Imputing right skewed bounded biomarkers in partially measured cohorts

Nicola Orsini

Department of Global Public Health
Karolinska Institutet

2025 Northern European Stata Conference, Stockholm

August 29, 2025

Outline

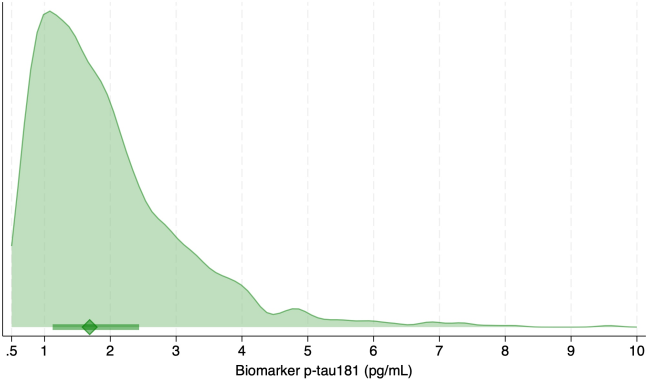
- Context
- Truncated Log Normal Imputation
- Logistic Quantile Imputation
- Simulation study
- Final remarks

Context

A growing number of studies show that blood biomarkers for Alzheimer's disease — such as **plasma phosphorylated tau-181 (ptau181)** — are associated with neuropathologic changes in the brain.
Biomarkers can be useful for:

- Speed and accuracy in diagnosis: improve sensitivity/specificity, enable earlier detection
- Risk stratification and prognosis: identify who is likely to develop, progress, or relapse
- Guide treatment choices: predict who will benefit and who might be harmed
- Monitor disease and therapy: track activity over time without waiting for clinical endpoint
- Trial enrichment: select participants with underlying Alzheimer's disease biology, boosting power and lowering sample sizes

Distribution of the biomarker



Plasma p-tau181 is a positive-valued, right-skewed, bounded to the range [0.5, 10] pg/mL.

The distribution has been simulated according to data from 2,000 Swedish adults (*Nature Medicine*, 2025).

Key features of the investigation

- Due to the high cost of essays the biomarker is typically measured in a small fraction of the available cohort
- The distribution of the biomarker ptau181 shifts upward with age and particularly with worse health conditions.
- The biomarker ptau181 is less likely to be measured among older and worse health conditions.
- The incidence of dementia is likely to increase with ptau181 up to about 2 pg/mL and then levels off upon adjustment for age, health conditions, and female sex.

Mechanisms underlying biomarker values, missingness, and outcomes

$$\textbf{Biomarker} \longleftarrow \text{Older Age} + \text{Worst Health}$$

$$\textbf{Missing Biomarker} \longleftarrow \text{Older Age} + \text{Worst Health}$$

$$\textbf{Dementia} \longleftarrow f(\textbf{Biomarker}) + \text{Older Age} + \text{Worst Health} + \text{Female}$$

Key questions

1. What is the impact of missing in studying the main distributional features of the biomarker?
2. What is the impact of missing in investigating a possible non-linear effect of the biomarker on the incidence of dementia?

A mechanism underlying the truncated biomarker

Define $A_i \in \{0, 1\}$ (older age) and $W_i \in \{0, 1\}$ (worse health).

$$A_i \sim \mathrm{Bernoulli}(0.6)$$
$$W_i \sim \mathrm{Bernoulli}(0.4)$$

Here Y_i denotes plasma p-tau181 (pg/mL).

$$Y_i \mid A_i, W_i \sim \mathrm{LogNormal}(\mu_i, \sigma) \text{ truncated to } [0.5, 10] \text{ pg/mL}$$
$$\mu_i = \alpha_0 \; + \; \alpha_1 \, A_i \; + \; \alpha_2 \, W_i$$
$$\alpha_0 = 0.2 \; \alpha_1 = 0.3 \; \alpha_2 = 0.5$$
$$\sigma = 0.5$$

The above model implies a positive, right-skewed distribution with additive shifts by A_i and W_i on the natural log scale.

