CAUSAL INFERENCE FROM SELF-CONTROLLED CASE SERIES STUDIES USING TARGETED MAXIMUM LIKELIHOOD ESTIMATION

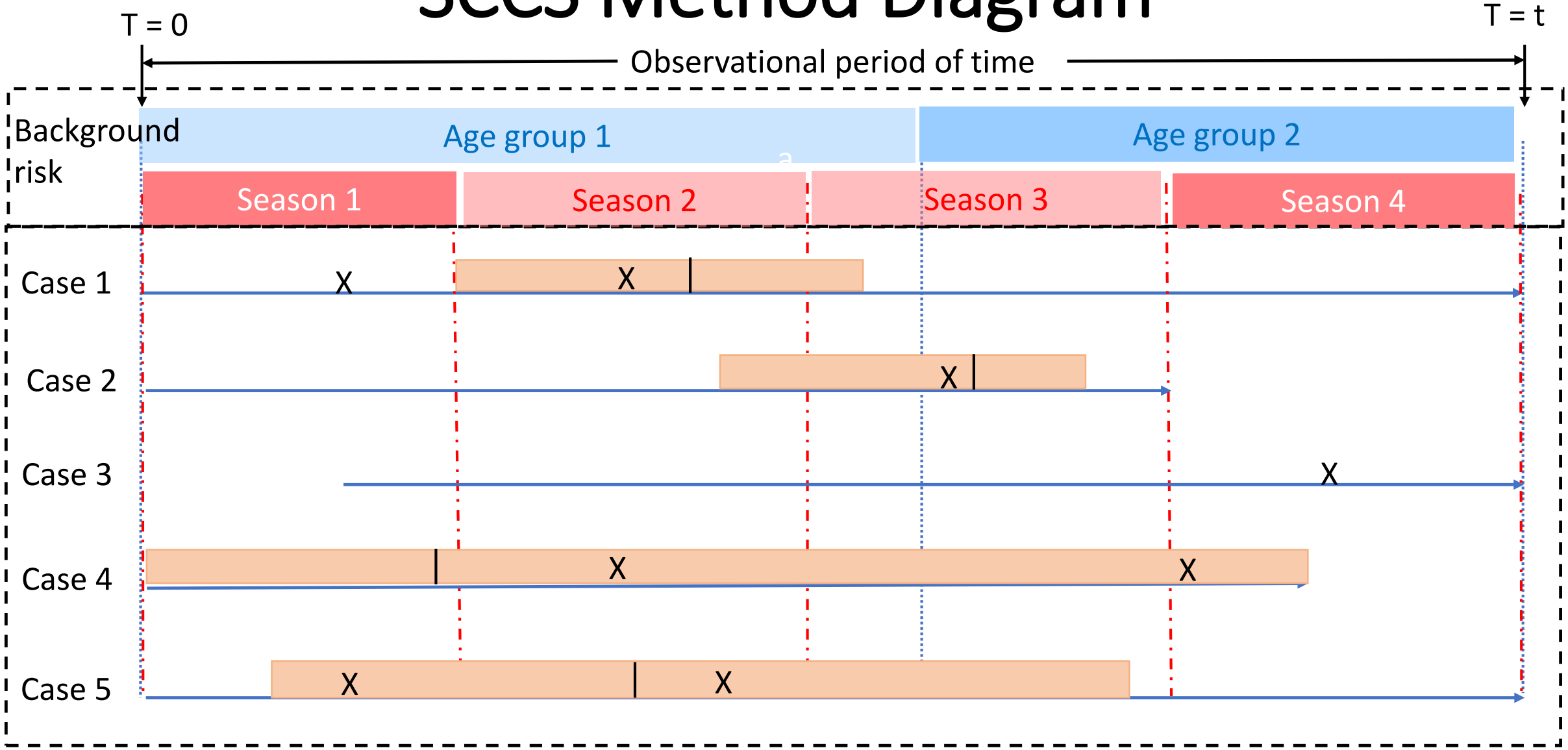
Yaru Shi, Fang Liu, Jie Chen
September 25, 2020 ASA Biopharmaceutical Section
Regulatory-Industry Statistics Workshop

