IMPERIAL

Analysis of ordinal longitudinal data under case-control sampling: studying the association between glycocalyx degradation and mortality in critically ill patients

Chiara Di Gravio
IBC2024 Atlanta, USA
10 December 2024

The CLOVERS Study

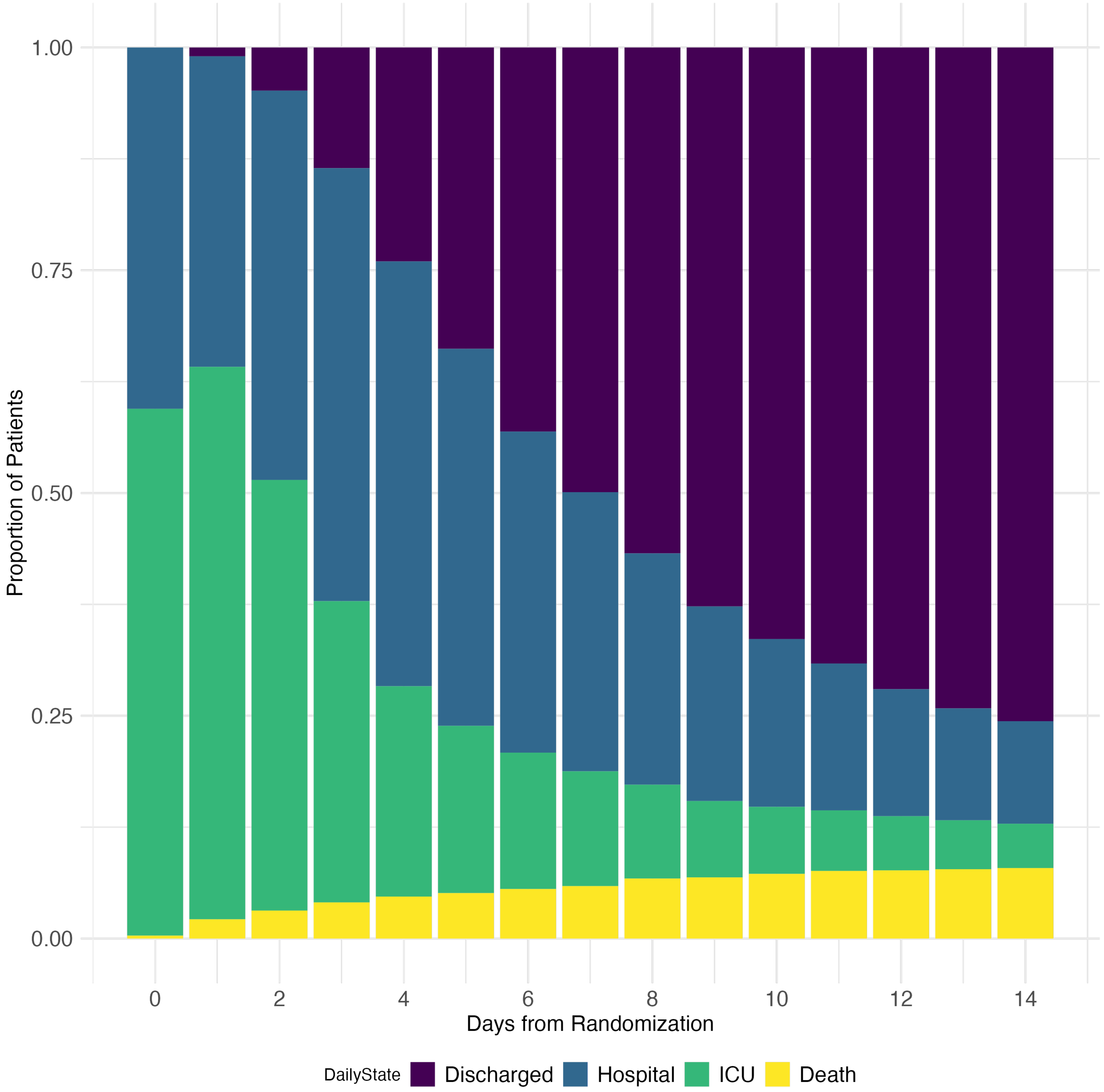
- The CLOVERS study was a randomized clinical trial comparing the effect of two resuscitation strategies on patients mortality and ARDS
- The trial recruited 1,563 critically-ill hospitalized patients with sepsis before being stopped due to inefficacy
- At recruitment, blood samples were collected and stored for later use. We want to use the collected blood samples to:
  - Measure levels of glycocalyx degradation
  - Study the relationship between glycocalyx degradation and mortality

IMPERIAL

For the first 14 days, we have information on where each patient was:

- Discharge/At Home
- Hospital
- Hospital + ICU
- Death

By day 14, 146 patients (9.5%) in the study were in the death state



Who are we going to sample?

- Budget and time constraints allowed us to collect information on glycocalyx degradation for 600 of the 1,563 patients enrolled in the CLOVERS study

Who are we going to sample?

- Budget and time constraints allowed us to collect information on glycocalyx degradation for 600 of the 1,563 patients enrolled in the CLOVERS study

Everyone who dies and/or develops ARDS are sampled with probability one. The remaining patients are sampled using simple random sampling until we reach a total of 600 patients

How are we analysing the data?

- We want to understand the relationship between glycocalyx degradation (yes/no) and mortality at the beginning (day 1) and at the end (day 14) of the study.