A plausible mechanism underlying *missing* biomarker

Let $R_i = 1$ if p-tau181 is *missing* for subject i and 0 otherwise. We assume Missing at Random (MAR) given predictors.

Missing biomarker increases among older individuals and with worst health conditions

$$\Pr(R_i = 1 \mid A_i, W_i) = \text{logit}^{-1}\{\text{logit}(0.30) + \log(2)A_i + \log(3)W_i\}$$

Implied missingness fractions (approx.)

Group	(A, W)	$\Pr(R = 1)$
Younger–Better Health	$(0, 0)$	0.30
Older–Better Health	$(1, 0)$	0.46
Younger–Worse Health	$(0, 1)$	0.56
Older–Worse Health	$(1, 1)$	0.72

Marginally, about **50%** of p-tau181 measurements are missing.

Truncated Normal Imputation with `mi impute truncreg`

Let $Z_i = \log(Y_i)$ be imputed on the log scale $[\ell, u] = [\log L, \log U]$ with

$$Z_i \mid X \sim \mathcal{N}_{[\ell,u]}(\mu_i, \sigma^2)$$

$$\mu_i = X^\top \beta$$

Steps

1. *Estimate* truncated normal regression on observed Z_i obtaining MLEs $\hat{\theta} = (\hat{\beta}, \widehat{\ln \sigma})$ and covariance $\hat{U}$
2. *Draw parameters* $\theta^\star \sim \mathcal{N}(\hat{\theta}, \hat{U})$
3. *Draw a value* from $Z_i^{(m)} \sim \mathcal{N}_{[\ell,u]}(\mu_i^\star, \sigma^\star)$ with $\mu_i^\star = X_i^\top \beta^\star$
4. *Back-transform* $Y_i^{(m)} = \exp\left(Z_i^{(m)}\right)$

Logistic Quantile Imputation with `mi impute lqreg`

It is based on quantile regression (Bottai & Zhen, 2013) upon transformation of the bounded variable $Y \in [L, U]$ using a logistic transformation (Bottai et al, 2010; Orsini & Bottai, 2011):

$$logit(Y) = \log\left(\frac{Y - L}{U - Y}\right)$$

For each missing value of Y, do the following:

1. *Draw* a random number p from a continuous uniform distribution

$$p \sim \text{Uniform}(0, 1)$$

2. *Estimate* the p-quantile for the $logit(Y)$ conditionally on predictors X

$$Q_{logit(Y)}(p \mid X) = X^\top \hat{\beta}_p$$

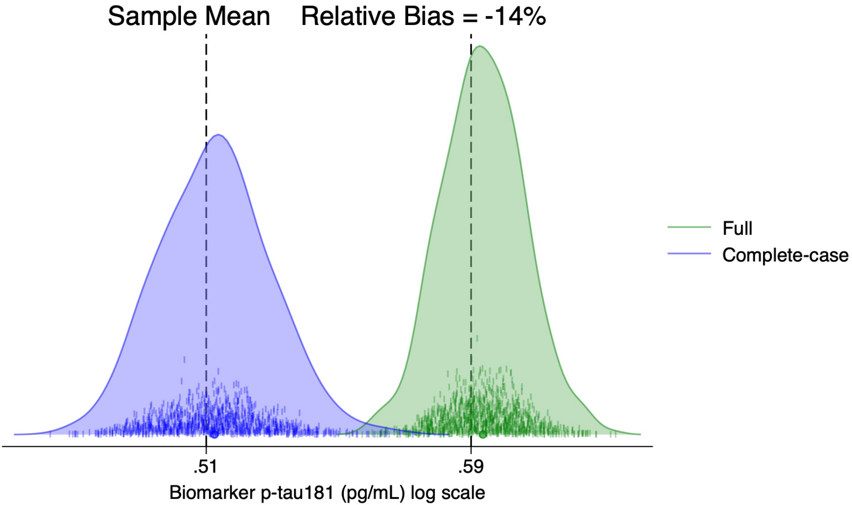
3. *Replace* the missing value with the inverse of the logit transformation:

$$Y_i^{(m)} = \frac{\exp(X^\top \hat{\beta}_p)\, U + L}{1 + \exp(X^\top \hat{\beta}_p)}$$

