Imputation of systematic missing data in individual participant data meta-analysis

Nicola Orsini

Department of Global Public Health
Karolinska Institutet

Northern European Stata Conference

September 1, 2023

Outline

- Rationale for a user-written imputation command
- `mi impute cqi`
- Application to sporadic and systematic missing data on smoking pack-years
- Application to Individual Participant Data (IPD) meta-analysis investigating effect modification
- Final remarks

What is the context?

- In epidemiological research, pooling or consortia projects are commonly conducted to answer complex research questions and to increase the statistical power to detect even small exposure/treatment effect
- Variation in the fraction of missing data across studies can be a challenge
- Nominal, categorical, discrete variables with missing data are widely used in research

Systematic missing data

- A key variable is 100% missing by design (not measured) in one or more studies
- Within-study imputation is not feasible in such studies

What imputation model to use for systematic missing data?

- Imputation model for the study with systematic missing has to be based on other studies with some information
- An imputation model, possibly similar across studies, should be in line with the complexity of outcome model
- Within-study estimates of the imputation model, possibly weighted by the precision of each study, should be combined before using them for imputation in other studies with systematic missing data

Direct method to generate imputations

- Inverse transform sampling
- Given a cumulative distribution function, a random sample from such distribution can be obtained by drawing random sample from a continuous uniform distribution over the interval (0, 1).
- The univariate conditional quantile imputation has been introduced for continuous variables (Bottai and Zhen, 2013)
- Here we will focus on categorical/discrete distributions with missing data

Quantile Imputation for discrete distributions

Let's denote with $Y^{(m)}$ the m-th imputation of a discrete distribution with missing values

1. Draw a value from a $U \sim \mathcal{U}(0,1)$
2. $Y^{(m)} = \hat{Q}_{Y|x}(U)$
3. Repeat Step 1 and 2 to generate M completed datasets

where $\hat{Q}_{Y|x}(U)$ is the conditional quantile function, inverse of the conditional cumulative distribution function, $\hat{F}^{-1}_{Y|x}(y)$, combining studies with data on Y based on a set of predictors x.

Estimate the conditional cumulative distribution function

An estimate of the conditional cumulative distribution function $\hat{F}_{Y|x}(y)$ for a categorial/discrete random variable with k possible outcomes and probability mass function $p(Y|x) = P(Y = y|x)$ with a sample space $y = y_1, y_2, ..., y_k$ can be obtained by

1. estimating a multinomial logistic regression model within each study with complete or partial data
2. combining the parameters across studies with a multivariate meta-regression model
3. predicting conditional cumulative probabilities $\hat{F}_{Y|x}(y_j) = \sum_{i \leq j} \hat{p}(y_i|x)$ in studies with systematic missing data

Example: Cigarette smoking status and pack-years

Consider a composite variable Y about smoking status (Never, Ex, Current) and pack-years (< 20, 20-39, ≥ 40) with $k = 7$ levels and the following probability mass function:

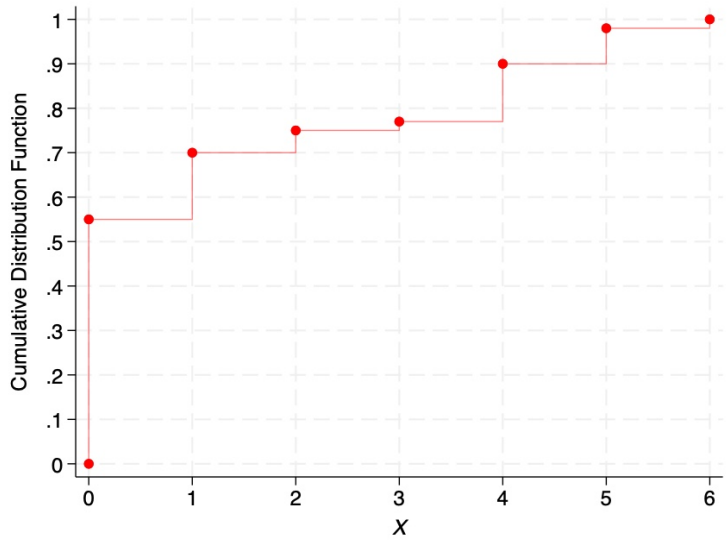
$$p(Y) = \begin{cases} 0.55, & \text{if } y = 0 \text{ (Never)} \\ 0.15, & \text{if } y = 1 \text{ (Ex } < 20 \text{ pack-years)} \\ 0.05, & \text{if } y = 2 \text{ (Ex 20-39 pack-years)} \\ 0.02, & \text{if } y = 3 \text{ (Ex } \geq 40 \text{ pack-years)} \\ 0.13, & \text{if } y = 4 \text{ (Cur } < 20 \text{ pack-years)} \\ 0.08, & \text{if } y = 5 \text{ (Cur 20-39 pack-years)} \\ 0.02, & \text{if } y = 6 \text{ (Cur } \geq 40 \text{ pack-years)} \end{cases}$$