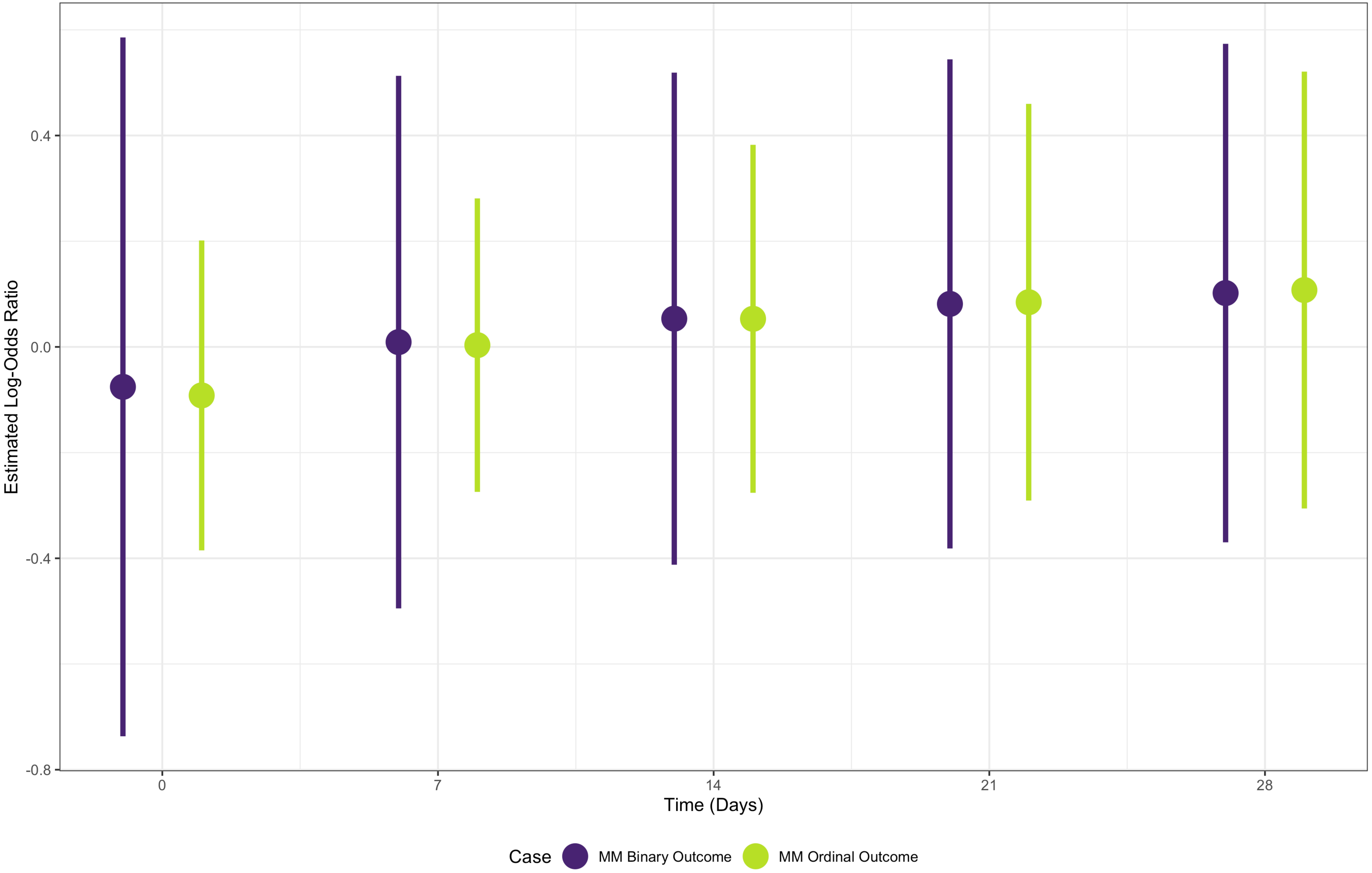
- We define the outcome in two ways:

    - **Longitudinal binary outcome:** death vs alive (hospital, hospital + ICU, discharge/at home)

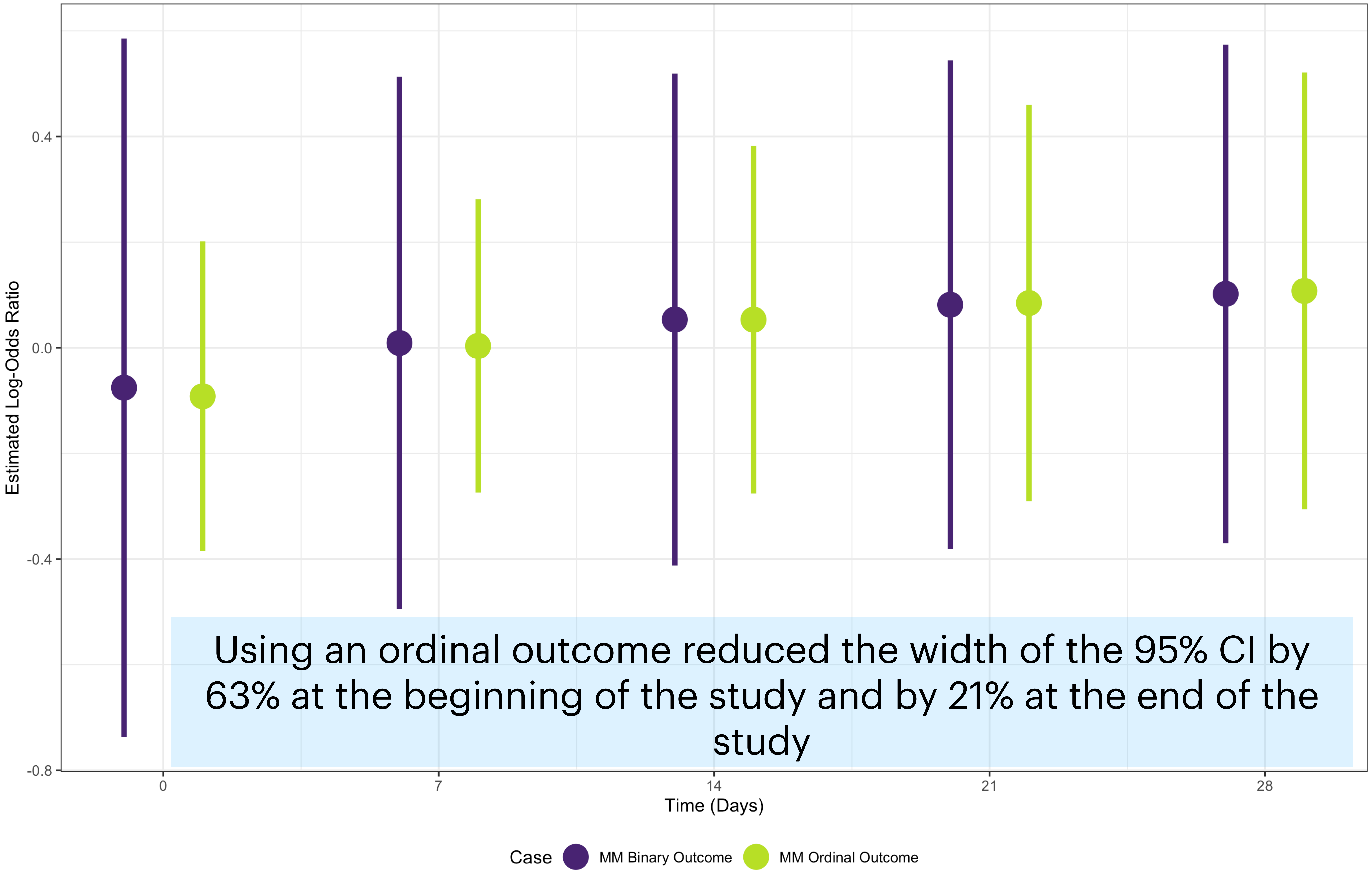
    - **Longitudinal ordinal outcome:** death, hospital + ICU, hospital, discharge/at home