MERCK
INVENTING FOR LIFE

A Motivating Example

- Investigate the casual relationship between colitis incidences and immunotherapy
- Colitis is one of the immune related AEs observed in immunotherapy medications (1.7-3% incidence rates)
- Using Melanoma Flatiron dataset (a Medicare database available on RWDEX Platform).
- Provide a natural Self-Controlled Case Series (SCCS) data set-up

SCCS Method Diagram



Notations

- For case i, the observational time $[0, T_i)$ can be partitioned into m_i subintervals $0 = T_1 < \cdots < T_{m_i-1} < T_{m_i} = T_i$, and in time window t $(0 < t \leq m_i)$,
- We observe $O_i(t) = \left(\boldsymbol{W}_i(t), A_i(t), D_i(t), Y_i(t)\right)$ for case i in subinterval t
- $\boldsymbol{W}_i$(t) denotes a vector of time-varying covariates that do not change in the subinterval $D_i(t)$
- $A_i(t)$ is the binary variable risk period associated with exposure and $Y_i(t)$ an outcome variable
- Number of events: $Y_i(t) \sim Poisson\left(\lambda_i(t) D_i(t)\right), D_i(t) = T_{t+1} - T_t$
- $\hat{\lambda}_i(t) = Y_i(t)/D_i(t)$
- $\boldsymbol{X}_i$ denotes a vector of time-invariant baseline covariates
- $\boldsymbol{O_i} = \left(\boldsymbol{X}_i, \left(\boldsymbol{W}_i(1), D_i(1), A_i(1), Y_i(1)\right), \ldots, \left(\boldsymbol{W}_i(m_i), D_i(m_i), A_i(m_i), Y_i(m_i)\right)\right)$

SCCS Model (Whitaker et al. (2006, 2009) and Farrington et al. (2018))

- Intensity parameter $\lambda_i(t)$ can be written as a Poisson regression model,

$$\log[\lambda_i(t)] = \log\left[\frac{\mu_i(t)}{D_i(t)}\right] \propto \boldsymbol{\alpha_i X_i} + \beta \boldsymbol{W_i}(t) + \phi A_i(t)$$

- The number of events within subinterval $D_i(t)$ follow a multinomial distribution with probabilities

$$\frac{\exp(\boldsymbol{\alpha_i X_i} + \beta \boldsymbol{W_i}(t) + \phi A_i(t))D_i(t)}{\sum_{t=1}^{m_i} \exp(\boldsymbol{\alpha_i X_i} + \beta \boldsymbol{W_i}(t) + \phi A_i(t))D_i(t)}$$

