (other second line drugs; SLGT2i or DPP4) on dementia?*

- Translating this question into a formal causal parameter requires carefully defining an ideal hypothetical experiment:

1. **Target population**
2. **Treatment regimes of interest** (ideal protocols)
   - *What variables would you ideally intervene to control and how?*
     - Can include measurement frequency, follow-up
   - *What variables do you not intervene on?*
     - Ex. Competing risks, adherence (if part of the effect of interest)
3. **Outcome**
4. **Target Causal parameter:** Population-level summary measure used to contrast counterfactual outcomes under different treatment regimes

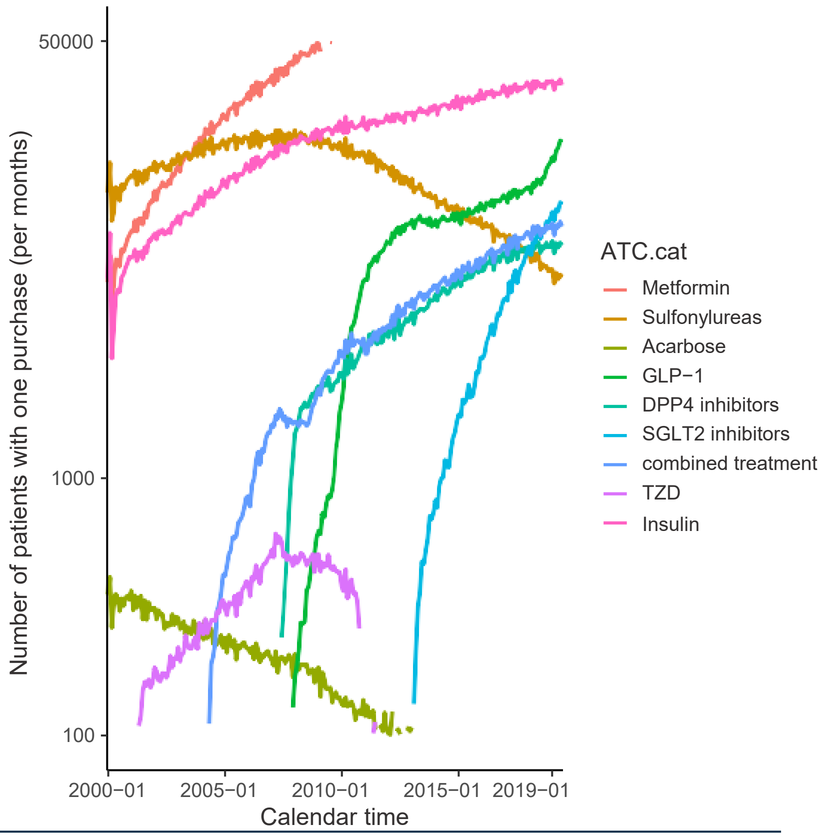
Treatment regimes of interest

In an ideal experiment, what variables would we want to intervene on and how to contrast dementia risk?

Option 1: Ex. A hypothetical intervention on drug use at a single time point (baseline):

- "Treatment": Initiate GLP-1 at baseline
- "Active comparator": Initiate other 2nd-line treatments at baseline
- In this ideal experiment, participants could still interrupt or switch treatments post-baseline
- **Not ideal for getting at cumulative effects**

Changing usage of T2DM second-line treatments in the Danish registry



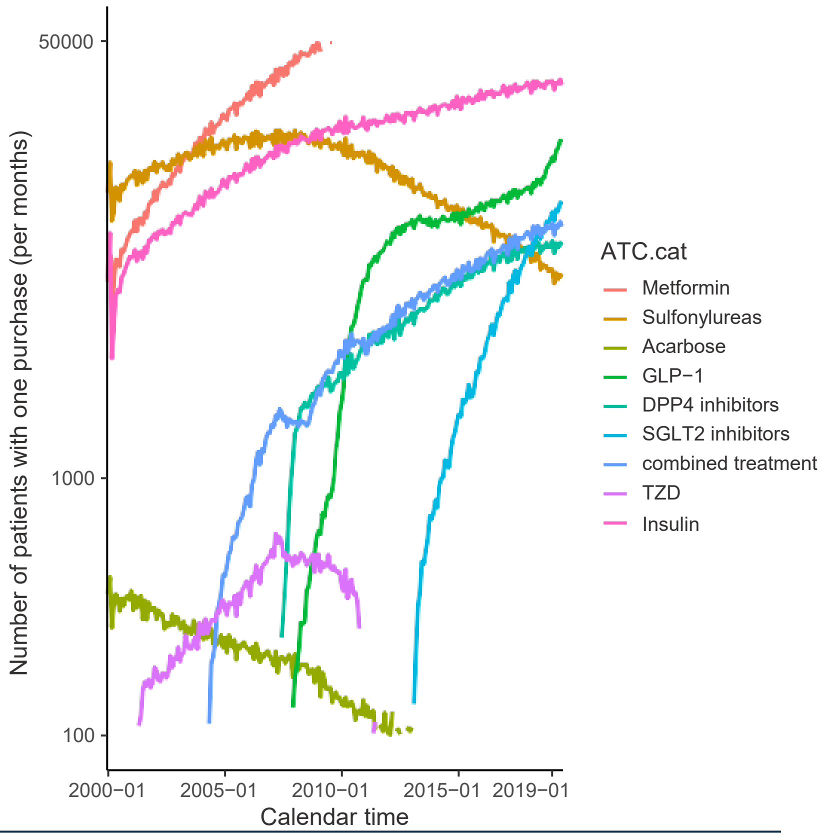
Treatment regimes of interest

In an ideal experiment, what variables would we want to intervene on and how to contrast dementia risk?

Option 2: Intervene on second-line treatment use at **multiple time points**

- "Treatment": Initiate GLP-1 at baseline **and** stay on it at least 5 years
- "Active comparator":
  - Comparator #1: Initiate SGLT2 or DPP4 and stay on at least 5 years
  - Comparator #2: "Standard of care": Treat per standard care with everything but GLP-1
    - Full range of second line drugs

Changing usage of T2DM second-line treatments in the Danish registry



Treatment regimes of interest

In an ideal experiment, what variables would we want to intervene on and how to contrast dementia risk?

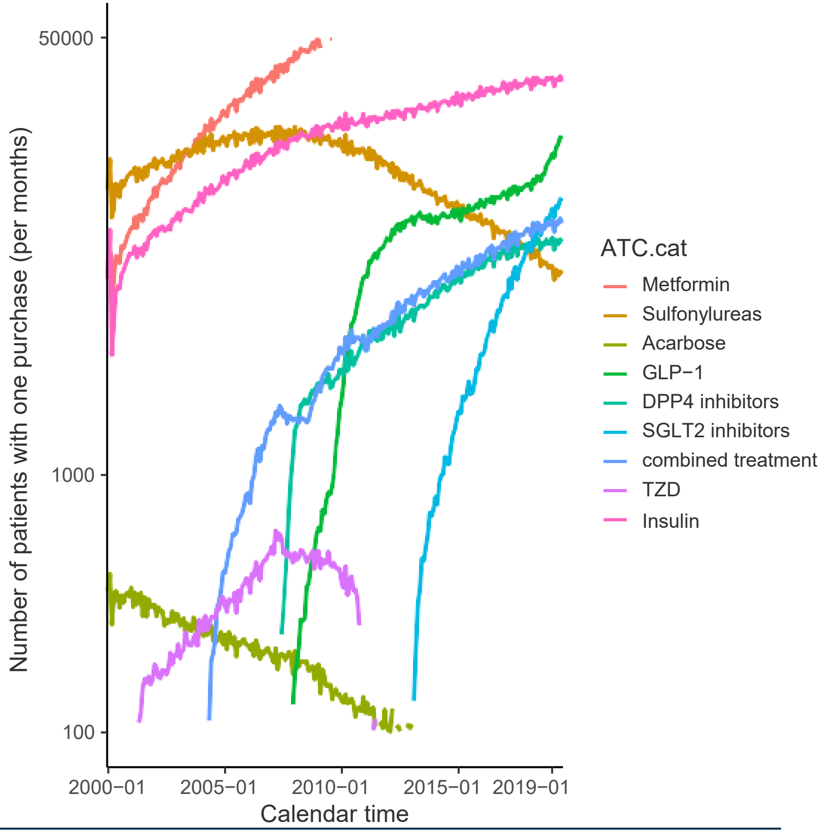
Censoring:

- In ideal experiment: intervene on follow-up (prevent right censoring):
- Ensures that we follow all participants until the max time point we care about (eg 5 years)

Ideal treatment regimes of interest:

- **$\bar{a}$=1**: Initiate GLP1 and remain on it for five years of follow -up
- **$\bar{a}$=0**: Initiate "active comparator": and remain on it for five years of follow -up

Changing usage of T2DM second-line treatments in the Danish registry



Roadmap step I: Target causal parameter

Scientific question: *What is the effect of cumulative exposure to GLP-1RAs vs. active comparator (other second line drugs; SLGT2i or DPP4) on dementia?*

- Translating this question into a formal causal parameter requires carefully defining an ideal hypothetical experiment:

1. **Target population**
2. **Treatment regimes of interest** (ideal protocols)
   - *What variables would you ideally intervene to control and how?*
     - Can include measurement frequency, follow-up
   - *What variables do you not intervene on?*
     - Ex. Competing risks, adherence (if part of the effect of interest)
3. **Outcome**
4. **Target Causal parameter:** Population-level summary measure used to contrast counterfactual outcomes under different treatment regimes

Roadmap step 1: Target causal parameter

Scientific question: *What is the effect of cumulative exposure to GLP-1RAs vs. active comparator (other second line drugs; SLGT2i or DPP4) on dementia?*

Outcome (endpoint)**:**

- Diagnosis of dementia
  - Let Y denote an indicator of Dementia diagnosis by 5 years
    - Y(t): indicator diagnosis by time t, t=0,..., 5 years; let Y denote Y(5 years)
- Death as a competing risk:
  - Don’t “hypothetically intervene” to prevent death
  - GLP1 reduces death- that means we will to some extent underestimate the biological effect of GLP1 on dementia as it keeps people alive and they have more chance to get dementia.