IMPERIAL

Why do we want to include every health state?



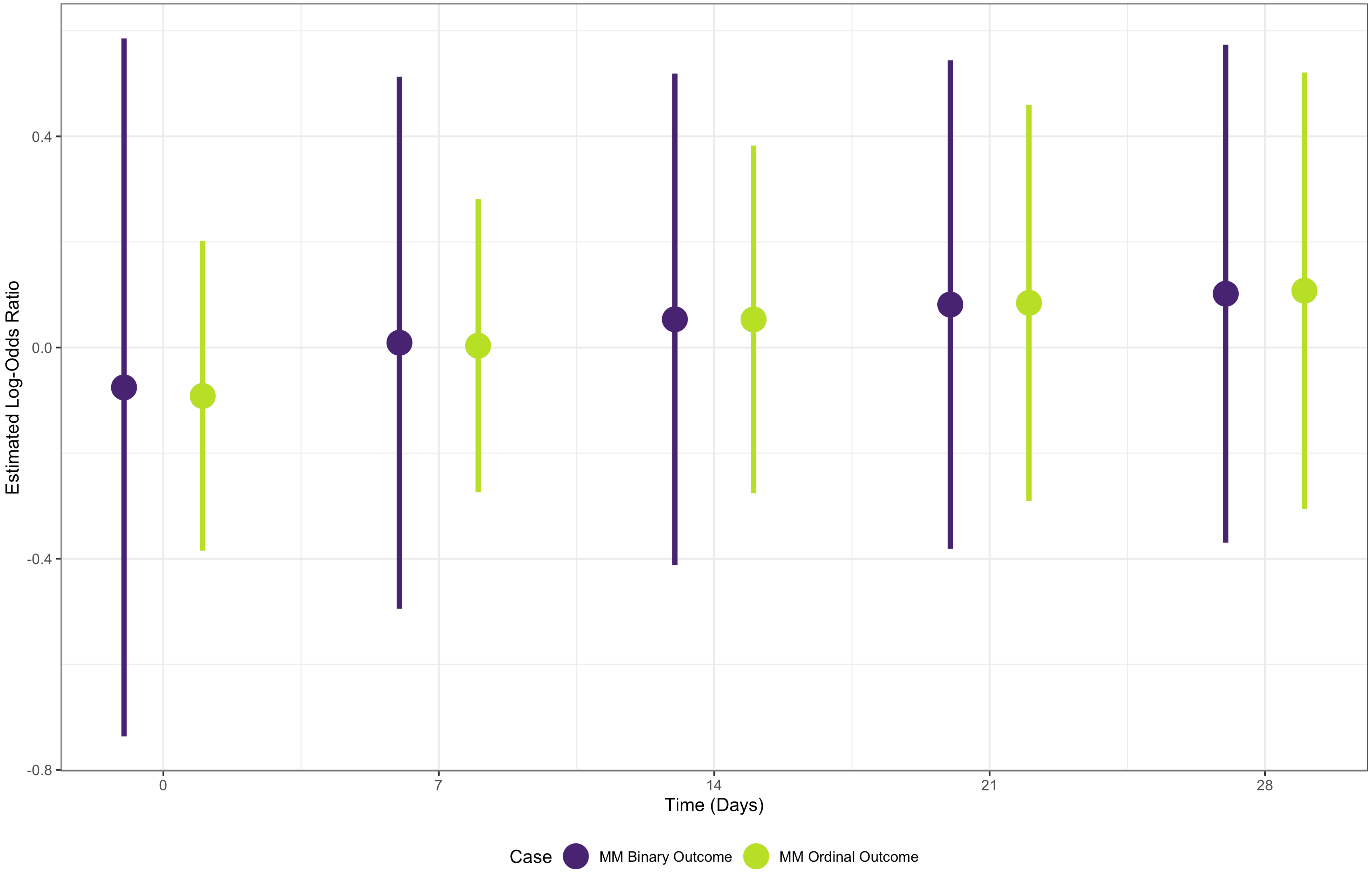
Why do we want to include every health state?



Using an ordinal outcome reduced the width of the 95% CI by 63% at the beginning of the study and by 21% at the end of the study

IMPERIAL

Why do we want to include every health state?