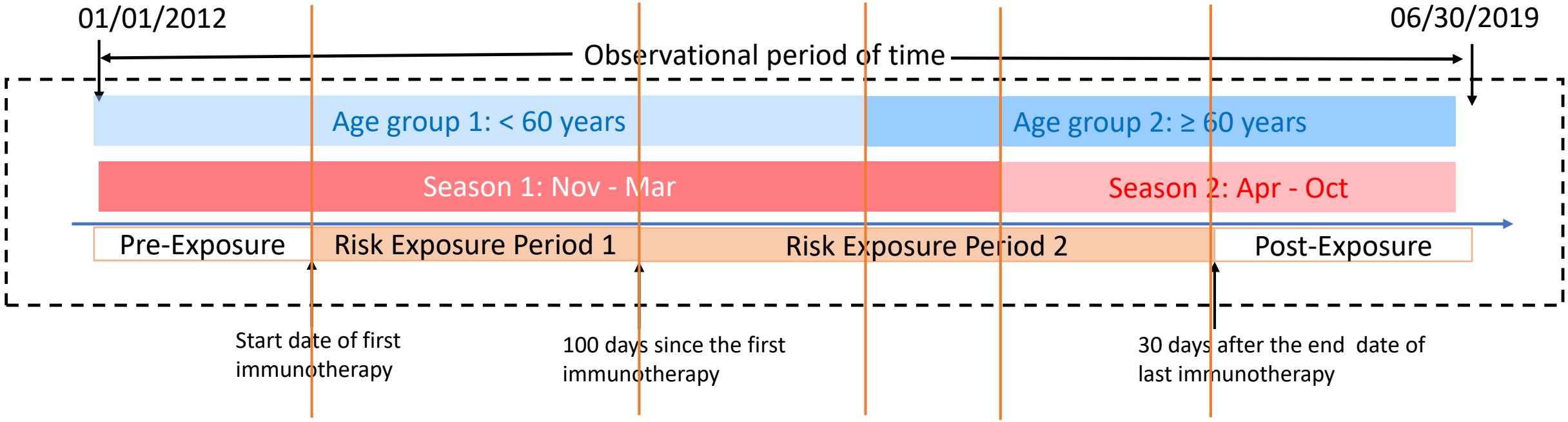
- Likelihood becomes

$$L = \prod_{i=1}^{N} \prod_{t=1}^{m_i} \left(\frac{\exp(\beta \boldsymbol{W_i}(t) + \phi A_i(t))D_i(t)}{\sum_{t=1}^{m_i} \exp(\beta \boldsymbol{W_i}(t) + \phi A_i(t))D_i(t))}\right)^{Y_i(t)}$$

- $\boldsymbol{\alpha_i X_i}$ is canceled out, no need to adjust for time-independent variables

Apply SCCS Method to Melanoma data

- Observation Period 01/01/2012 – 06/30/2019
- A total of 205 patients were diagnosed with colitis
- Among them 172 patients have received one immunotherapy medication



Preliminary Results of SCCS Model

Parameter	Estimate	Relative Risk	P value
Risk Period 1	5.01	151.6	<0.001
Risk Period 2	2.95	19.0	<0.001
Age group 2: ≥ 60 years	-0.14	0.9	0.78
Season 1: Nov - Mar	0.52	1.7	0.07
Risk Period 1 × Season 1	-0.58	0.6	0.15
Risk Period 2 × Season 1	-0.03	1.0	0.94

Overview of the Causal Inference Framework

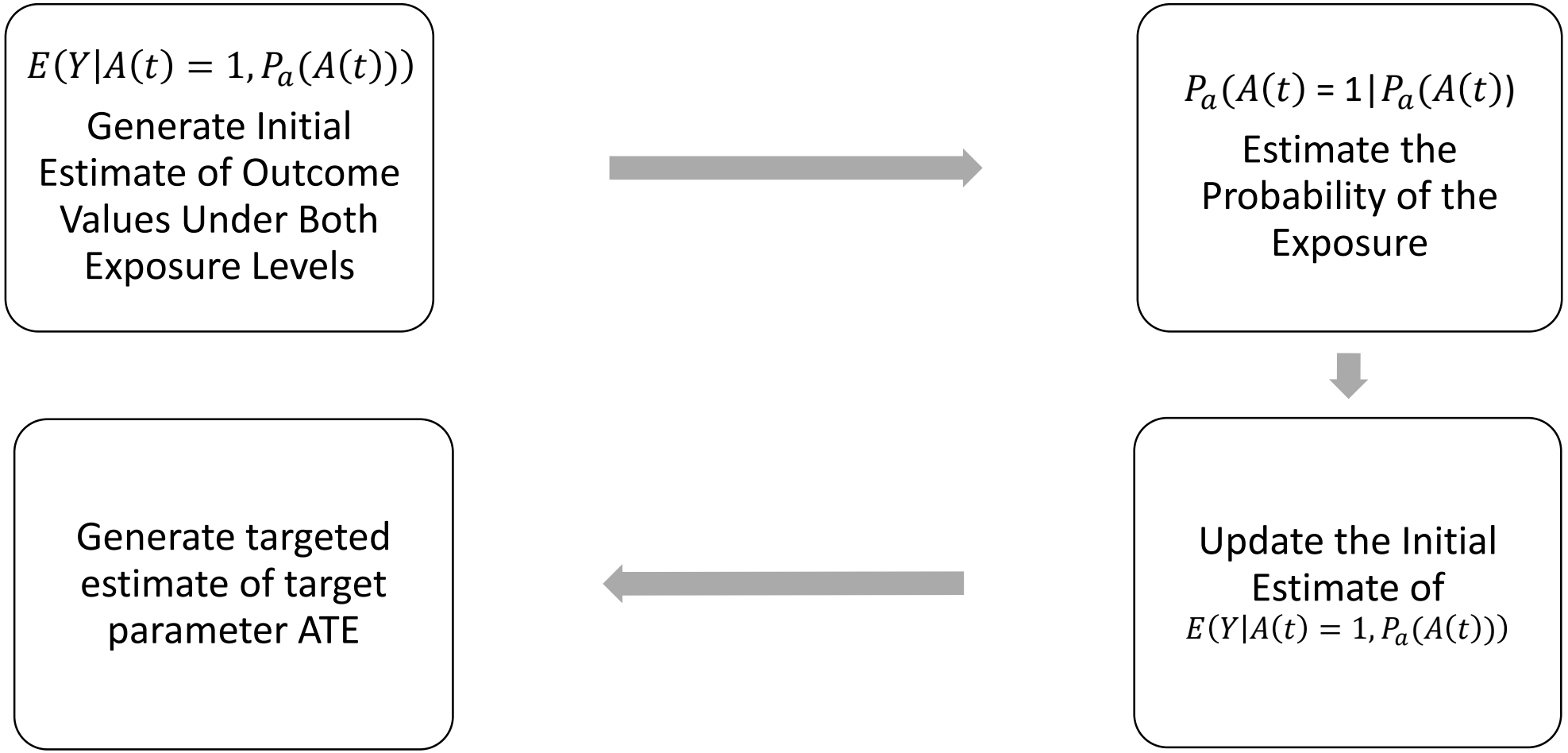
- Consider a binary exposure $A_i(t) \in \{0,1\}$
- Average treatment effect (ATE) of $A(t)$ on Y(t), adjusting for pre-treatment history is