Alternative option: Composite outcome (death *or* dementia diagnosis)

Roadmap step I: Target causal parameter

Scientific question: *What is the effect of cumulative exposure to GLP-1RAs vs. active comparator (other second line drugs; SLGT2i or DPP4) on dementia?*

- **Counterfactual Outcomes**: $Y_{\bar{a}}$: Indicator whether dementia would have been diagnosed by 5 years after baseline under **ideal hypothetical intervention ā**
  - $Y_{\bar{a}=1}$: Counterfactual dementia status after 5 years if every eligible subject had received “treatment” regime: continuous GLP-1 diabetes treatment
  - $Y_{\bar{a}=0}$: Counterfactual dementia status after 5 years if every eligible subject had received “active comparator” regime: other continuous second-line diabetes treatment

Roadmap step I: Target causal parameter

Scientific question: *What is the effect of cumulative exposure to GLP-1RAs vs. active comparator (other second line drugs; SLGT2i or DPP4) on dementia?*

- Translating this question into a formal causal parameter requires carefully defining an ideal hypothetical experiment:

1. **Target population**
2. **Treatment regimes of interest** (ideal protocols)
   - *What variables would you ideally intervene to control and how?*
     - Can include measurement frequency, follow-up
   - *What variables do you not intervene on?*
     - Ex. Competing risks, adherence (if part of the effect of interest)
3. **Outcome**
4. **Target Causal parameter:** Population-level summary measure used to contrast counterfactual outcomes under different treatment regimes

Roadmap step I: Target causal parameter

Scientific question: *What is the effect of cumulative exposure to GLP-1RAs vs. active comparator (other second line drugs; SLGT2i or DPP4) on dementia?*

- **Target Causal Parameter:** Function of the *unobserved* counterfactual outcome distributions.
  - Choice of population-level summary measure for contrasting the "treatment" and "active comparator" interventions
    - Ex. $\boldsymbol{EY}_{\bar{\mathbf{a}}=\mathbf{1}} - \boldsymbol{EY}_{\bar{\mathbf{a}}=\mathbf{0}}$: Causal risk difference of dementia diagnosis by five years if all patients had complied with the GLP-1 regime *vs.* active control regimes
    - Ex. Full counterfactual survival curves
      - ○ Under "Treatment" and "Active comparator" ideal protocols
      - ○ For both dementia diagnosis and death (competing risk)

Presentation outline

1. Background
2. Causal Roadmap:
   - I. Target causal parameter
   - II. Observed Data
   - III. Causal Model and Identification.
   - IV. Estimation
   - V. Interpretation
3. Conclusion

Roadmap- Overview

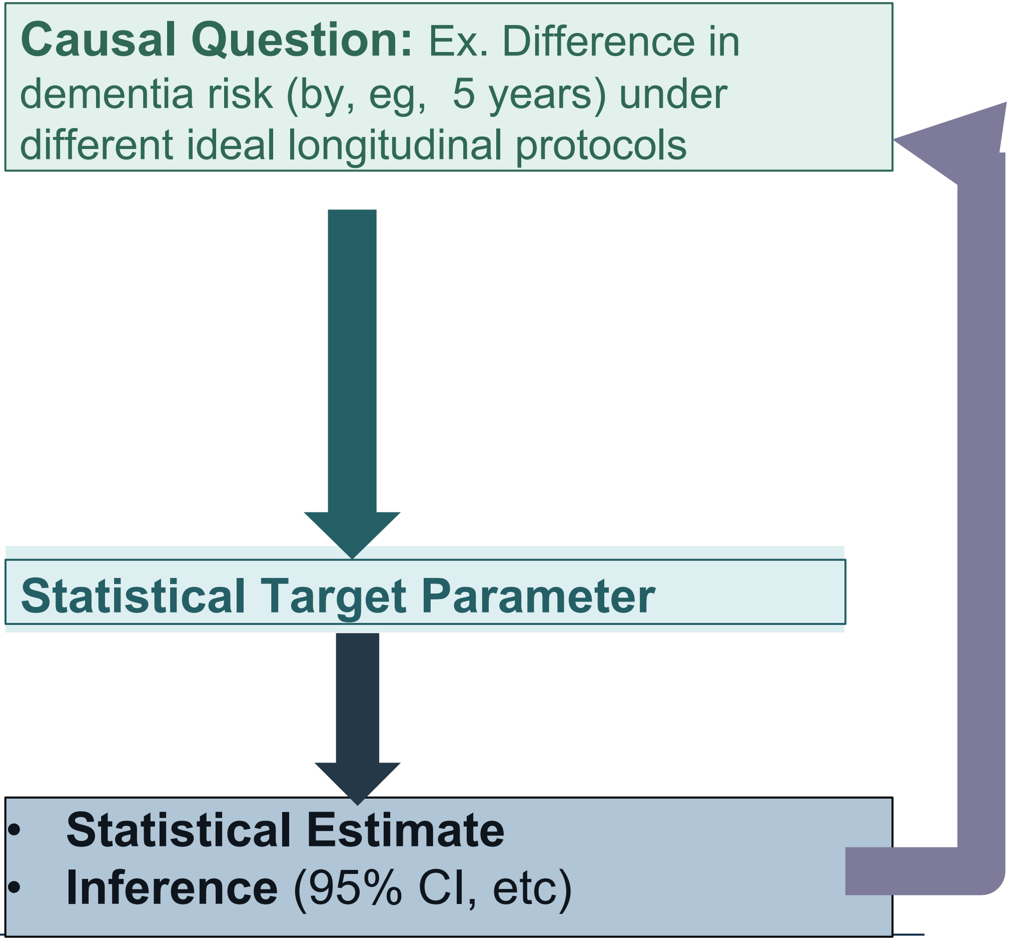
1. **Causal question**
   - Translate scientific question into causal parameter (defined in terms of counterfactual outcomes)

2. **Observed data & statistical model**
   - Model should reflect uncertainty

3. **Identify**
   - Translate causal parameter to statistical parameter under explicit causal assumptions

4. **Estimate**

5. **Interpret**



Causal Question: Ex. Difference in dementia risk (by, eg, 5 years) under different ideal longitudinal protocols

Statistical Target Parameter

- **Statistical Estimate**
- **Inference** (95% CI, etc)

II. Observed Data & Statistical Model

General Longitudinal Data Structure for Complex Observational Studies

We observe *n* i.i.d. copies of a longitudinal data structure

$$O = (L(0), A(0), . . . , L(K\ ), A(K\ ), Y = L(K)),$$

- ***K:*** max follow up time (eg, 5 years)
- ***A(t)*: intervention variables** The variables we would intervene on in our ideal experiment
  - A1(t): antihyperglycemic treatment use
  - A2(t): right censoring
- **L(t)**: **non-intervention variables**
  - L(0): Baseline (non-time) varying characteristics
  - L(t): Time-varying characteristics,
  - Y(t): Dementia diagnosis by time t
- ***Y*** is a final **outcome** of interest (Dementia diagnosis by 5 years)

Statistical model: Set of possible distributions for the Observed data

- Avoid any unsupported assumptions -> Work in large (semiparametric) statistical model

II. Observed Data: “Intervention” variables

General Longitudinal Data Structure for Complex Observational Studies

A_1(t) : Second line treatment used at time t

- GLP1 vs. active comparator vs. neither
- Measured via antidiabetic drug purchase at time t
- In observed data (unlike ideal protocol): treatment modifications and prolonged interruptions occur over time for some persons due to side effects, new options becoming available, etc.

A_2(t) : Indicator still under follow-up at time t

- In Danish registry: major source of right censoring is “administrative censoring”
    - Driven by variability in calendar date of baseline time point across participants

II. Observed Data: “Non-intervention” variables

General Longitudinal Data Structure for Complex Observational Studies

L(0) : Non-time varying covariates (at baseline: time of second line initiation)

- eg demographics, baseline medical history

L(t) : Time-varying covariates

- eg, Medical history: medical purchases, ICD-10 codes, labs, hospital admissions
- Includes death by time (t) (or alternatively, include death as a second Y(t) node)
- Includes diagnosis of dementia by time (t) (Y(t))

II. Observed Data: “Non-intervention” variables

General Longitudinal Data Structure for Complex Observational Studies

Y(t): Indicator of dementia diagnosis by time t (included in L(t))

- Dementia is based on diagnoses from the Danish Hospital Registry
- Dementia onset may occur before diagnosis
    - If the times at which dementia status assessed are variable across persons and affected by treatment-> could result in less interesting target parameter
    - <u>Solution</u>: add hypothetical intervention to ensure comparable assessment frequency
        - This is part of defining your ideal experiment: Back to step 1 of the roadmap

Presentation outline

1. Background
2. Causal Roadmap:
   - I. Target causal parameter
   - II. Observed Data
   - III. Causal Model and Identification.
   - IV. Estimation
   - V. Interpretation
3. Conclusion