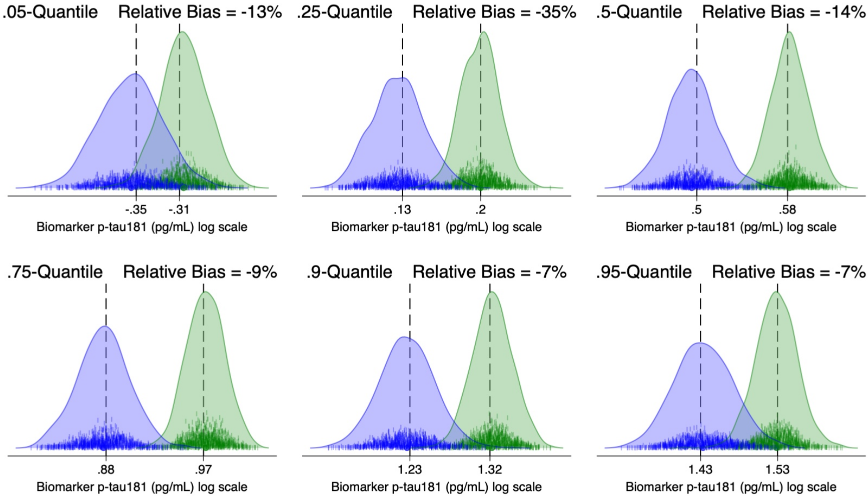
Pseudo-code to generate one sample

```
Input:
  N = 2000
  Parameters: a0 = 0.20, a1 = 0.30, a2 = 0.50, sigma = 0.50
  Bounds (original scale): L = 0.5, U = 10 (log scale): l = ln(L), u = ln(U)

For i = 1,...,N:
  Draw A_i ~ Bernoulli(0.6)        # Older age (old)
  Draw W_i ~ Bernoulli(0.4)        # Worst health (bh)

  # Linear predictor on log scale
  mu_i = a0 + a1*A_i + a2*W_i

  # Truncation CDF limits under Normal(mu_i, sigma^2)
  Fa_i = Phi( (l - mu_i)/sigma )
  Fb_i = Phi( (u - mu_i)/sigma )

  # Inverse-CDF draw on the truncated normal for Z_i = ln(Y_i)
  U_i = Uniform(0,1)
  Z_i  = mu_i + sigma * Phi^{-1}( Fa_i + U_i * (Fb_i - Fa_i) )

  # Back-transform to original scale (pg/mL)
  Y_i  = exp(Z_i)   # p-tau181

  # MAR
  Draw R_i ~ Bernoulli(logit(0.3)+ln(2)*A_i + ln(3)*W_i)

Output:
  Dataset {Y_i, A_i, W_i, R_i}
```

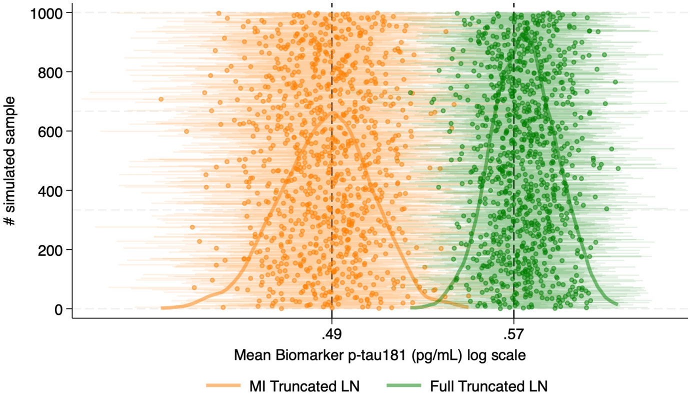
Sample mean biomarker in complete-case data is lower than full data