$$\psi = E\left(E\left(Y_{A(t)=1}\middle|A(t)=1, P_a(A(t))\right) - E\left(Y_{A(t)=0}\middle|A(t)=0, P_a(A(t))\right)\right)$$

$P_a(A(t))$ are all parent variables observed prior to $A(t)$

- Assumptions in order for ψ to have a causal interpretation
  - ✓ No unmeasured confounders
  - ✓ Stable unit treatment value assumption
  - ✓ $0 < P(A(t) = 1) < 1$

Steps for Implementing TMLE (Rose S. and van der Laan M. J. (2014))



Advantage of Estimating ATE Using TMLE

- Doubly robust method and yields unbiased estimates if either $E(Y|A(t)=1, P_a(A(t)))$ or $\mathrm{P}(A(t)=1|P_a(A(t))$ is consistently estimated (e.g., correctly specified in the case of parametric regression).
- Great flexibility to incorporate a variety of algorithms, including various machine learning methods and regression methods to estimate the outcome and exposure mechanism; Machine learning algorithms can help minimize bias in comparison with use of mis-specified regressions.
- Super learning: an assembling method that implement a collection of algorithms and assign weights based on prespecified selection criteria (e.g. mean squared error)

Estimate expected outcome

- Given the

$$\boldsymbol{O_i} = \Big(\big(\boldsymbol{W}_i(1), D_i(1), A_i(1), Y_i(1)\big), \dots, \big(\boldsymbol{W}_i(m_i), D_i(m_i), A_i(m_i), Y_i(m_i)\big)|Y_i\Big)$$

- The expected outcome

$$E(Y_a(t)|A(t) = a, \boldsymbol{W}(t), D(t))$$
$$= E(yP_0(Y(t) = y|A(t) = a, \boldsymbol{W}(t), D(t))|\boldsymbol{W}(t))$$

- is identifiable under the assumptions 1) time ordering 2) consistency: $Y_a(t) = Y(t)$ for any $A(t) = a$ 3) no unmeasured confounding : $Y_a(t) \perp A(t)|\boldsymbol{W}(t), D(t)$.
- A multinomial regression could be used here to calculate the $P_0(Y(t) = y|A(t) = a, \boldsymbol{W}(t), D(t))$

Work in Progress

1. Follow the proposed framework to calculate the ATE using TMLE
2. Evaluate how the ATE differs from the treatment effect estimated by SCCS in simulations
3. Causal relationship always needs to be established with cautious!

References

- Farrington C. P. (1995). Relative incidence estimation from case series for vaccine safety evaluation. Biometrics 51: 228–35.
- Keogh R. H., Daniel R. M., Vanderweele T. J., Vansteelandt S. (2017). Analysis of longitudinal studies with repeated outcome measures: adjusting for time-dependent confounding using conventional methods. American Journal of Epidemiology 187 (5): 1085
- van der Laan M. J. (2008). Estimation based on case-control designs with known prevalence probability. The International Journal of Biostatistics 4(1): 17.
- Rose S. and van der Laan M. J. (2009). Causal inference for nested case-control studies using target maximum likelihood estimation. U. C. Berkeley Division of Biostatistics Working Paper Series: 253.
- Rose S. and van der Laan M. J. (2014). A double robust approach to causal effects in case-control studies. American Journal of Epidemiology 179(6): 663–669.
- Whitaker H. J., Hocine M. N., Farrington C. P. (2009). The methodology of self-controlled case series studies. Statistical Methods in Medical Research 18: 7-26.
- Whitaker H. J. and Ghebremichael-Weldeselassi Y. (2019). Self-controlled case series methodology. Annual Review of Statistics and Its Application 6: 241-261.