Roadmap- Overview

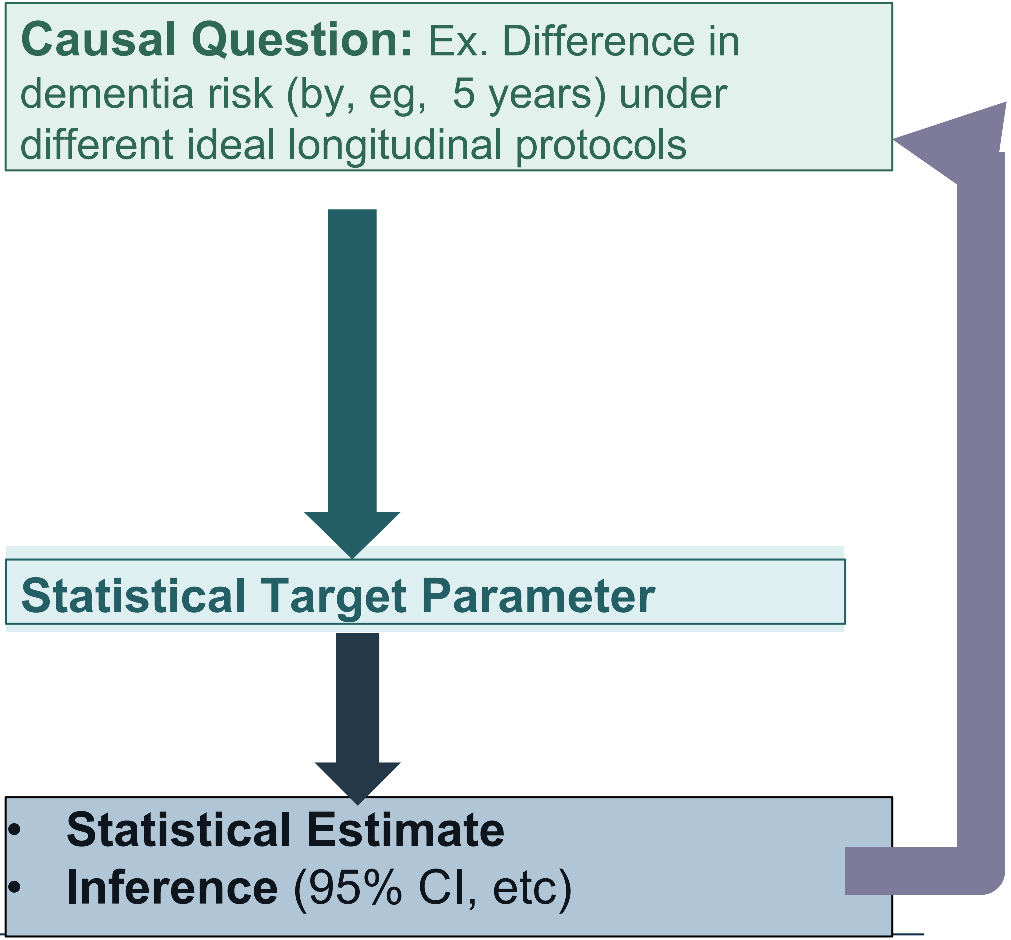
1. **Causal question**
   - Translate scientific question into causal parameter (defined in terms of counterfactual outcomes)

2. **Observed data & statistical model**
   - Model should reflect uncertainty

3. **Identify**
   - Translate causal parameter to statistical parameter under explicit causal assumptions

4. **Estimate**

5. **Interpret**



Causal Question: Ex. Difference in dementia risk (by, eg, 5 years) under different ideal longitudinal protocols

Statistical Target Parameter

- **Statistical Estimate**
- **Inference** (95% CI, etc)

IIIa. Causal Model and Identification.

- Causal identification: translate target causal parameter into a parameter of the observed data distribution (a statistical target parameter) so we can estimate it.
- The Causal Model is a tool to do this by expressing our knowledge about the data generating process -> facilitates this translation process.

- Key subject matter expertise:
  - Why are patients started on GLP1 vs other second line treatments?
  - What are key reasons patients switch off GLP1 to something else over time?
  - What are reasons for, or potential predictors of longer term (>9 month) interruption?
  - What are key causes of (or predictors of dementia)?

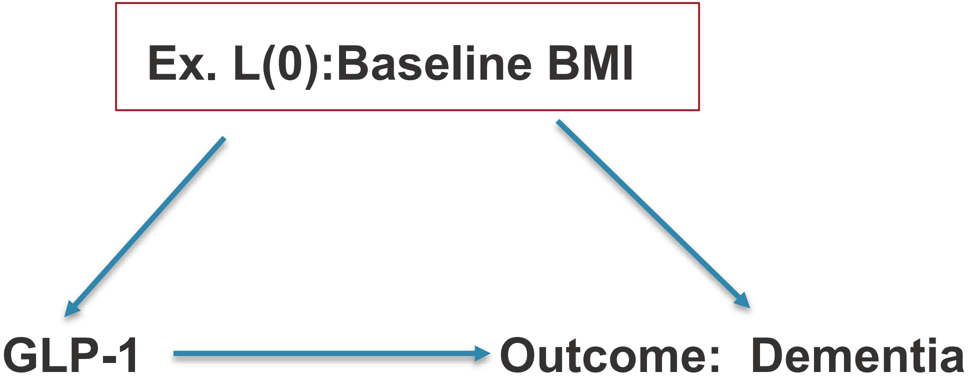
Identification: Baseline Confounding

Identification assumption for single time point intervention:
“Randomization assumption” (or conditional exchangeability)

$Y_{\bar{a}} \perp A | L(0)$

Baseline covariates (L(0) sufficient to control for confounding

Holds if measured baseline covariates block all “backdoor paths” $A \rightarrow Y$



Ex. L(0):Baseline BMI

GLP-1

Outcome: Dementia

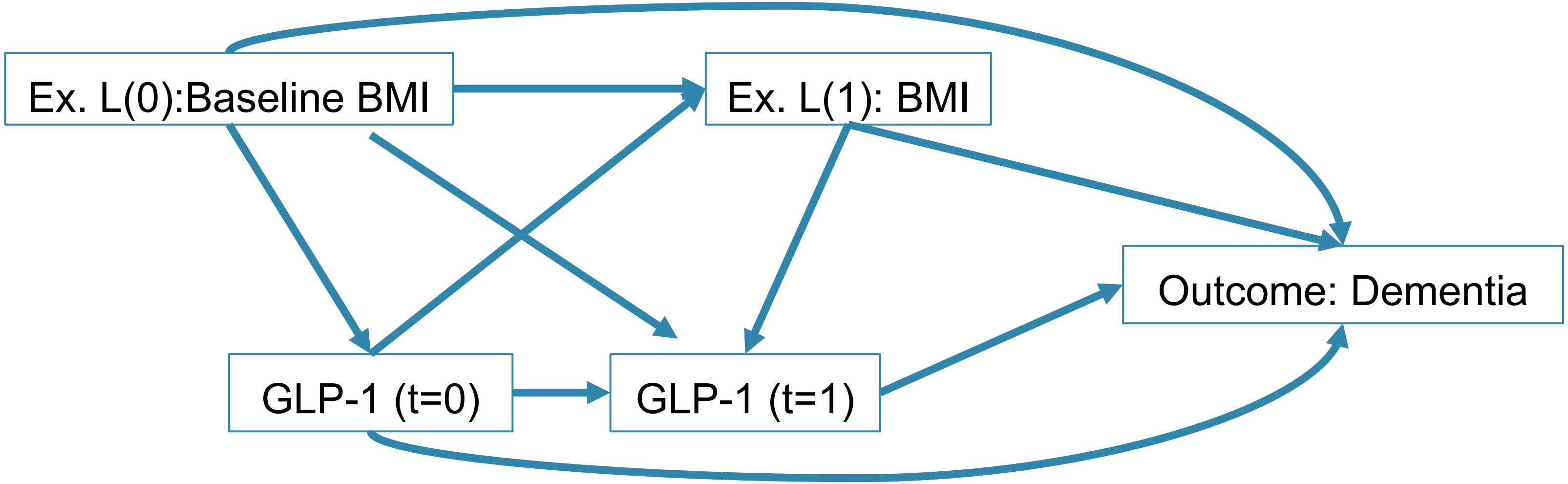
Confounders

Potential Baseline confounders: Variables measured at time of second line therapy initiation

- **Demographics**
  - Age
  - Sex
  - Income
  - Region of Denmark
  - Education
  - Household size (living alone)
  - Marital status
  - Employment status
- **Medical history at baseline**:
  - Time since diagnosis of diabetes;
  - Pre-baseline treatments,
  - Baseline labs (HA1C)
  - Comorbidities,
  - BMI….

More complex longitudinal structure

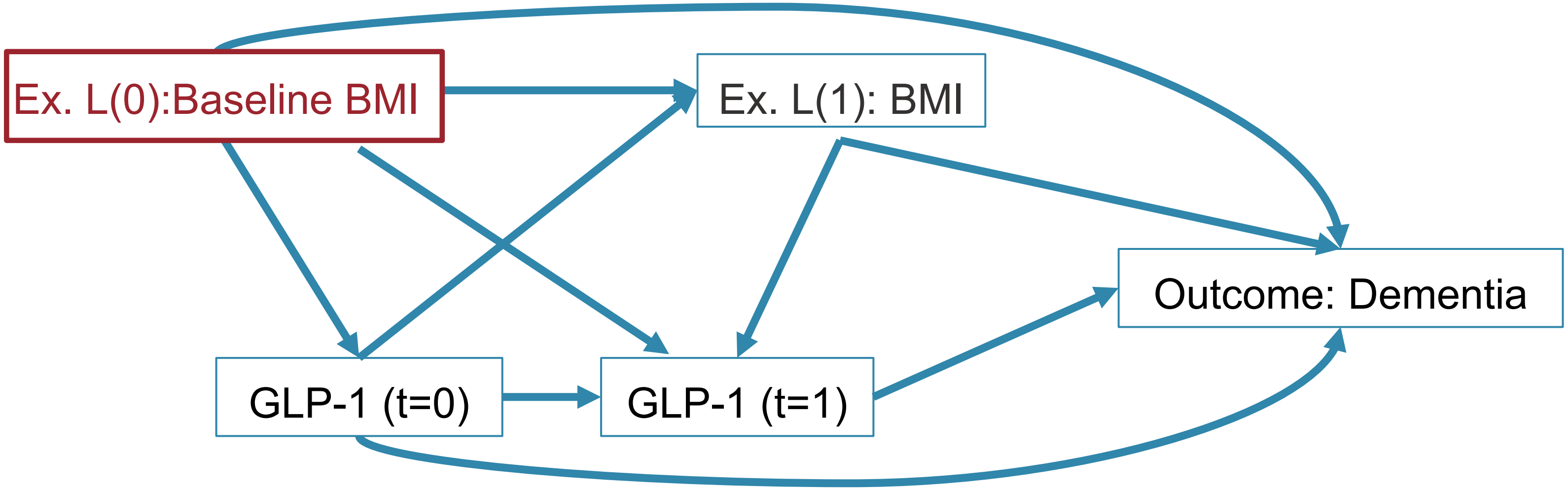
The challenge of post-baseline confounding



More complex longitudinal structure

Classical outcome regressions: no way forward!