The Marginalised Transition Model

The Model

The marginalized transition model is identified by two generalized linear models:

$$h\{E(Y_{ij}\,|\,X_i, \boldsymbol{Z}_i, T_{ij})\} = \alpha_0 + \beta_x X_i + \beta_t T_{ij} + \beta_{xt} X_i T_{ij} + \boldsymbol{\beta}_z^{\mathrm{T}} \boldsymbol{Z}_i$$

$$g\{E(Y_{ij}\,|\,X_i, \boldsymbol{Z}_i, Y_{i(j-1)})\} = \Delta_{ijk} + \sum_{s=1}^{K-1} \gamma^{ks} I\{Y_{i(j-1)} = s\}$$

where:

- Y_{ij} is the outcome state for subject i at time j
- K is the total number of states (K = 1, 2, 3, 4)
- X_i is an indicator of the presence of glycocalyx degradation
- $\boldsymbol{Z}_i$ is a matrix of baseline covariates: SOFA score, age, gender, ARDS
- Δ_{ijk} links the marginal and the conditional mean model

The Model

The marginalized transition model is identified by two generalized linear models:

Marginal Mean Model

$$h\{E(Y_{ij}\,|\,X_i, \boldsymbol{Z}_i, T_{ij})\} = \alpha_0 + \beta_x X_i + \beta_t T_{ij} + \beta_{xt} X_i T_{ij} + \boldsymbol{\beta}_z^{\mathrm{T}} \boldsymbol{Z}_i$$

$$g\{E(Y_{ij}\,|\,X_i, \boldsymbol{Z}_i, Y_{i(j-1)})\} = \Delta_{ijk} + \sum_{s=1}^{K-1} \gamma^{ks} I\{Y_{i(j-1)} = s\}$$

where:

- Y_{ij} is the outcome state for subject i at time j
- K is the total number of states (K = 1, 2, 3, 4)
- X_i is an indicator of the presence of glycocalyx degradation
- $\boldsymbol{Z}_i$ is a matrix of baseline covariates: SOFA score, age, gender, ARDS
- Δ_{ijk} links the marginal and the conditional mean model

The Model

The marginalized transition model is identified by two generalized linear models:

Marginal Mean Model $h\{E(Y_{ij}|X_i, \boldsymbol{Z}_i, T_{ij})\} = \alpha_0 + \beta_x X_i + \beta_t T_{ij} + \beta_{xt} X_i T_{ij} + \boldsymbol{\beta}_z^{\mathrm{T}} \boldsymbol{Z}_i$

Conditional Response Dependence Model $g\{E(Y_{ij}|X_i, \boldsymbol{Z}_i, Y_{i(j-1)})\} = \Delta_{ijk} + \sum_{s=1}^{K-1} \gamma^{ks} I\{Y_{i(j-1)} = s\}$

where:

- Y_{ij} is the outcome state for subject i at time j
- K is the total number of states (K = 1, 2, 3, 4)
- X_i is an indicator of the presence of glycocalyx degradation
- $\boldsymbol{Z}_i$ is a matrix of baseline covariates: SOFA score, age, gender, ARDS
- Δ_{ijk} links the marginal and the conditional mean model

A natural choice for the link functions is the logit link:

$$\operatorname{logit}\{P(\boldsymbol{Y}_i \leq k \,|\, X_i, T_{ij}, \boldsymbol{Z}_i)\} = \alpha_{0,k} + \beta_x X_i + \beta_t T_{ij} + \beta_{xt} X_i T_{ij} + \boldsymbol{\beta}_z^{\mathrm{T}} \boldsymbol{Z}_i$$

$$\log\left\{\frac{P(Y_{ij} = k \,|\, Y_{i(j-1)}, X_i, T_{ij}, \boldsymbol{Z}_i)}{P(Y_{ij} = K \,|\, Y_{i(j-1)}, X_i, T_{ij}, \boldsymbol{Z}_i)}\right\} = \Delta_{ijk} + \sum_{s=1}^{K-1} \gamma^{ks} I\{Y_{i(j-1)} = s\}$$

When K = 2 (binary case), the marginal mean model and the conditional, response dependence model become logistic regression models

When K > 2 (ordinal case), the marginal mean model is a proportional odds model while the conditional, response dependence model is a multinomial regression

Dealing with Absorbing States

The marginal mean model assumes that the association between the ordinal outcome and time is captured by a single coefficient. This is **not true** when there are absorbing states

We relax the proportional odds assumption for time in the marginal mean model

$$\begin{aligned}\operatorname{logit}\{P(\boldsymbol{Y}_i \le k \mid X_i, T_{ij}, \boldsymbol{Z}_i)\} = \alpha_0 + \beta_x X_i + T_{ij}[\beta_{t,1} + \beta_{t,2} I(Y=2) + \beta_{t,3} I(Y=3)] \\ + \beta_{xt} X_i T_{ij} + \boldsymbol{\beta}_z^{\mathrm{T}} \boldsymbol{Z}_i\end{aligned}$$

The difference in the log-odds of death for those with and without glycocalyx degradation at time T_{ij} is:

$$\beta_x + T_{ij}\beta_{xt}$$

We can relax the proportional odds assumptions for other variables

Dealing with Absorbing States

The marginal mean model assumes that the association between the ordinal outcome and time is captured by a single coefficient. This is **not true** when there are absorbing states

We relax the proportional odds assumption for time in the marginal mean model

$$\text{logit}\{P(\boldsymbol{Y}_i \leq k \,|\, X_i, T_{ij}, \boldsymbol{Z}_i)\} = \alpha_0 + \beta_x X_i + T_{ij}[\beta_{t,1} + \beta_{t,2} I(Y=2) + \beta_{t,3} I(Y=3)]$$
$$+\beta_{xt} X_i T_{ij} + \boldsymbol{\beta}_z^{\mathrm{T}} \boldsymbol{Z}_i$$

The difference in the log-odds of death for those with and without glycocalyx degradation at time T_{ij} is:

$$\beta_x + T_{ij}\beta_{xt}$$

We can relax the proportional odds assumptions for other variables

Dealing with Absorbing States

The marginal mean model assumes that the association between the ordinal outcome and time is captured by a single coefficient. This is **not true** when there are absorbing states