All empirical quantiles of the biomarker are shifted downward in the complete-case data



Ignoring why data are missing misleads inference

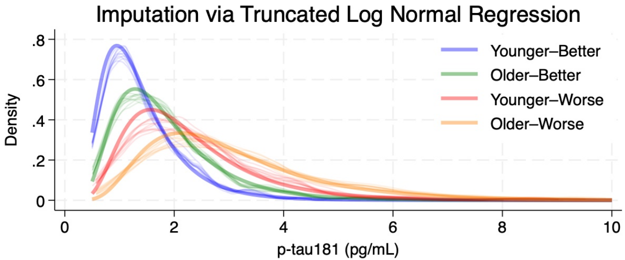


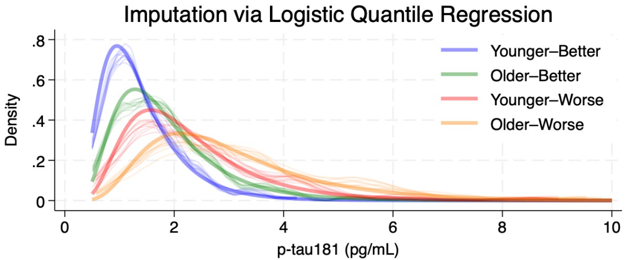
None of the MI-based 95% confidence intervals **include** the full-data mean of the biomarker.

Key syntax for imputation conditionally on covariates

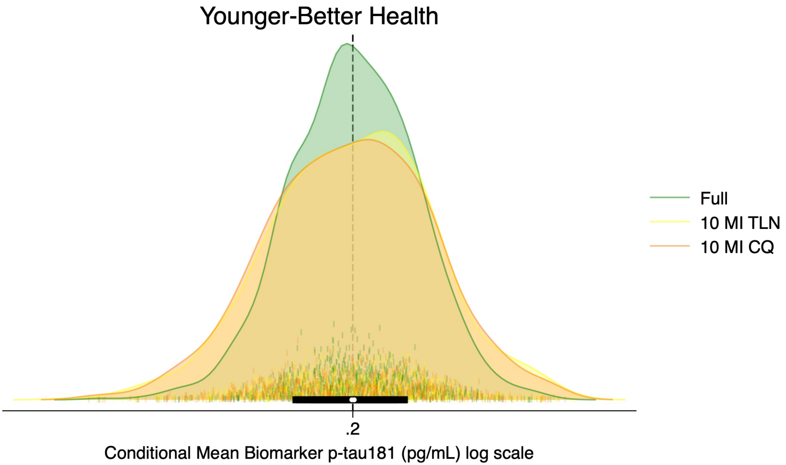
```
* Truncated Log Normal Imputation

mi impute truncreg ln_ptau181 old wh , ll(-0.693) ul(2.303)

* Conditional Logistic Quantile Imputation

mi impute lqreg ptau181 old wh , ll(0.5) ul(10)
```