Thank you!

Background

- Self-Controlled Case Series (SCCS) method can be used to investigate associations between outcomes (e.g., AE) and exposure, using case only data (e.g., experienced the outcome of interest).
  - High efficacy relative to other methods as it is self-controlled: time-independent confounders (e.g., gender, underlying health status) are controlled implicitly.
  - Time-dependent variables (e.g., age, season) can be included in the model.
- Could be a good data to investigate the causal relationship between outcomes and exposure, further adjust the time-dependent confounders.

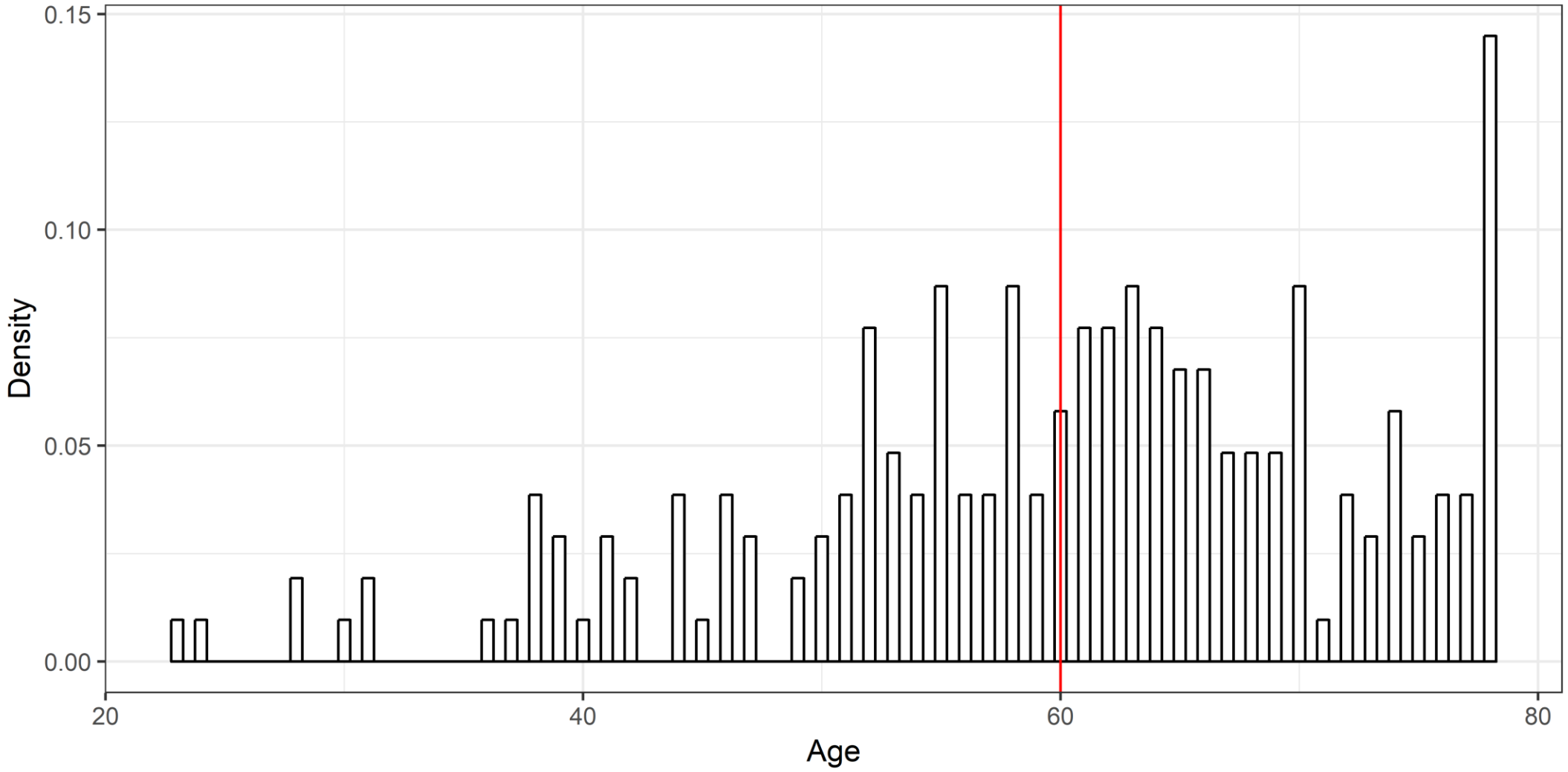
Limitation of SCCS method

1. It requires that the probability of exposure is not affected by the occurrence of an outcome event