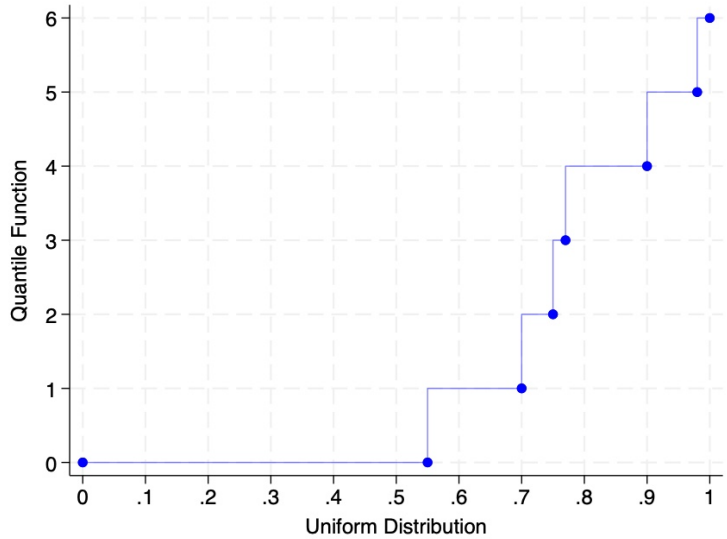
Example: Cigarette smoking status and pack-years

The cumulative distribution function is

$$
\begin{aligned}
P(Y \leq 0) &= 0.55 \\
P(Y \leq 1) &= 0.55 + 0.15 = 0.70 \\
P(Y \leq 2) &= 0.55 + 0.15 + 0.05 = 0.75 \\
P(Y \leq 3) &= 0.55 + 0.15 + 0.05 + 0.02 = 0.77 \\
P(Y \leq 4) &= 0.55 + 0.15 + 0.05 + 0.02 + 0.13 = 0.90 \\
P(Y \leq 5) &= 0.55 + 0.15 + 0.05 + 0.02 + 0.13 + 0.08 = 0.98 \\
P(Y \leq 6) &= 0.55 + 0.15 + 0.05 + 0.02 + 0.13 + 0.08 + 0.02 = 1
\end{aligned}
$$

Cumulative Distribution Function and Quantile Function





Sample size and fraction of missing data vary from 10% to 90% across 5 studies

packyc	study 1	2	3	4	5	Total
0	277	127	1,480	647	4,453	6,984
	27.70	6.35	49.33	16.18	44.53	34.92
1	74	38	400	193	1,200	1,905
	7.40	1.90	13.33	4.83	12.00	9.53
2	21	4	130	50	394	599
	2.10	0.20	4.33	1.25	3.94	3.00
3	6	2	54	21	158	241
	0.60	0.10	1.80	0.53	1.58	1.21
4	73	20	333	161	1,021	1,608
	7.30	1.00	11.10	4.03	10.21	8.04
5	42	17	240	83	649	1,031
	4.20	0.85	8.00	2.08	6.49	5.16
6	8	6	64	23	158	259
	0.80	0.30	2.13	0.57	1.58	1.29
.	499	1,786	299	2,822	1,967	7,373
	49.90	89.30	9.97	70.55	19.67	36.86
Total	1,000	2,000	3,000	4,000	10,000	20,000
	100.00	100.00	100.00	100.00	100.00	100.00