Similarities of theoretical and imputed densities

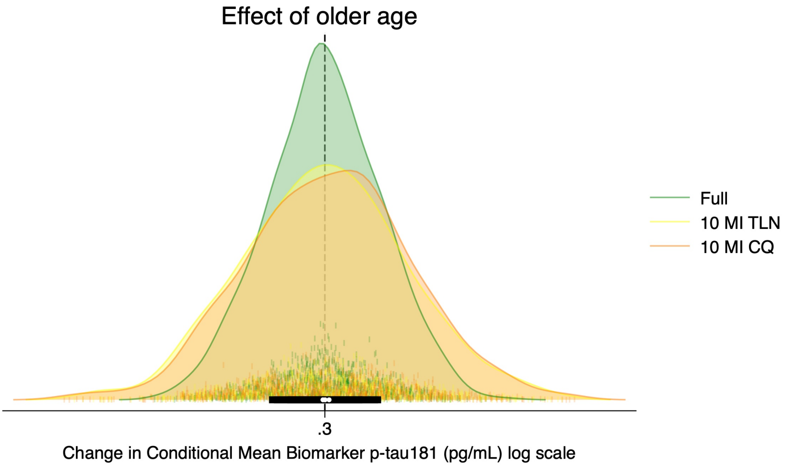




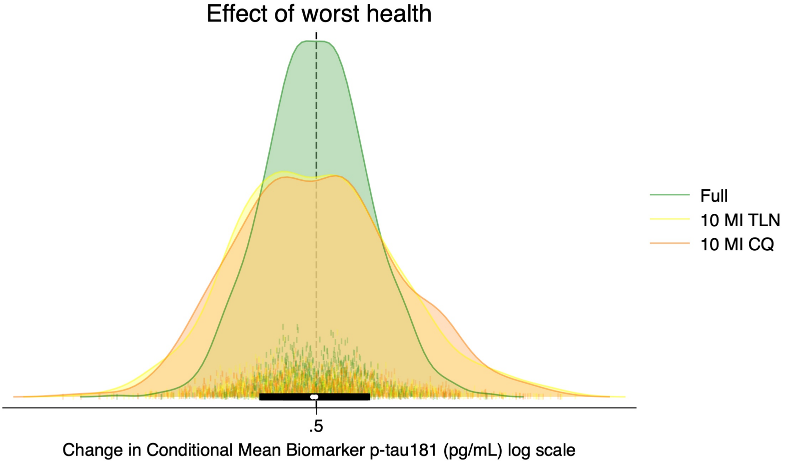
1,000 sample estimates $\hat{\alpha}_0$ generated under $\alpha_0 = 0.2$



1,000 sample estimates $\hat{\alpha}_1$ generated under $\alpha_1 = 0.3$



1,000 sample estimates $\hat{\alpha}_2$ generated under $\alpha_2 = 0.5$



Performance measures for $\hat{\alpha}_0$ generated under $\alpha_0 = 0.2$

Performance measure	Full	CC	MI LQ	MI TLN
Bias in point estimate	0.0002	-0.0004	-0.0012	-0.0005
% bias in point estimate	0.0974	-0.2178	-0.6146	-0.2444
Mean of point estimate	0.2002	0.1996	0.1988	0.1995
Empirical standard error	0.0216	0.0272	0.0277	0.0279
RMS model-based standard error	0.0216	0.0277	0.0281	0.0283
Coverage of 95% CI (%)	95.4	95.4	94.9	94.2

Performance measures for $\hat{\alpha}_1$ generated under $\alpha_1 = 0.3$

Performance measure	Full	CC	MI LQ	MI TLN
Bias in point estimate	-0.0002	0.0000	0.0015	-0.0001
% bias in point estimate	-0.0608	0.0077	0.5089	-0.0385
Mean of point estimate	0.2998	0.3000	0.3015	0.2999
Empirical standard error	0.0232	0.0330	0.0339	0.0339
RMS model-based standard error	0.0240	0.0335	0.0340	0.0345
Coverage of nominal 95% CI (%)	95.7	95.6	94.5	95.0

Performance measures for $\hat{\alpha}_2$ generated under $\alpha_2 = 0.5$

Performance measure	Full	CC	MI LQ	MI TLN
Bias in point estimate	-0.0000	-0.0000	0.0008	0.0001
% bias in point estimate	-0.0100	-0.0033	0.1568	0.0121
Mean of point estimate	0.5000	0.5000	0.5008	0.5001
Empirical standard error	0.0243	0.0369	0.0382	0.0381
RMS model-based standard error	0.0237	0.0369	0.0376	0.0379
Coverage of nominal 95% CI (%)	94.8	95.5	94.1	92.6

Key insights from performance tables

- **MI LQ** is nearly unbiased
- Model-based SEs are close to empirical SEs
- Coverage is near nominal

A mechanism underlying the survival outcome #1

Let $A_i, W_i, F_i \in \{0, 1\}$ denote old, worst health, and female, respectively.