We relax the proportional odds assumption for time in the marginal mean model

$$\text{logit}\{P(\boldsymbol{Y}_i \leq k \,|\, X_i, T_{ij}, \boldsymbol{Z}_i)\} = \alpha_0 + \beta_x X_i + T_{ij}[\beta_{t,1} + \beta_{t,2} I(Y=2) + \beta_{t,3} I(Y=3)]$$
$$+\beta_{xt} X_i T_{ij} + \boldsymbol{\beta}_z^{\mathrm{T}} \boldsymbol{Z}_i$$

The difference in the log-odds of death for those with and without glycocalyx degradation at time T_{ij} is:

$$\beta_x + T_{ij}\beta_{xt}$$

We can relax the proportional odds assumptions for other variables

Estimation Procedures

The Estimation Procedure

- Parameters estimation needs to account for the study design

- Together with the study design and the modelling choice, one can increase estimation efficiency by choosing an analysis procedure that uses all the information available

IMPERIAL

The Estimation Procedure

- Parameters estimation needs to account for the study design
- Together with the study design and the modelling choice, one can increase estimation efficiency by choosing an analysis procedure that uses all the information available

Complete Data Analysis (IPW)

Consider the outcome, exposure and covariates for the 600 patients with information on glycocalyx degradation. For each participant, weight their contribution to the likelihood by the inverse of their sampling probability

The Estimation Procedure

- Parameters estimation needs to account for the study design
- Together with the study design and the modelling choice, one can increase estimation efficiency by choosing an analysis procedure that uses all the information available

Complete Data Analysis (IPW)

Consider the outcome, exposure and covariates for the 600 patients with information on glycocalyx degradation. For each participant, weight their contribution to the likelihood by the inverse of their sampling probability

Full-Data Analysis (SMLE and MI)

Consider the outcome, exposure and covariate data for 600 patients with information on glycocalyx degradation together with the outcome and covariates data for the remaining patients

IMPERIAL

Full Data Analysis: The Sieve Maximum Likelihood Estimator (SMLE)

Let V be an indicator of whether a subject has the exposure X measured

$$\underbrace{\sum_{i=1}^{n} V_i \left\{ log P_\beta(\boldsymbol{Y}_i \,|\, X_i, \boldsymbol{Z}_i) G(X_i \,|\, \boldsymbol{Z}_i) \right\}}_{\text{Contribution of Sampled Subjects}} + \underbrace{\sum_{i=1}^{n} (1 - V_i) \left[log \int_x P_\beta(\boldsymbol{Y}_i \,|\, \boldsymbol{x}, \boldsymbol{Z}_i) G(\boldsymbol{x} \,|\, \boldsymbol{Z}_i) \right]}_{\text{Contribution of Unsampled Subjects}}$$

We extend the **Sieve Maximum Likelihood Estimator (SMLE)** from Tao et al (2017).

We estimate $P_\beta(\boldsymbol{Y}_i \,|\, X_i, \boldsymbol{Z}_i)$ parametrically using a marginalized transition model

We estimate $G(X_i \,|\, \boldsymbol{Z}_i)$ non-parametrically by considering the distinct observed values of glycocalyx degradation and using the method of sieves and B-spline basis

Full Data Analysis: The Sieve Maximum Likelihood Estimator (SMLE)

To estimate $G(X|\mathbf{Z})$ we use B-spline basis to construct the approximating function. If $B_l^q(\mathbf{Z}_i)$ is the lth B-spline of order q, then:

$$logG(X_i|\mathbf{Z}_i) \approx \sum_{w=1}^{m} I(X_i = x_w) \sum_{l=1}^{s_n} B_l^q(\mathbf{Z}_i) logp_{wl}$$

$$G(x_i|\mathbf{Z}_i) \approx \sum_{w=1}^{m} I(X_i = x_w) \sum_{l=1}^{s_n} B_l^q(\mathbf{Z}_i) p_{wl}$$

where

- s_n is the total number of functions in the B-spline basis
- p_{wl} is the coefficient associated with the B-spline term $B_l^q(\mathbf{Z}_i)$ at $X_i = x_w$

Full Data Analysis: The Sieve Maximum Likelihood Estimator (SMLE)

$$\sum_{i=1}^{n} V_i \left\{ logP_\beta(\mathbf{Y}_i|X_i,\mathbf{Z}_i) + \sum_{w=1}^{m}\sum_{l=1}^{s_n} I(X_i = x_w)B_l^q(\mathbf{Z}_i)logp_{wl} \right\} + \sum_{i=1}^{n}(1-V_i)\left[log\left(\sum_{w=1}^{m} I(X_i = x_w)P_\beta(\mathbf{Y}_i|x_w,\mathbf{Z}_i)\sum_{l=1}^{s_n}B_l^q(\mathbf{Z}_i)\right)\right]$$

Direct maximisation of this likelihood is difficult

We introduce a latent variable $W = \{1/s_n, \ldots, 1\}$ such that the second term can be interpreted as the log-likelihood of $(Y_i, \mathbf{Z}_i)$ assuming that the complete data consist of $(Y_i, X_i, \mathbf{Z}_i, W_i)$ but X_i and W_i are missing

We estimate the parameters β using the EM algorithm

We estimate $Cov(\beta)$ using the profile likelihood method from Murphy et al (2000)

Full Data Analysis: Multiple Imputation (MI)

- Participants with data on glycocalyx degradation were selected based on their observed outcome