- Fit regression (or Cox model) of outcome on GLP-1 use, adjusting for baseline BMI?



Ex. L(0):Baseline BMI

Ex. L(1): BMI

Outcome: Dementia

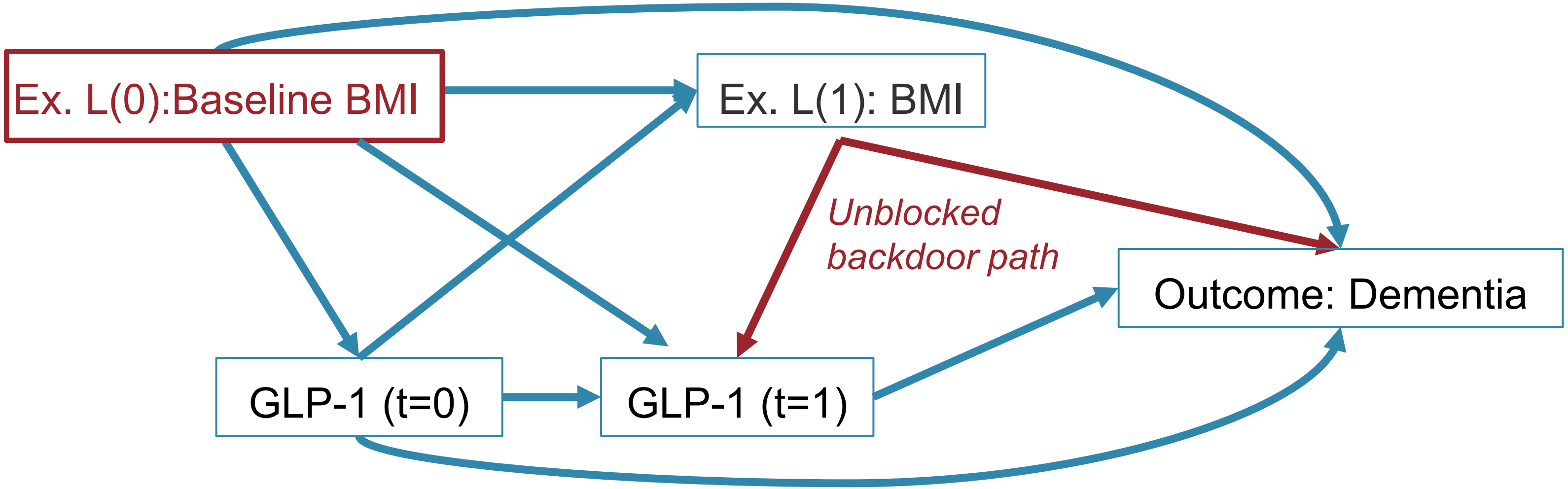
GLP-1 (t=0)

GLP-1 (t=1)

More complex longitudinal structure

Classical outcome regressions: no way forward!

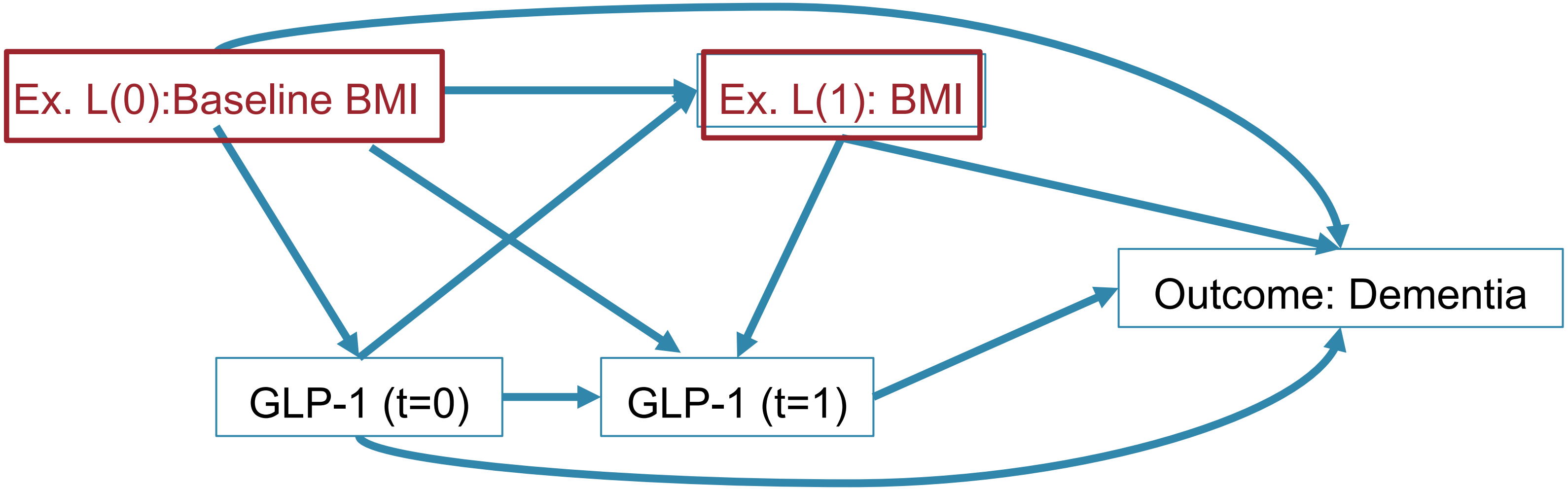
- Fit regression (or Cox model) of outcome on GLP-1 use, adjusting for baseline BMI?
  - **No!** Have not accounted for fact that those not on GLP-1 initially may experience more weight gain, affecting chance of future switch to GLP-1



More complex longitudinal structure

Classical outcome regressions: no way forward!

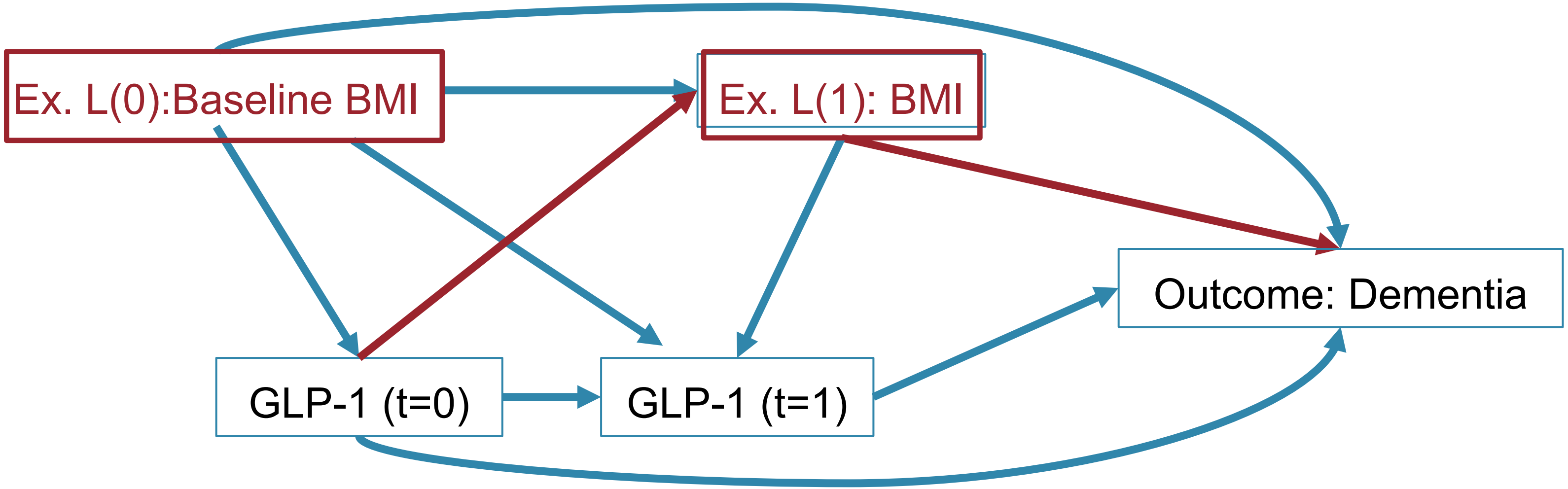
- Fit regression (or Cox model) of outcome on GLP-1 use, adjusting for baseline BMI and post-baseline BMI?



More complex longitudinal structure

Classical outcome regressions: no way forward!