Let's continue to denote Y_i the biomarker p-tau181 (pg/mL).

The linear predictor underlying the (log) dementia rate λ_i is

$$\log \lambda_i = \gamma_0 + \underbrace{\gamma_1 \, Y_i - \gamma_2 \, (Y_i - k)_+}_{\text{piecewise linear spline at } k} + \gamma_3 \, A_i + \gamma_4 \, W_i + \gamma_5 \, F_i$$

where $(Y_i - k)_+ = \max(Y_i - k, 0)$ is a linear spline with a knot at $k = 2$ pg/mL.

The (conditional) dementia rate increases by 20% for each 1 pg/mL increase in p-tau181 up to 2 pg/mL, after which the effect plateaus ($\gamma_2 = -\gamma_1$). Older age, worse health, and female sex lead to higher dementia rates independently of the biomarker. The regression coefficients are set to

$$\log \lambda_i = \log(-1.817) \mathbf{+ \log(1.2)} \; Y_i \mathbf{- \log(1.2)} \; (Y_i - k)_+ + \log(1.6) \, A_i + \log(2) \, W_i + \log(1.5) \, F_i$$

A mechanism underlying the survival outcome #2

Time elapsed from entry into the study to diagnosis of dementia is generated from an Exponential survival distribution $S(T_i) = e^{-\lambda_i T_i}$:

$$T_i \mid (A_i, W_i, F_i, Y_i) \sim \mathrm{Exponential}(\lambda_i)$$

by inverting the cumulative distribution function:

$$T_i \; = \; -\frac{\log(U_i)}{\lambda_i}, \qquad U_i \sim \mathrm{Unif}(0, 1)$$

Adding an administrative censoring at $C = 5$ years, we obtain the dementia-free time (years) and dementia indicator:

$$\widetilde{T}_i = \min(T_i, C) \qquad D_i = \mathbb{1}\{T_i < C\}$$

Target parameters in this simulation are γ_1 and γ_2 jointly defining the (adjusted) piecewise-linear effect of the biomarker ptau181 on the rate of dementia.

Imputation model for the biomarker

Based on the plausible missing mechanism underlying the biomarker and the survival model underlying dementia rate, the linear predictor X_i for the imputation model for ptau181 includes A (old age), W (Worst Health), log Cumulative Hazard (H), Dementia (D), and Female (F):

$$\begin{aligned} \mu_i &= \beta_0 + \beta_A A_i + \beta_W W_i + \beta_H H_i + \beta_D D_i + \beta_F F_i \\ &= X_i^\top \beta \end{aligned}$$

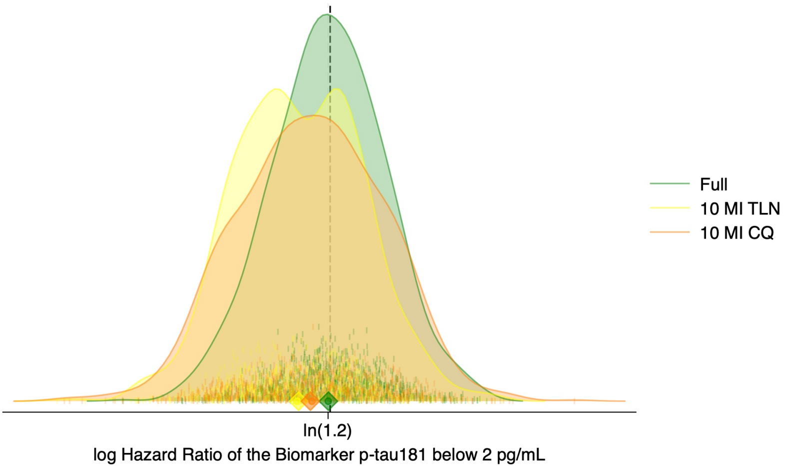
where $X_i = (1,\ A_i,\ W_i,\ H_i,\ D_i,\ F_i)^\top$.