Example: mi impute cqi

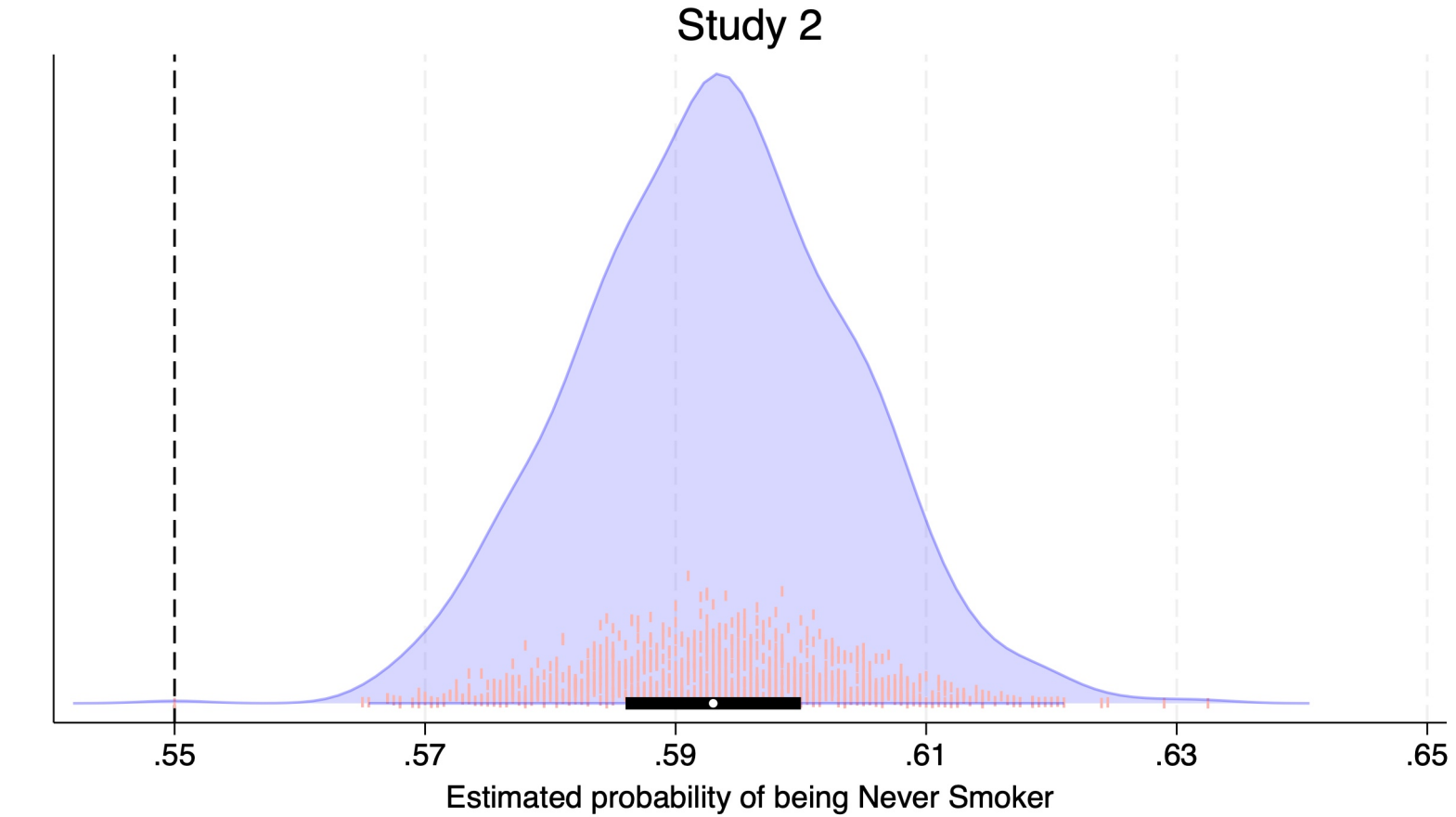
```
use data_packyc.dta

mi set wide

mi register imputed packyc

mi impute cqi packyc , add(1) id(study)
```

Study-specific comparison of observed vs imputed values

```
. tab packyc if study == 2

     packyc |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        127       59.35       59.35
          1 |         38       17.76       77.10
          2 |          4        1.87       78.97
          3 |          2        0.93       79.91
          4 |         20        9.35       89.25
          5 |         17        7.94       97.20
          6 |          6        2.80      100.00
------------+-----------------------------------
      Total |        214      100.00

. tab _1_packyc if study == 2

  _1_packyc |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      1,206       60.30       60.30
          1 |        337       16.85       77.15
          2 |         43        2.15       79.30
          3 |         17        0.85       80.15
          4 |        180        9.00       89.15
          5 |        161        8.05       97.20
          6 |         56        2.80      100.00
------------+-----------------------------------
      Total |      2,000      100.00
```