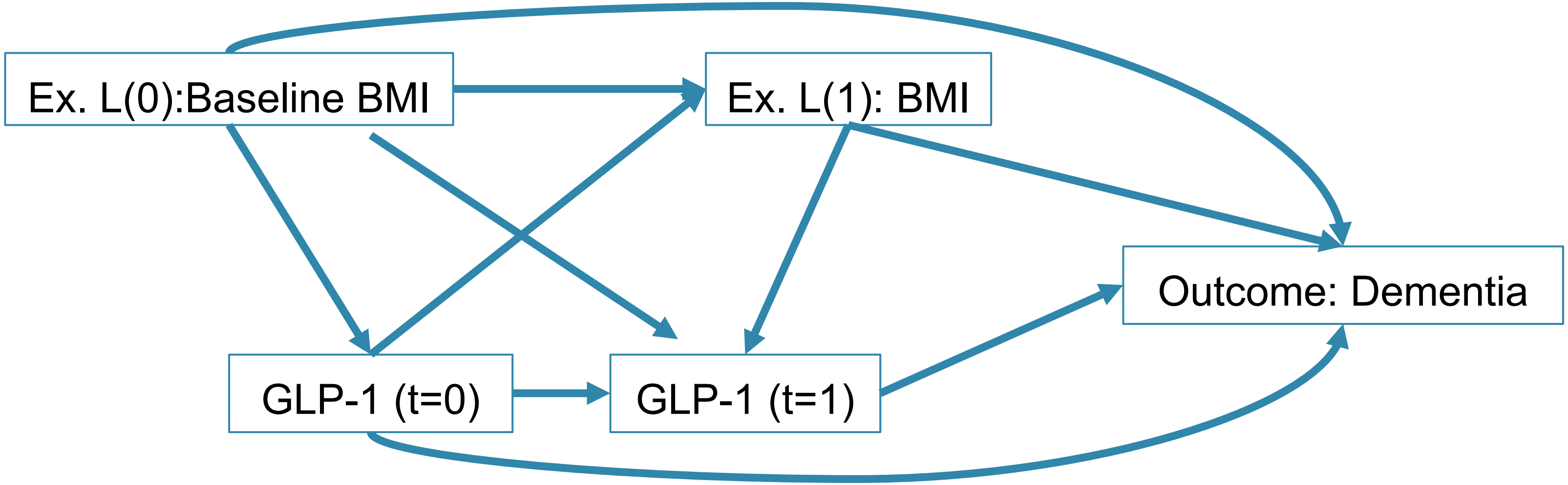
- Fit regression (or Cox model) of outcome on GLP-1 use, adjusting for baseline BMI and post-baseline BMI?
  - **No!** “Blocking” (adjusting away) part of the effect of interest



Identification: Time-dependent confounding

Sequential Randomization Assumption (SRA ,or sequential exchangeability) :

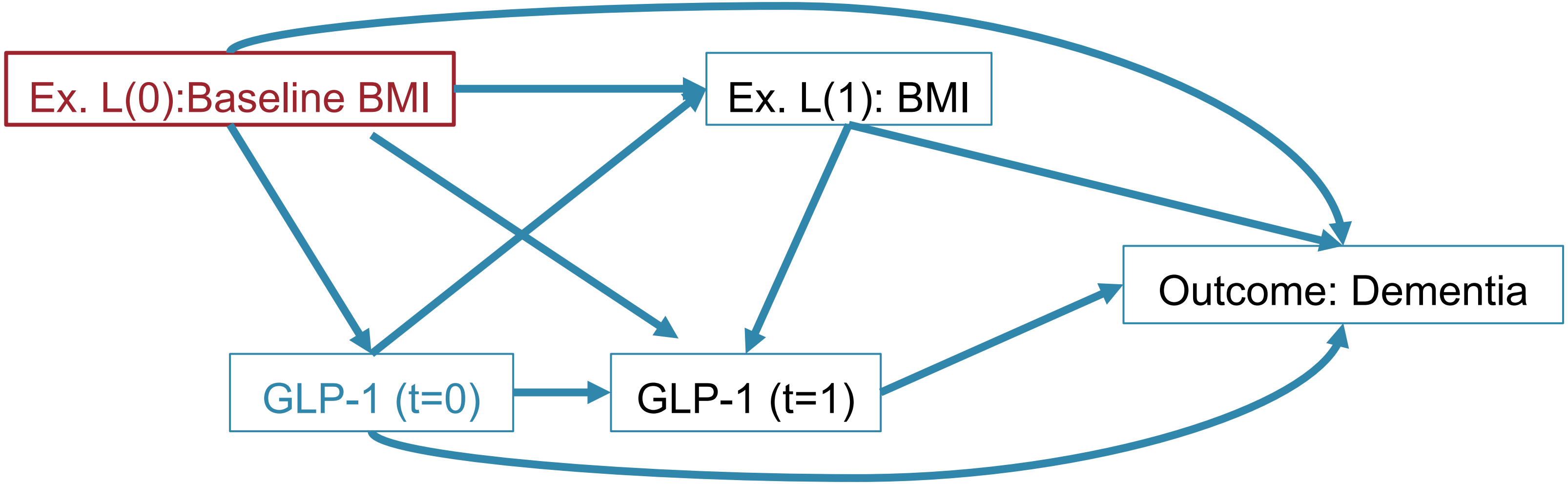
- $Y_{\bar{a}} \perp$ A(t) | Observed Past
- Holds if: For each intervention node A(t), the measured past is sufficient to block all back door paths from A(t) to future Y(τ) (τ≥t)



Identification: Time-dependent confounding

Sequential Randomization Assumption (SRA, or sequential exchangeability) :

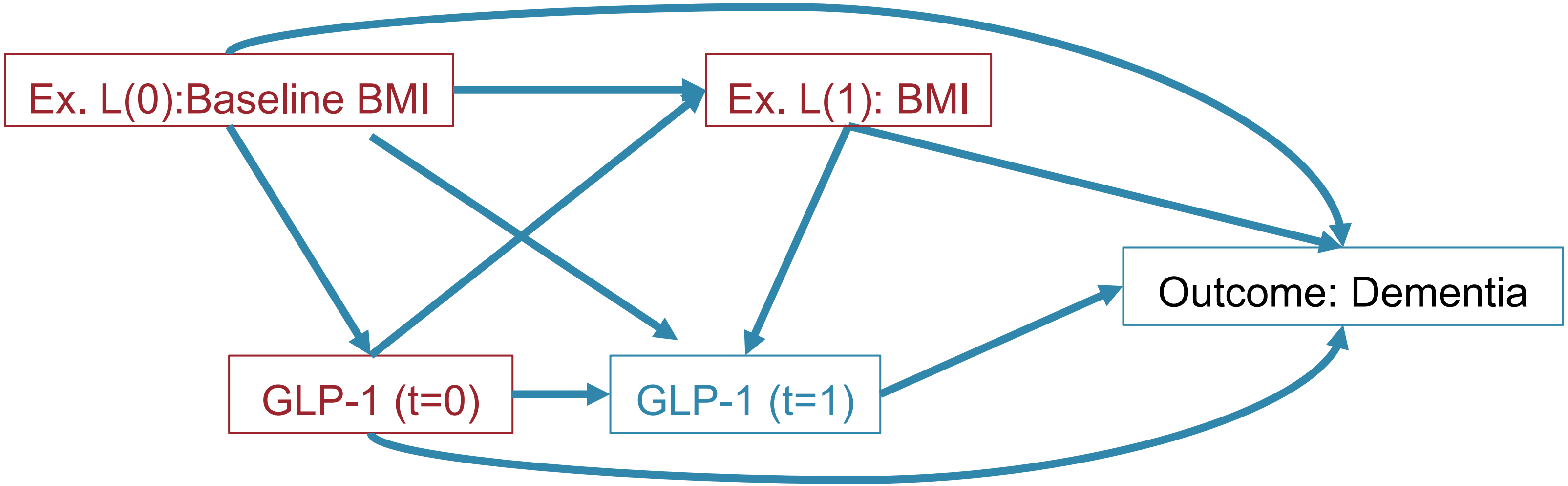
- $Y_{\bar{a}} \perp A(t)$ | Observed Past
- Holds if: For each intervention node A(t), the measured past is sufficient to block all back door paths from A(t) to future Y(τ) (τ≥t)



Identification: Time-dependent confounding

Sequential Randomization Assumption (SRA ,or sequential exchangeability) :

- $Y_{\bar{a}} \perp$ A(t) | Observed Past
- Holds if: For each intervention node A(t), the measured past is sufficient to block all back door paths from A(t) to future Y(τ) ($\tau \geq t$)



Confounders

Potential **time varying** confounders:

- Medical history: medical purchases, ICD-10 code, hospital admissions
  - CV history- MI, stroke, hypertension.
  - CV meds (ACE-I, other anti-HTN, statins, beta-blockers, ASA/other antiplatelet, etc.)
  - Comorbidity index
  - BMI
  - Renal disease
- Numbers of drugs prescribed
  - physicians may not want to increase drug burden on persons with early signs of dementia
- Hemoglobin A1C
- Visit frequency

IIIb. Positivity.

Key assumption for identification

- There must be some positive probability of continuing to follow each regime of interest (ie “ideal protocol”) at each time point, given that you have followed it so far, and irrespective of covariate history up to that time point.

  $P(A(t) = a(t) \mid \text{Observed Past}, \bar{A}(t-1)=\bar{a}(t-1)) > 0$ for $\bar{a} \in \{0, 1\}$

  - Example: if a patient’s glucose control gets poor enough, treatment will (essentially) always be intensified.
  - Ideal experiments that enforce no intensification would not be supported.
- This illustrates how careful definition of the ideal intervention of interest is important to ensure positivity.

IIIb. Positivity.

Key assumption for identification

- Additionally, the practical positivity assumption must be met: there must be support in the data to estimate to effect of complex combinations of covariates.
- We will need to define a class longitudinal regimes (ideal protocols) we will contrast and covariates we will adjust for that have adequate support in the data
    - We may choose these data adaptively- currently conducting an outcome blind analyses to look at those with support

Challenges:

- We need to preserve fine time scale in order to preserve causal ordering and maintain full confounder information (ie to optimize chance that SRA holds)
- However, for support (positivity), also want to limit the number of time points at which treatment can potentially change

Key identification result (longitudinal treatment)

*Longitudinal G-computation: Under the (sequential) randomization and positivity assumptions assumptions, we can express casual parameter as **a statistical target parameter***

$$E(Y_{\bar{a}}) = \sum_{\bar{l}} E(Y|\bar{A}(K) = \bar{a}(K), \bar{L}(K) = \bar{l}(K)) \times$$

$$\prod_{t=1}^{K} P(L(t) = l(t)|\bar{A}(t-1) = \bar{a}(t-1), \bar{L}(t-1) = \bar{l}(t-1))$$

- This parameter does not equal
  - a coefficient in a single parametric regression model
  - an exponentiated coefficient in a Cox PH model
  - a point treatment parameter from any estimation method
- We need estimators to solve the specific statistical problem at hand!