The above linear predictor is used for both Truncated Log Normal Imputation and Logistic Quantile Imputation.

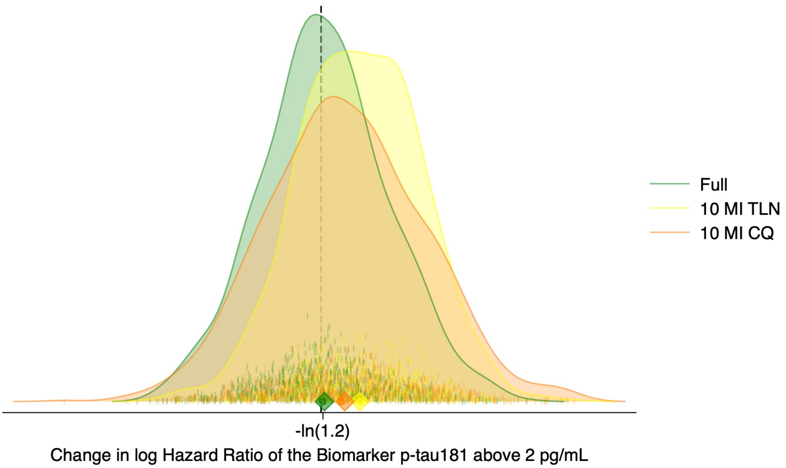
Key syntax for imputation

```
* Truncated Log Normal Imputation

mi impute truncreg ln_ptau181 old wh ///
   log_cumh dementia female, ll(-0.693) ul(2.303)

* Conditional Logistic Quantile Imputation

mi impute lqreg ptau181 old wh ///
   log_cumh dementia female, ll(0.5) ul(10)
```

1,000 sample estimates $\hat{\gamma}_1$ generated under $\gamma_1 = \log(1.2) = 0.182$



1,000 sample estimates $\hat{\gamma}_2$ generated under $\gamma_2 = -\log(1.2) = -0.182$



Performance measures for $\hat{\gamma}_1$ generated under $\gamma_1 = 0.182$

This is the linear trend for ptau181 before 2 pg/mL.

Performance measure	Full	CC	TLN	LQ
Bias in point estimate	-0.0019	-0.0023	-0.0355	-0.0222
% bias in point estimate	-1.0657	-1.2689	-19.4723	-12.1981
Mean of point estimate	0.1804	0.1800	0.1468	0.1601
Empirical standard error	0.0632	0.0905	0.0722	0.0812
RMS model-based standard error	0.0637	0.0877	0.0825	0.0853
Relative % error in standard error	0.7298	-3.0515	14.3342	5.1485
% coverage of 95% CI	94.7	95.0	96.2	96.1