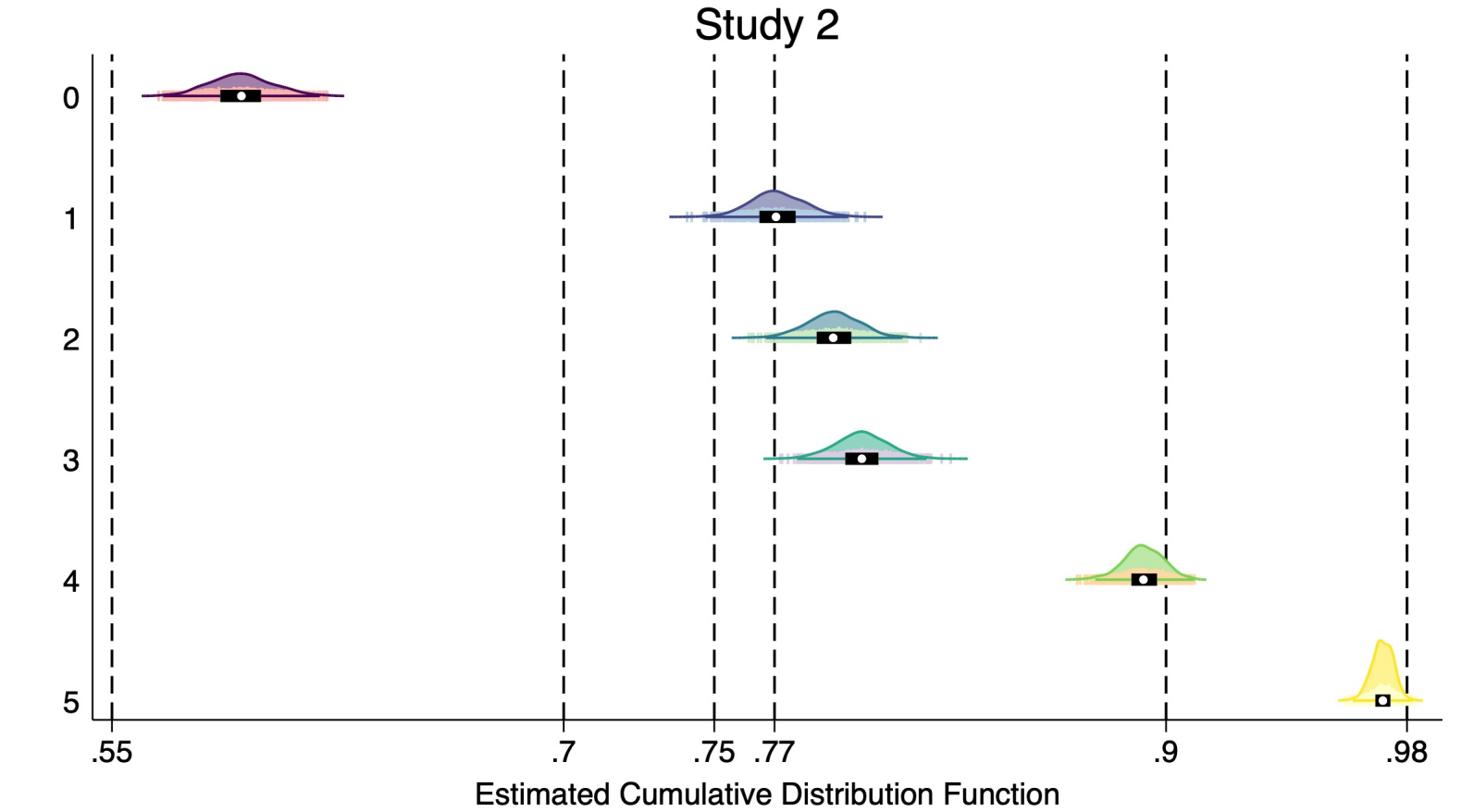
Imputed and observed probabilities are not close to the parameter value

In Study 2, sample of n=2000 and 90% missing data, the imputed probabilities of being never smoker are centered about the observed probability (0.59). However, such probability should be 0.55.