- **Data on glycocalyx degradation are missing at random**

- Multiple imputation is a valid alternative to SMLE

IMPERIAL

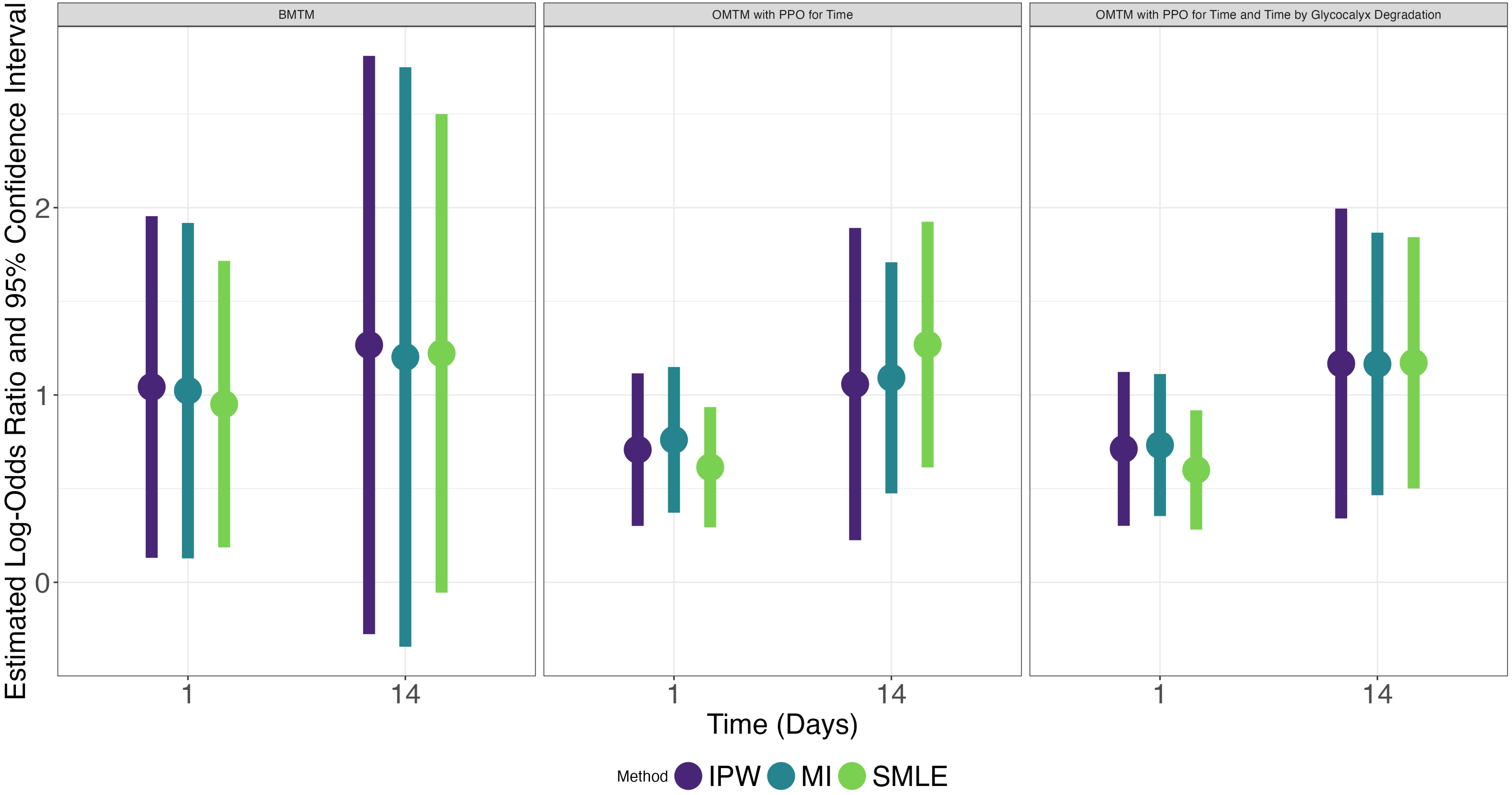
Results

We consider **three** models:

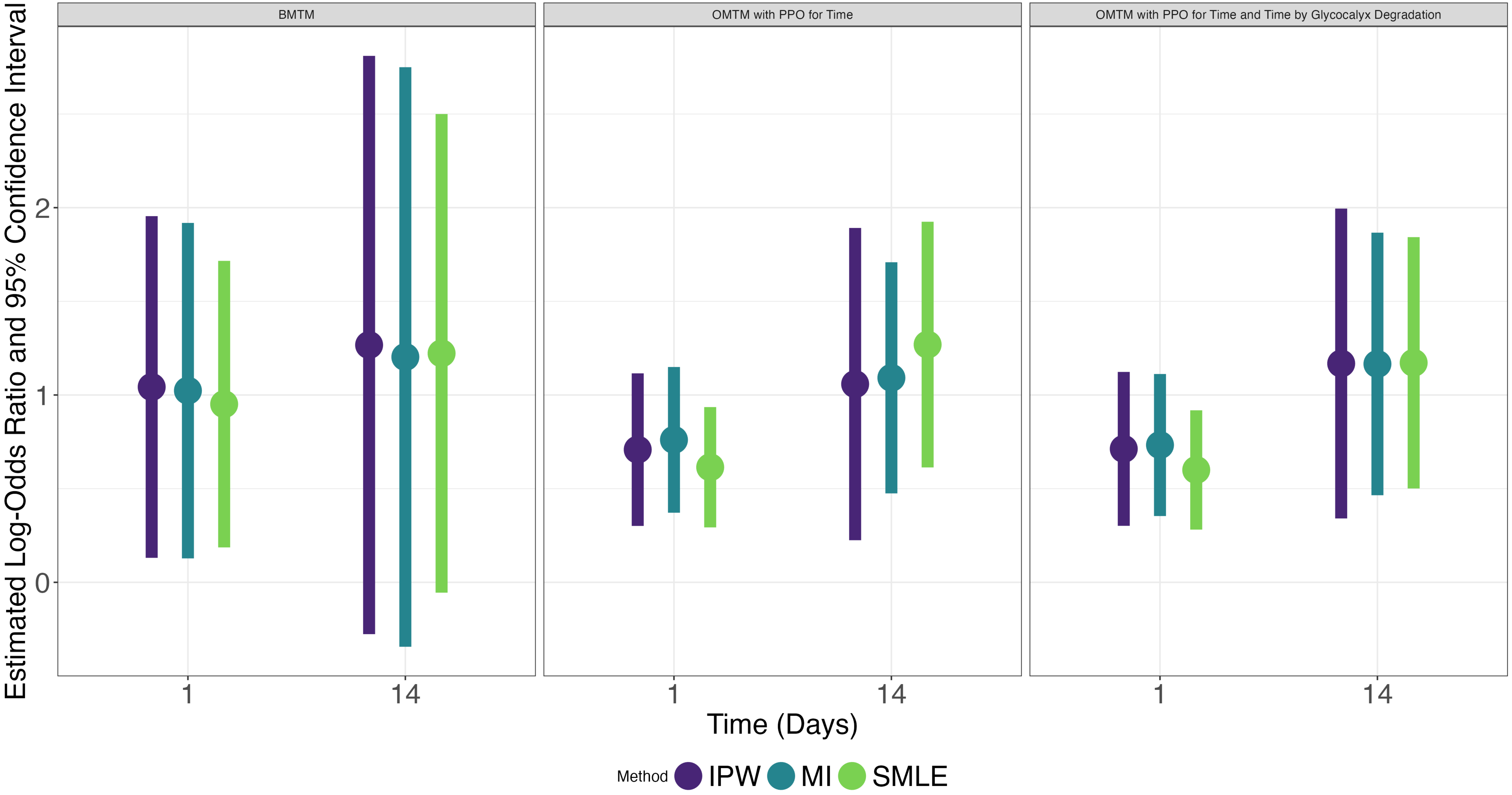
- Marginalised transition model with a binary outcome (death vs alive) (BMTM)
- Two marginalised transition models with an ordinal outcome (OMTM)
    - Relax the proportional odds assumption for association between time and mortality
    - Relax the proportional odds assumption for the association between time and mortality and the association between glycocalyx degradation and mortality