Presentation outline

1. Background
2. Causal Roadmap:
   - I. Target causal parameter
   - II. Observed Data
   - III. Causal Model and Identification.
   - IV. Estimation
   - V. Interpretation
3. Conclusion

Roadmap- Overview

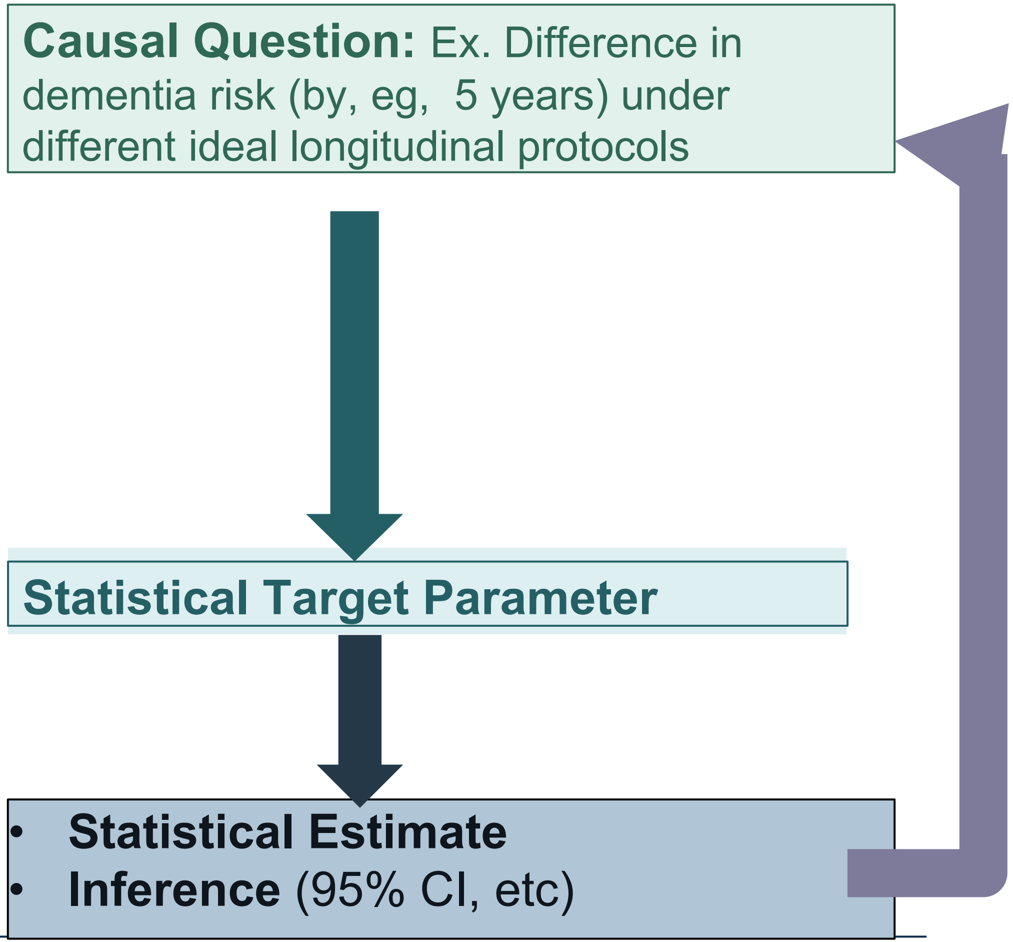
1. **Causal question**
   - Translate scientific question into causal parameter (defined in terms of counterfactual outcomes)

2. **Observed data & statistical model**
   - Model should reflect uncertainty

3. **Identify**
   - Translate causal parameter to statistical parameter under explicit causal assumptions

4. **Estimate**

5. **Interpret**



Causal Question: Ex. Difference in dementia risk (by, eg, 5 years) under different ideal longitudinal protocols

Statistical Target Parameter

- **Statistical Estimate**
- **Inference** (95% CI, etc)

Identification vs. estimation

$$\begin{tikzcd} \theta^* \arrow[r, "\text{Assumptions}"] & \theta \arrow[r, "\text{Estimator}"] & \hat{\theta} \end{tikzcd}$$

	Causal Estimand	Statistical Estimand	Statistical Estimate
	What hypothetical quantity, or parameter, are we interested in?	Can we write this causal quantity in terms of observable data?	What algorithm will best approximate this statistical quantity?
Common flawed analysis:	We are interested in the causal relative risk	We plan on estimating an odds ratio from incomplete data (unobserved confounding)	We use a logistic regression model and ignore interaction and non-linearity

Estimation

- Need estimators that provide best statistical performance (bias, variance, valid inference) for the statistical estimation problem defined using the roadmap
- **Statistical Estimation problem**
  - Observed Data: complex longitudinal data
  - Statistical Model (set of allowed distributions for observed data)
  - Statistical Target Parameter: Equal, under specific identification assumptions, to causal parameter
- **Challenges of the Estimation problem at hand**
  - High dimensional confounder set
  - Extended follow up with fine time scale
  - Complex statistical parameter (longitudinal G-computation formula)
  - Large statistical model (limited knowledge)- parametric model-based adjustment strategies -> bias

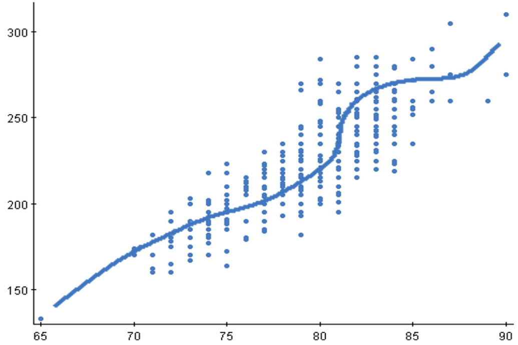
Common analysis approaches

- If the goal is **inference** (e.g., an effect size with a confidence interval), use an **interpretable, usually parametric, model** and explain what the coefficients and their standard errors mean.
- If the goal is **prediction**, use **data-adaptive machine learning algorithms** and then look at performance metrics, with the understanding that standard errors, and sometimes even coefficients, no longer exist.

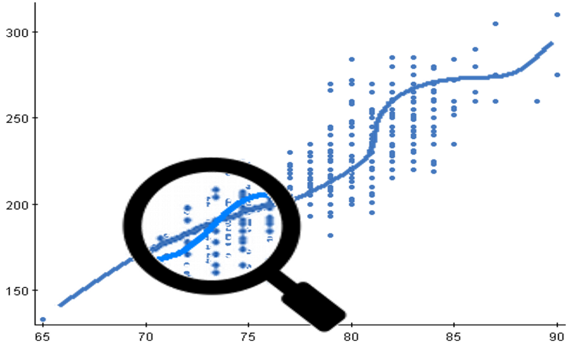
WRONG!

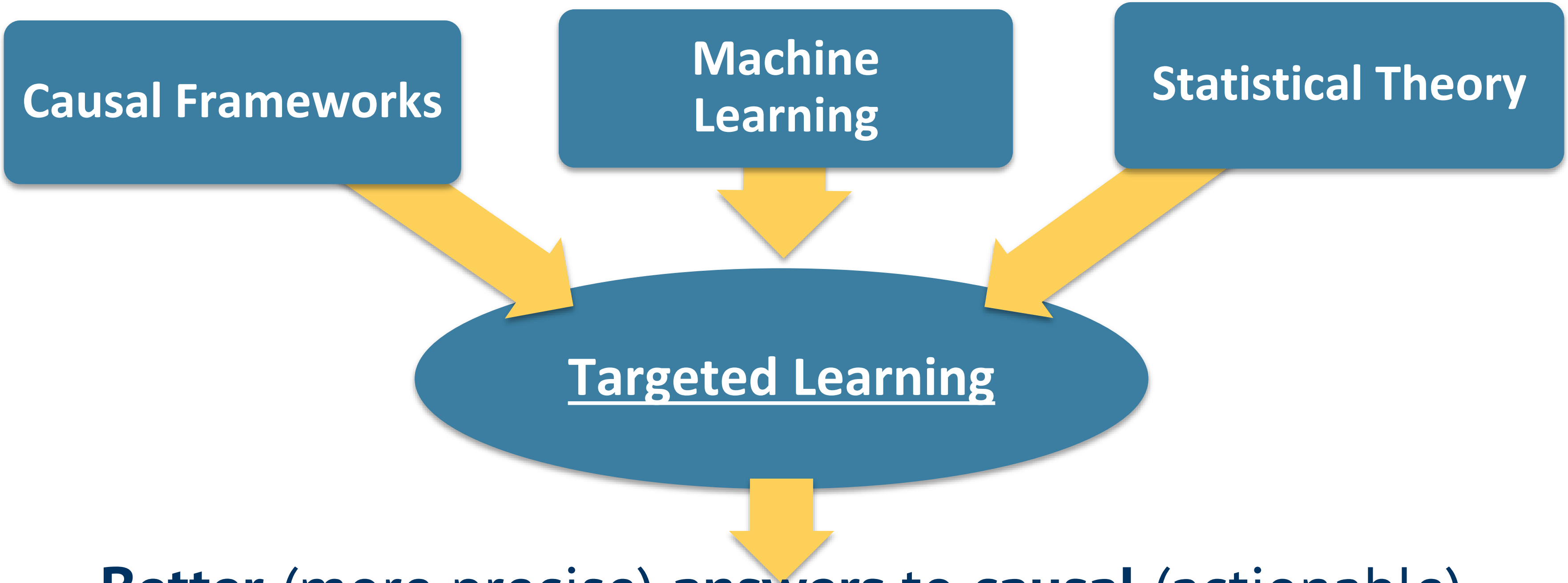
Targeted Machine Learning Estimation

1. **Super Learning-** Ensemble Machine Learning
   - Fit the data flexibly
   - Pre-specify rigorous, automated way to choose between (and combine) candidate approaches
     - Ex: Different parametric regression models
     - Ex: Machine learning methods



2. **Targeting**
   - Focus on the estimand
     - One aspect of the data—the target
   - Update Super learning fit to give best estimate for this quantitity