2. It doesn’t produce estimates of absolute incidence, only estimates of relative incidence.
3. Timing of one outcome should be independent to the next outcome; the occurrence of a first event should not affect the incidence of subsequent outcomes within an individual.
   - One simple work-around is to study first outcomes only.
4. Outcome shouldn’t influence the start or end time of the observation period. When the outcome could cause death, observation periods depends on event time (death). In this case, it will create bias.
5. It does not yield a causal relationship between exposure and outcome.

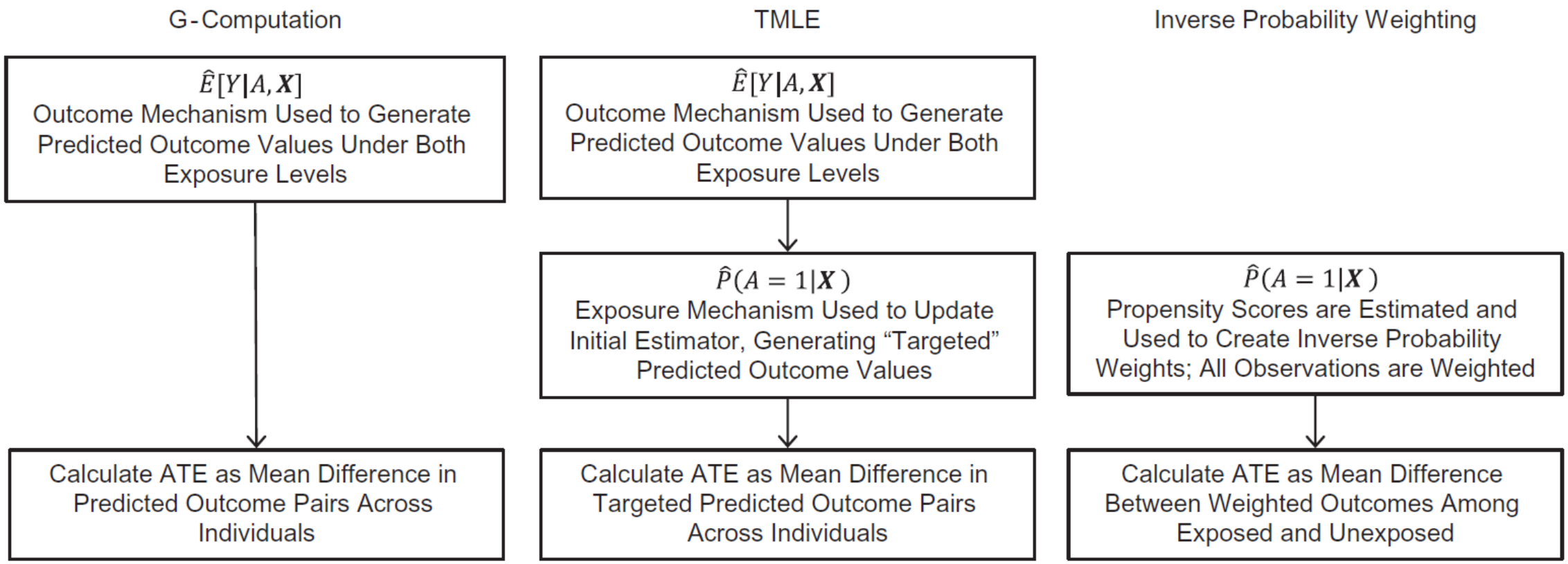
Methods to Estimate Causal Effects (Continue)