We estimate the association between glycocalyx degradation and mortality at day 1 and at day 14 with IPW, SMLE and MI

All models are adjusted for glycocalyx degradation, time, age, sex, ARDS, SOFA score and time by glycocalyx degradation

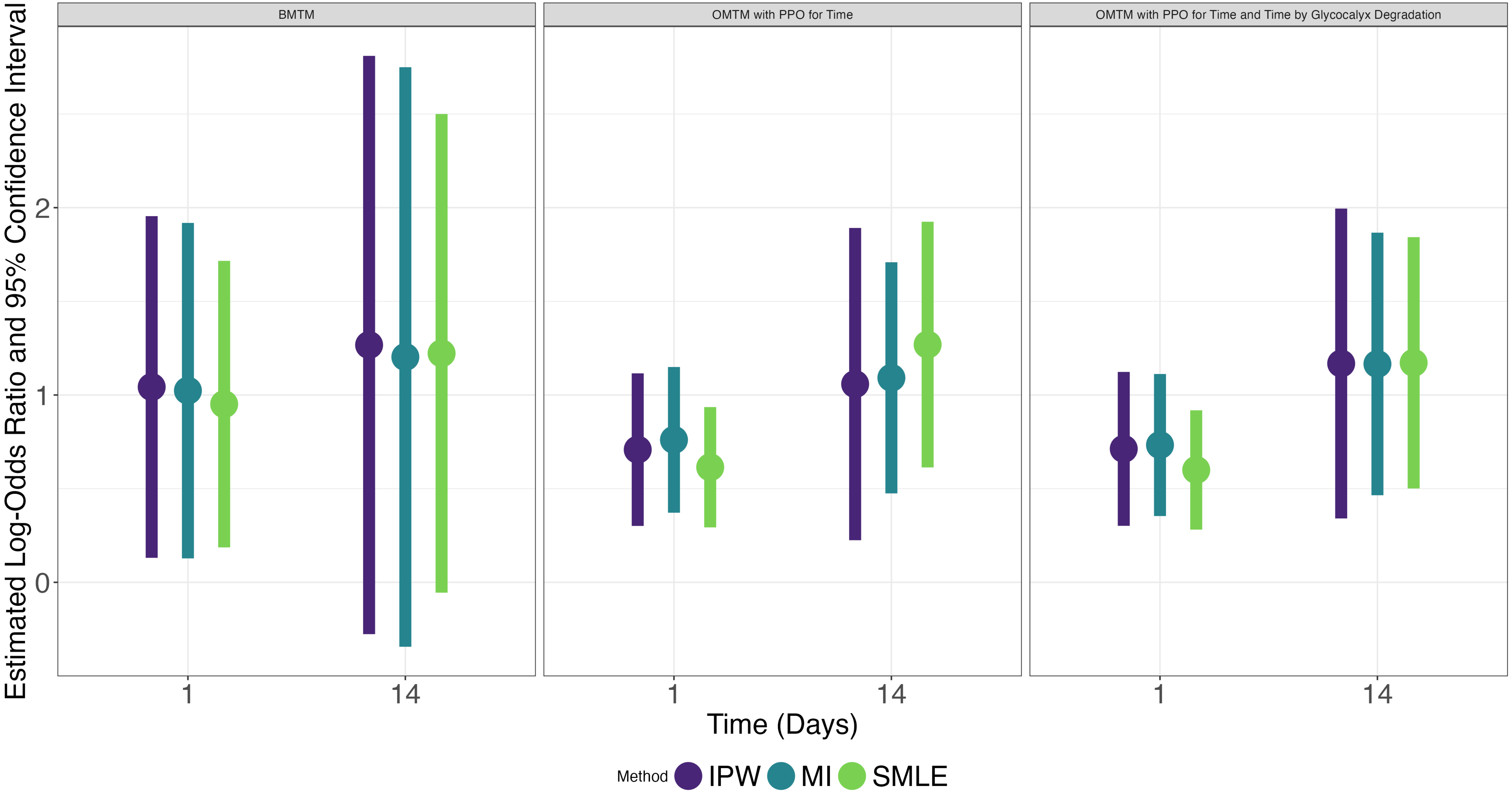

BMTM
OMTM with PPO for Time
OMTM with PPO for Time and Time by Glycocalyx Degradation
Estimated Log-Odds Ratio and 95% Confidence Interval
0
1
2
1
14
Time (Days)
Method
IPW
MI
SMLE