Causal Frameworks

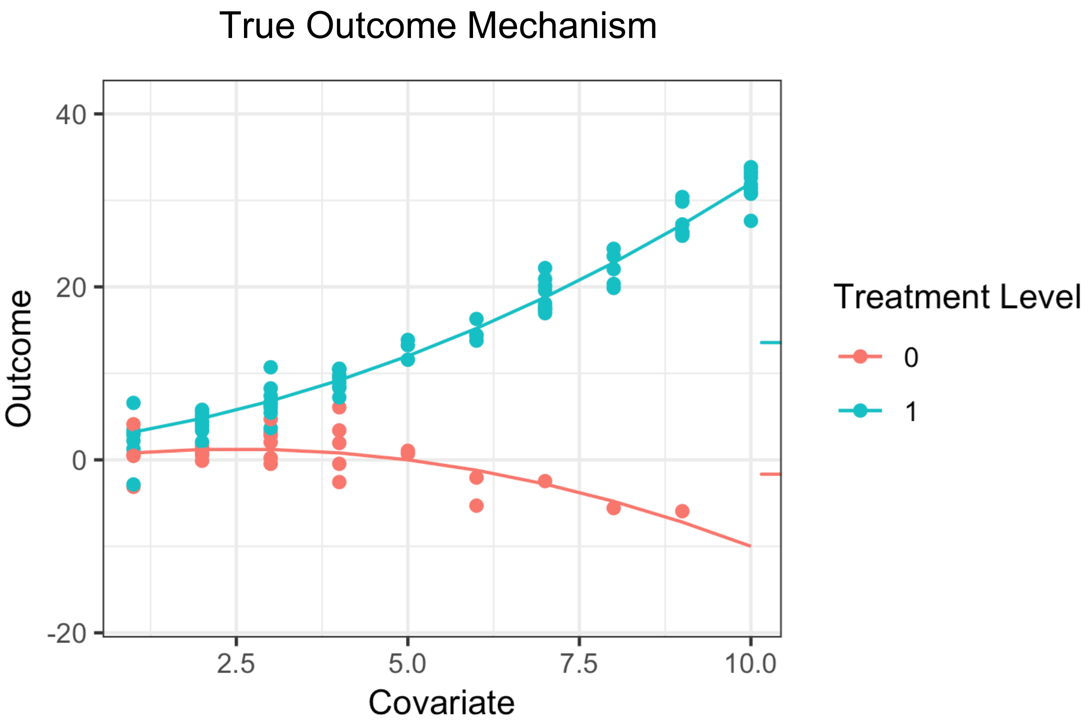
Machine Learning

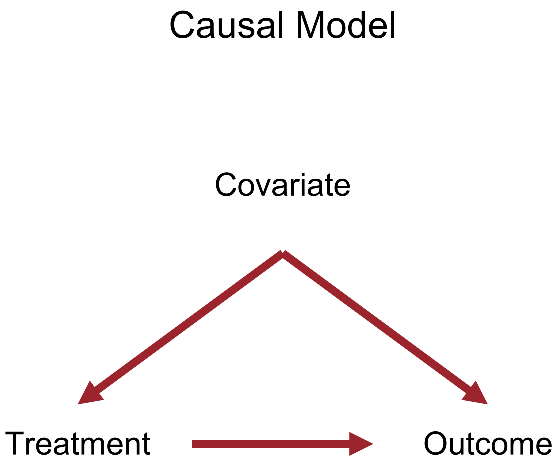
Statistical Theory

Targeted Learning

Better (more precise) **answers** to **causal** (actionable) **questions** with **accurate quantification of uncertainty** (signal from noise)

Targeted Learning Schematic





Causal Model

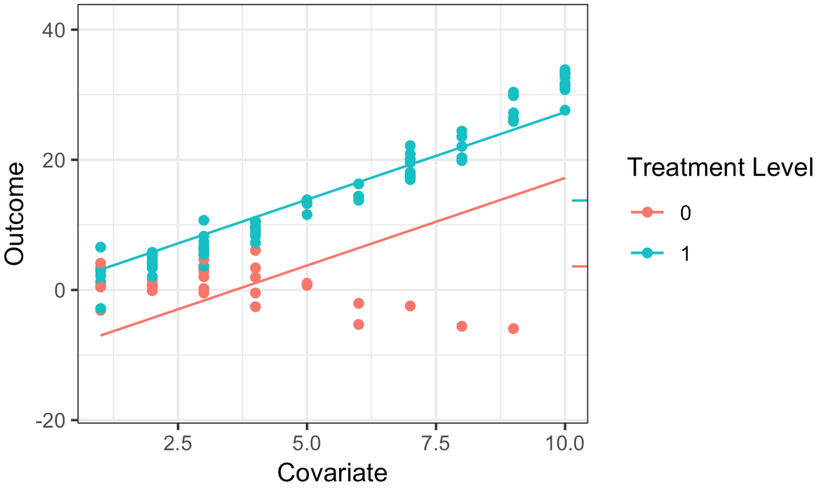
Covariate

Treatment

Outcome

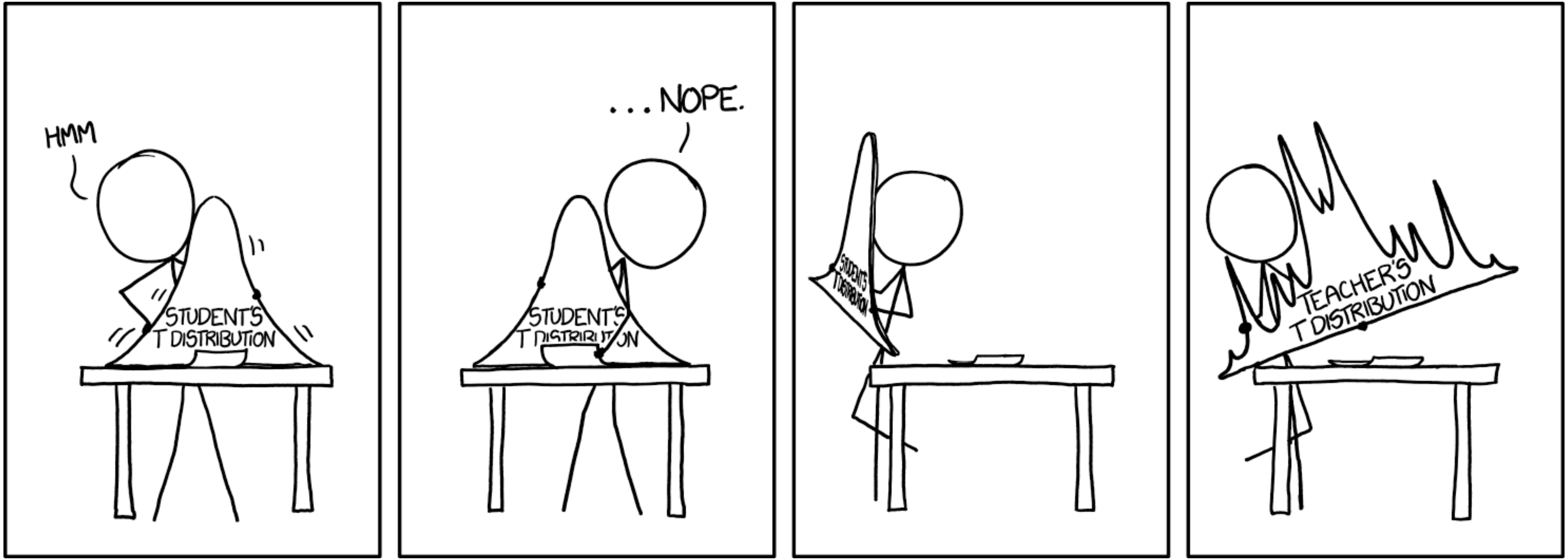
Targeted Learning Schematic

Outcome Mechanism
Estimated using Linear Regression



Underlying distributions

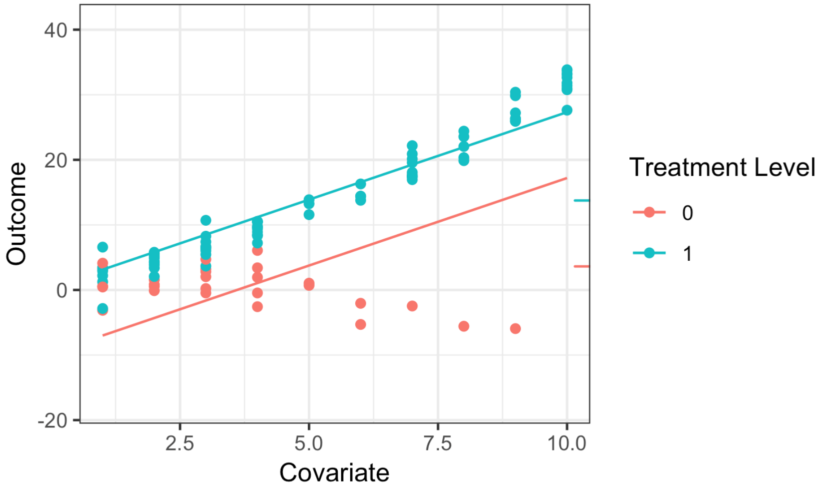
Assumptions vs. reality



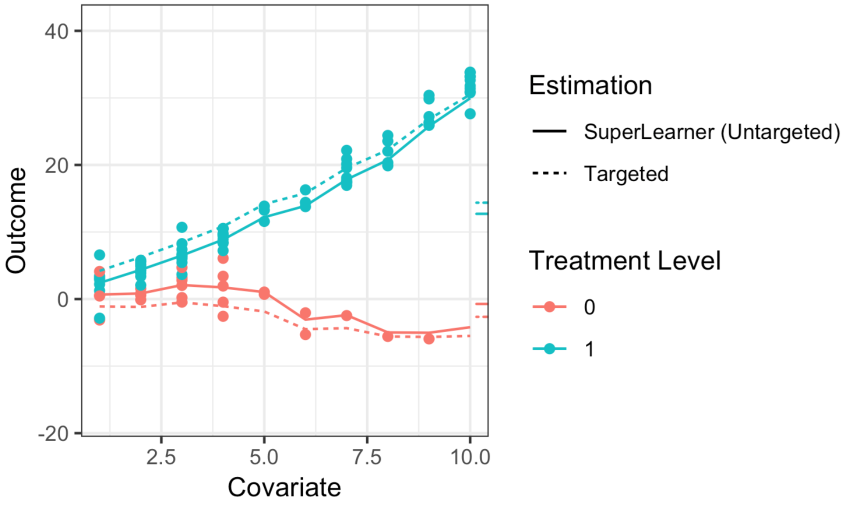
https://xkcd.com/1347/

Semiparametric estimation methods like TMLE can rely on machine learning to avoid making unrealistic parametric assumptions about the underlying distribution of the data (e.g. multivariate normality).