Performance measures for $\hat{\gamma}_2$ generated under $\gamma_2 = -0.182$

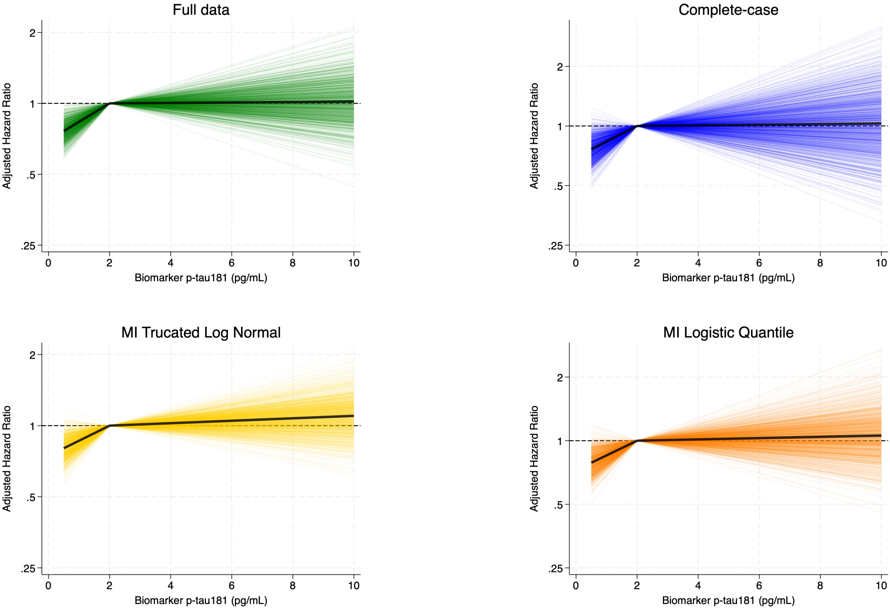
This is the change in linear trend for ptau181 above 2 pg/mL.

Performance measure	Full	CC	TLN	LQ
Bias in point estimate	0.0042	0.0061	0.0474	0.0292
% bias in point estimate	-2.3073	-3.3428	-26.0000	-16.0326
Mean of point estimate	-0.1781	-0.1762	-0.1349	-0.1531
Empirical standard error	0.0778	0.1175	0.0754	0.0956
RMS model-based standard error	0.0775	0.1121	0.0990	0.1052
% coverage of 95% CI	94.6	94.1	97.1	96.3

Summary: piecewise-linear biomarker effects (γ_1 pre-2 pg/mL, γ_2 change post-2 pg/mL)

- **Bias**
  - **Full, CC**: near-unbiased ($\approx$ 1–3%).
  - **MI TLN**: marked attenuation toward 0: γ_1 -19.5%, γ_2 -26.0%.
  - **MI LQ**: less biased than TLN: γ_1 -12.2%, γ_2 -16.0%.
- **Variance**
  - **Full**: smallest SEs ($\gamma_1 : 0.0632,\ \gamma_2 : 0.0778$).
  - **CC**: largest SEs ($\gamma_1 : 0.0905,\ \gamma_2 : 0.1175$).
  - **MI LQ**: improves vs CC but less precise than TLN ($\gamma_1 : 0.0812,\ \gamma_2 : 0.0956$).
- **Coverage**
  - **MI TLN**: model SEs $>$ empirical (rel. error $+14$–31%); coverage $\approx$ 96–97%.
  - **MI LQ**: modest SE overestimation ($+5$–14%); coverage $\approx$ 96%.

Piecewise-linear effect: a graphical comparison



Final comments

Based on this simulation study of and the current implementation of logistic quantile imputation:

- `mi impute lqreg` is a distribution-free imputation method based on quantile regression while respecting the bounds/truncations
- `mi impute lqreg` is computationally demanding (one estimation for each missing for each imputation)
- `mi impute lqreg` requires some observed data to estimate the imputation model
- `mi impute from` can be used to impute using external data (Thiesmeier, Bottai, Orsini, *SJ*, in press).
- A limitation of this simulation study is the limited number of imputations (M=10) relative to the fraction of missing data (about 50%). More simulation studies are needed.

Acknowledgement: Ongoing work with Robert Thiesmeier and Professor Matteo Bottai.

References

- Bottai, M., Cai, B. and McKeown, R. E. (2010). Logistic quantile regression for bounded outcomes. *Statistics in Medicine* 29, 2, 309–317.
- Bottai, M. and Zhen, H. (2013). Multiple imputation based on conditional quantile estimation. *Epidemiology, Biostatistics and Public Health*, 10(1), e8758.
- Thiesmeier R, Bottai M, Orsini N. (2025). Imputation when data cannot be pooled. *Stata Journal*. In Press.