IMPERIAL



At day 1 people with glycocalyx degradation had higher odds of mortality

IMPERIAL

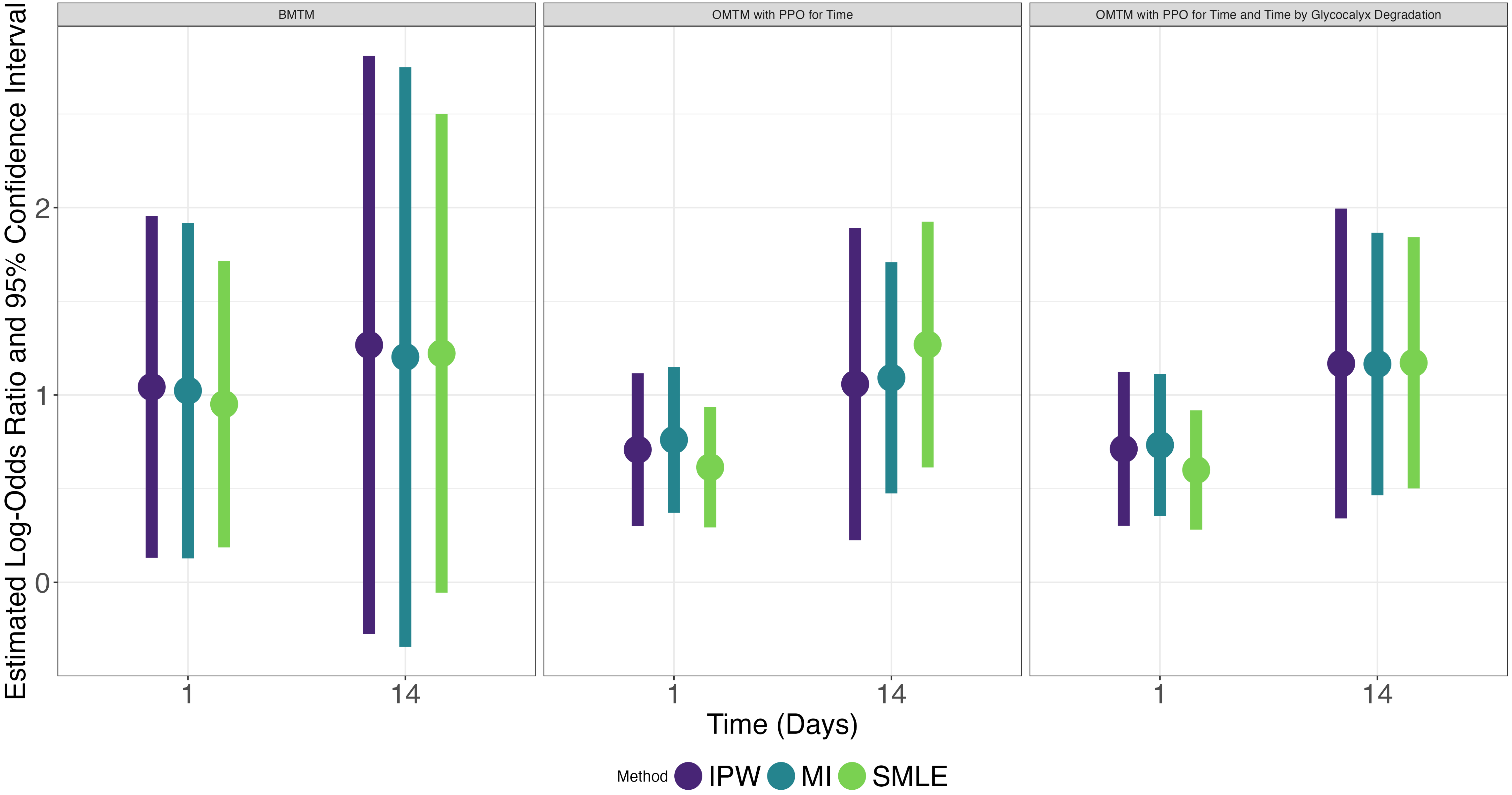


At day 1 people with glycocalyx degradation had higher odds of mortality

Using an ordinal outcome reduced the width of the CI at day 1 and day 14.

The two OMTM models had similar efficiency

IMPERIAL



At day 1 people with glycocalyx degradation had higher odds of mortality

Using an ordinal outcome reduced the width of the CI at day 1 and day 14.

The two OMTM models had similar efficiency

In OMTM, SMLE and MI reduced the width of the CI by 17% at day 14

Conclusion

- Marginalized transition models with longitudinal binary or ordinal outcomes were used to estimate the association between glycocalyx degradation and mortality
- Estimation efficiency can be increased when using an ordinal outcome rather than a binary outcome
- When all available information is included in the estimation procedure (SMLE or MI), we observed efficiency gains compared to methods that only include participants with complete data on the outcome, exposure and all other covariates.

Thank you!

Email: c.di-gravio@imperial.ac.uk

Reference

Schildcrout et al (2022). Model-assisted analyses of longitudinal, ordinal outcomes with absorbing states. Statistics in Medicine 41(14).

NHLBIP and Early Treatment of Acute Lung Injury Clinical Trial Network (2023). Early Restrictive or liberal fluid management for sepsis-induced hypotension. NEJM.

Di Gravio et al (2024) Efficient Designs and Analysis of Two-Phase Studies with Longitudinal Binary Outcome. Biometrics 80(1)

Tao R, Zeng D, Lin DY. Efficient Semiparametric Inference Under Two-Phase Sampling, With Applications to Genetic Association Studies. J Am Stat Assoc. 2017;112(520):1468-1476. doi: 10.1080/01621459.2017.1295864. Epub 2017 Feb 28. PMID: 29479125; PMCID: PMC5823539.