Targeted Learning Schematic



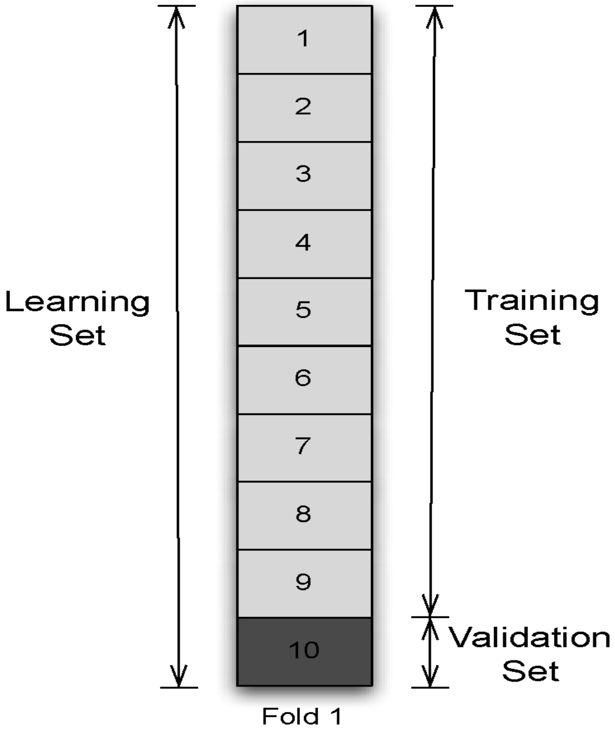
Outcome Mechanism
Estimated using Linear Regression



Outcome Mechanism
Estimated using SuperLearner and TMLE

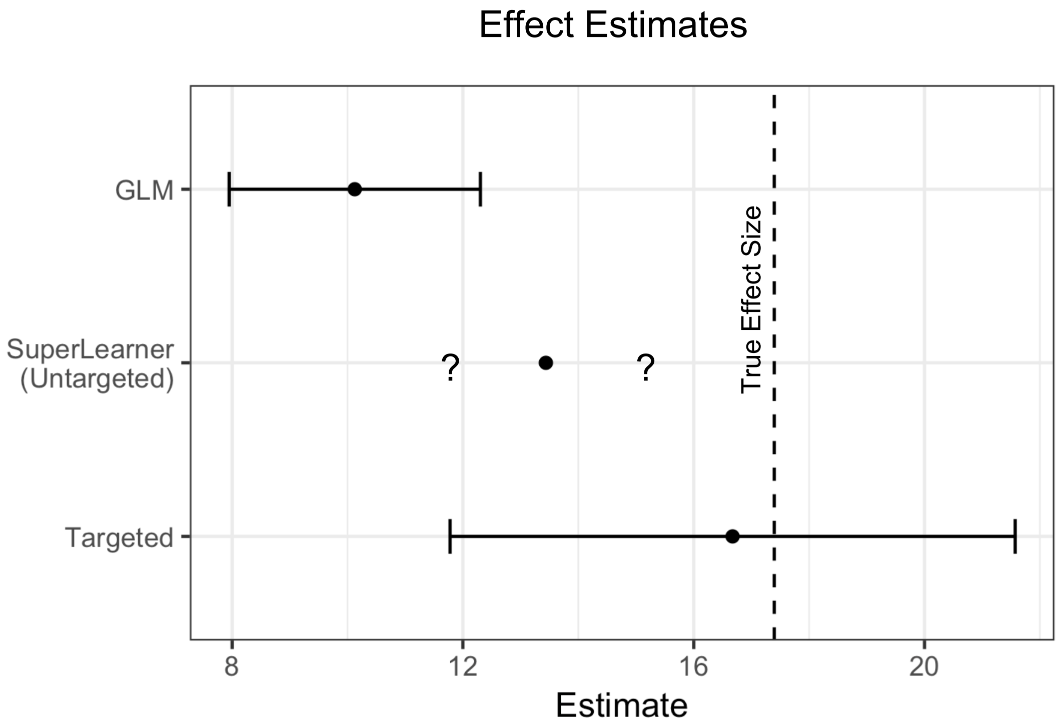
Super Learning: Ensemble Machine Learning

- “Competition” of algorithms
  - Parametric models
  - Data-adaptive (ex. Random forest, Neural nets)
- Best “team” wins
  - Convex combination of algorithms
- Performance judged on independent data
  - V-fold cross validation (Internal data splits) to avoid overfitting
  - Seek to minimize a specified loss function, for example, the mean squared error (MSE)
- Also called stacking, stacked generalizations, and weighted ensembling



Van der Laan, Polley, 2007

Results: removing bias AND robust inference



GLM did not learn the correct outcome mechanism, so its estimate is very biased

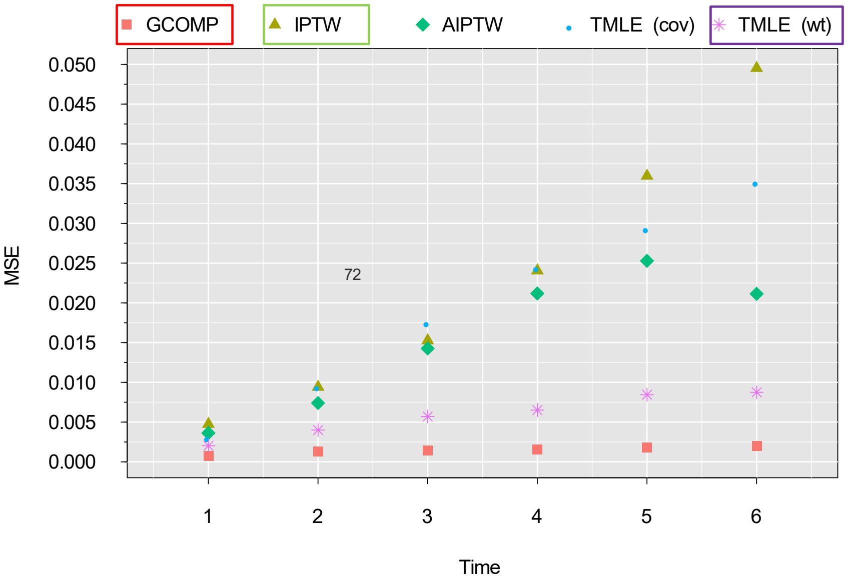
SuperLearner does a better job of estimating the outcome mechanism, but does not allow valid inference

TMLE combines good outcome mechanism estimation with targeting to get valid inference

Simulations’ role in the causal roadmap

Simulations can be used to inform statistical approach prior to finalizing statistical analysis plan

- Prior data or outcome blind data can be used to decide on statistical target parameter supported by data.
  - eg selection of ideal hypothetical interventions with adequate support in the data
- Prior data or outcome blind data can be used to set up realistic simulation
- Benchmark specifications of TMLE
  - Confidence interval coverage
  - Type I error control.
  - Provides a principled approach to navigating options in selecting a TMLE
- Can use simulation to select estimator prior to pre-registering analysis plan and peeking at real data



Presentation outline

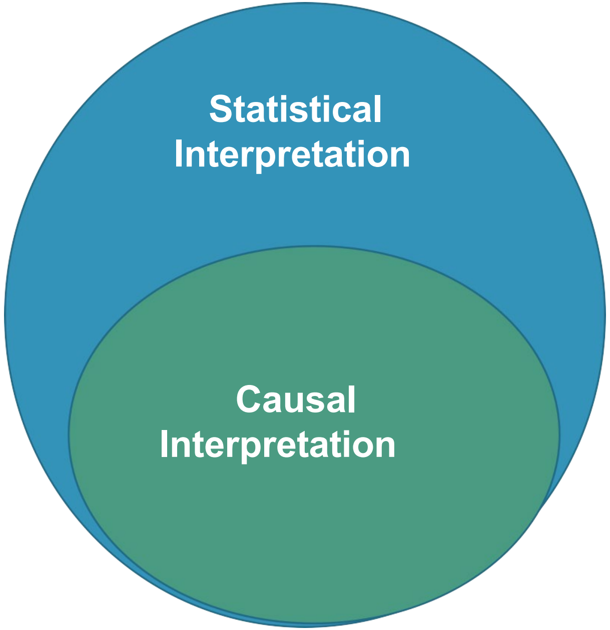
1. Background
2. Causal Roadmap:
   - I. Target causal parameter
   - II. Observed Data
   - III. Causal Model and Identification.
   - IV. Estimation
   - V. Interpretation
3. Conclusion

Step 5. Interpret

Good Practice: Be transparent, state your assumptions
Roadmap: Optimize Interpretability and Transparency/Reproducibility

A Hierarchy of interpretations

- Statistical Interpretation
  - Targeted Learning-> Reliable statistical inference!
  - For an estimand carefully chosen based on causal question
- Causal Interpretation
  - Under causal assumptions about the data generating process
  - Causal graphs (DAGs) can help make these assumptions interpretable



Presentation outline

1. Background
2. Causal Roadmap:
    - I. Target causal parameter
    - II. Observed Data
    - III. Causal Model and Identification.
    - IV. Estimation
    - V. Interpretation
3. Conclusion

Conclusion:

How does the roadmap lead to better causal, statistical, and scientific answers?

- We are clear on our causal question and its interpretation. We are forced to be completely specific about what we want to know
- We have a clear way to integrate contextual knowledge in defining the estimation problem
- The statistical parameter is designed to come as close as possible to the motivating question
- We can choose estimators with optimal statistical properties
- We verify their statistical performance and ensure valid inference through simulations

Questions?