- TMLE's double robustness ensures unbiasedness of the ATE if either the exposure or the outcome mechanism is consistently estimated
- G-Computation and Inverse Probability Weighting can be adversely affected by model misspecification, particularly misspecification arising from an omitted confounder
- Inverse Probability Weighting and G-computation have typically been implemented with parametric regression
- TMLE has been incorporated with machine learning algorithms since invented

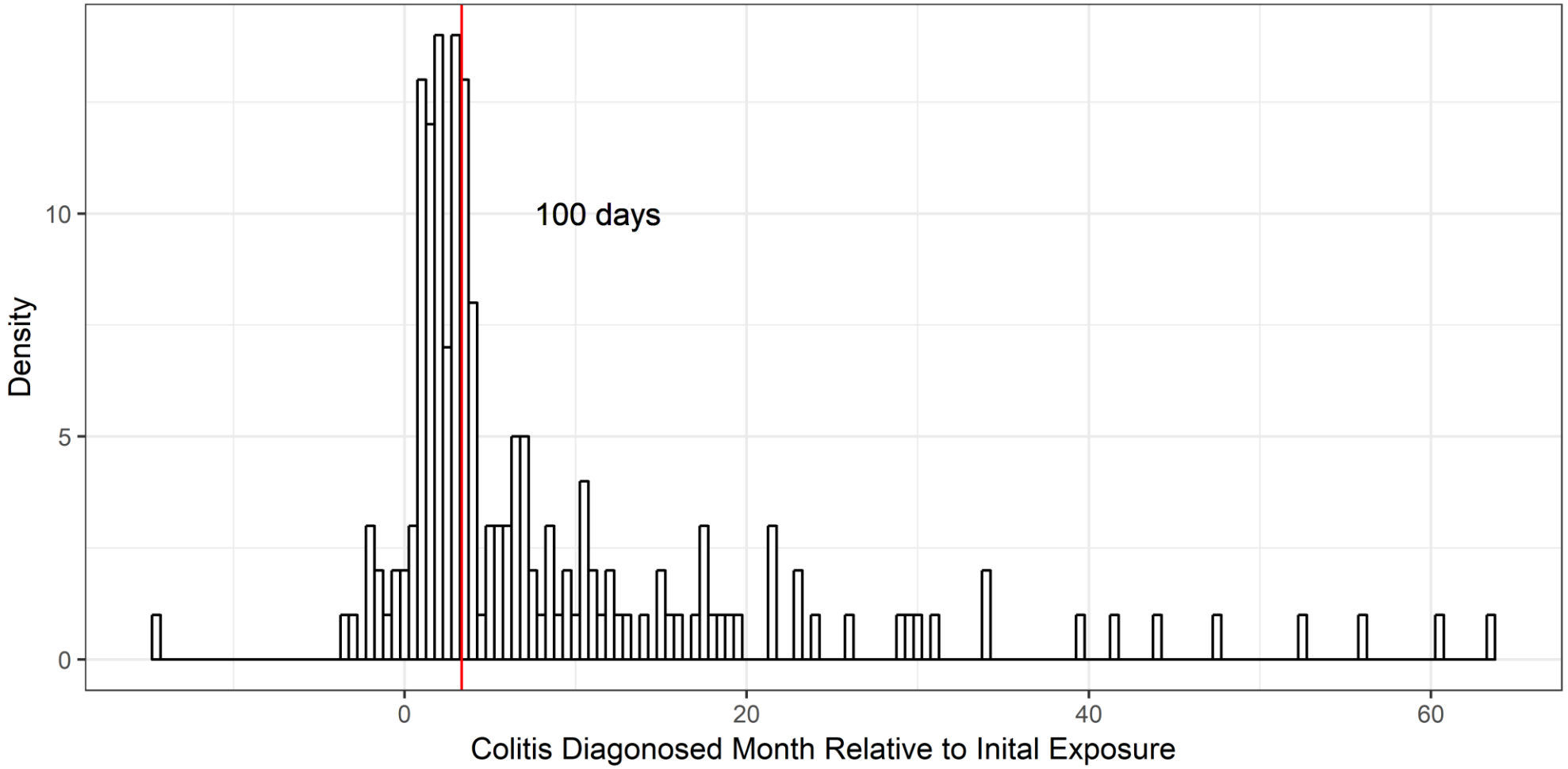
Distribution of Age at 01/01/2012



Methods to Estimate Causal Effects



Time to Onset of Colitis relative to the first dose



Season of Colitis incidence