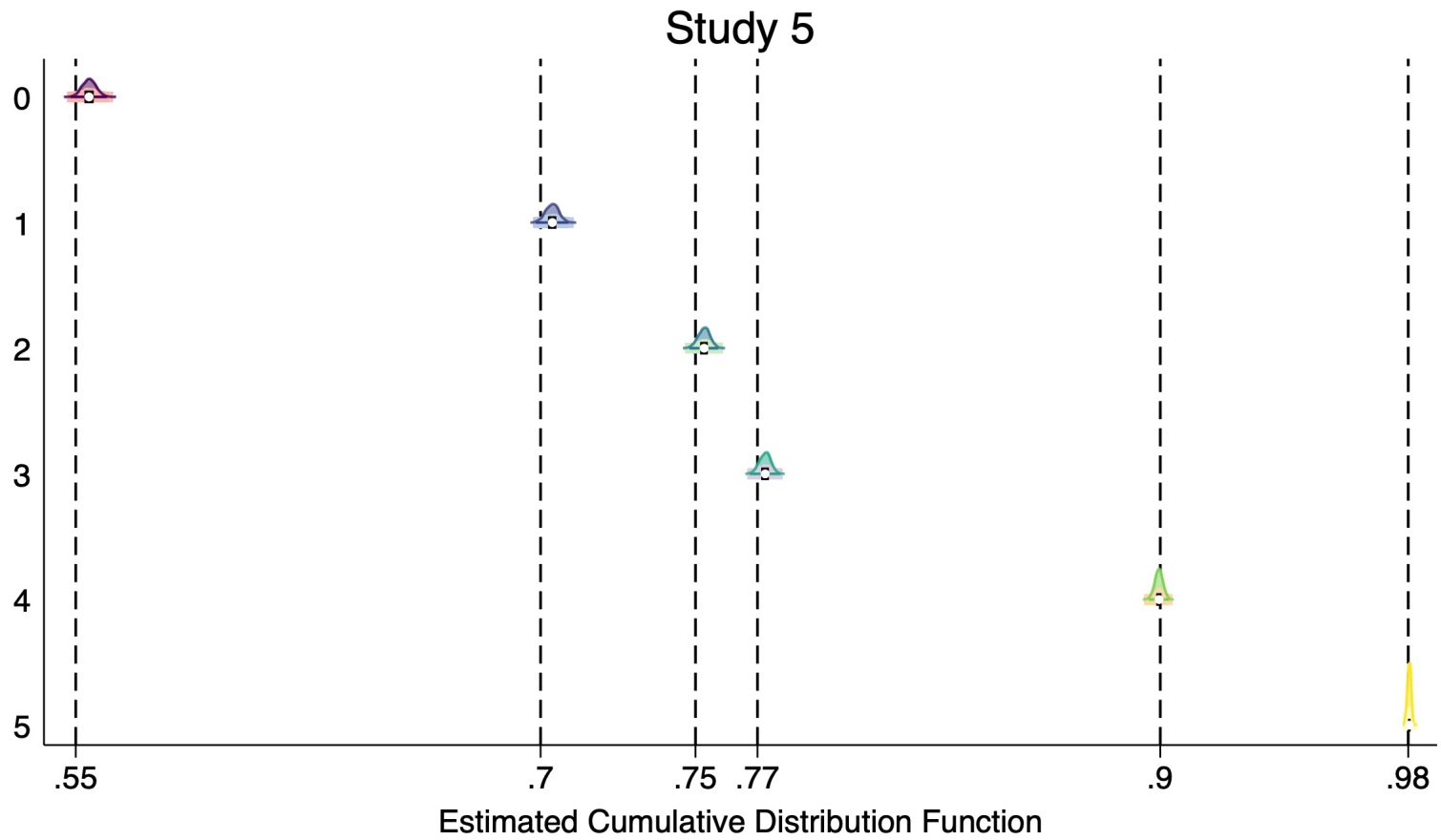
Imputed CDF is centered about the observed CDF but not the population CDF

In Study 2, the imputed cumulative probabilities of smoking pack-years are, overall, not centered about their corresponding parameter values (0.55 , 0.70 , 0.75 , 0.77 , 0.90 , 0.98).



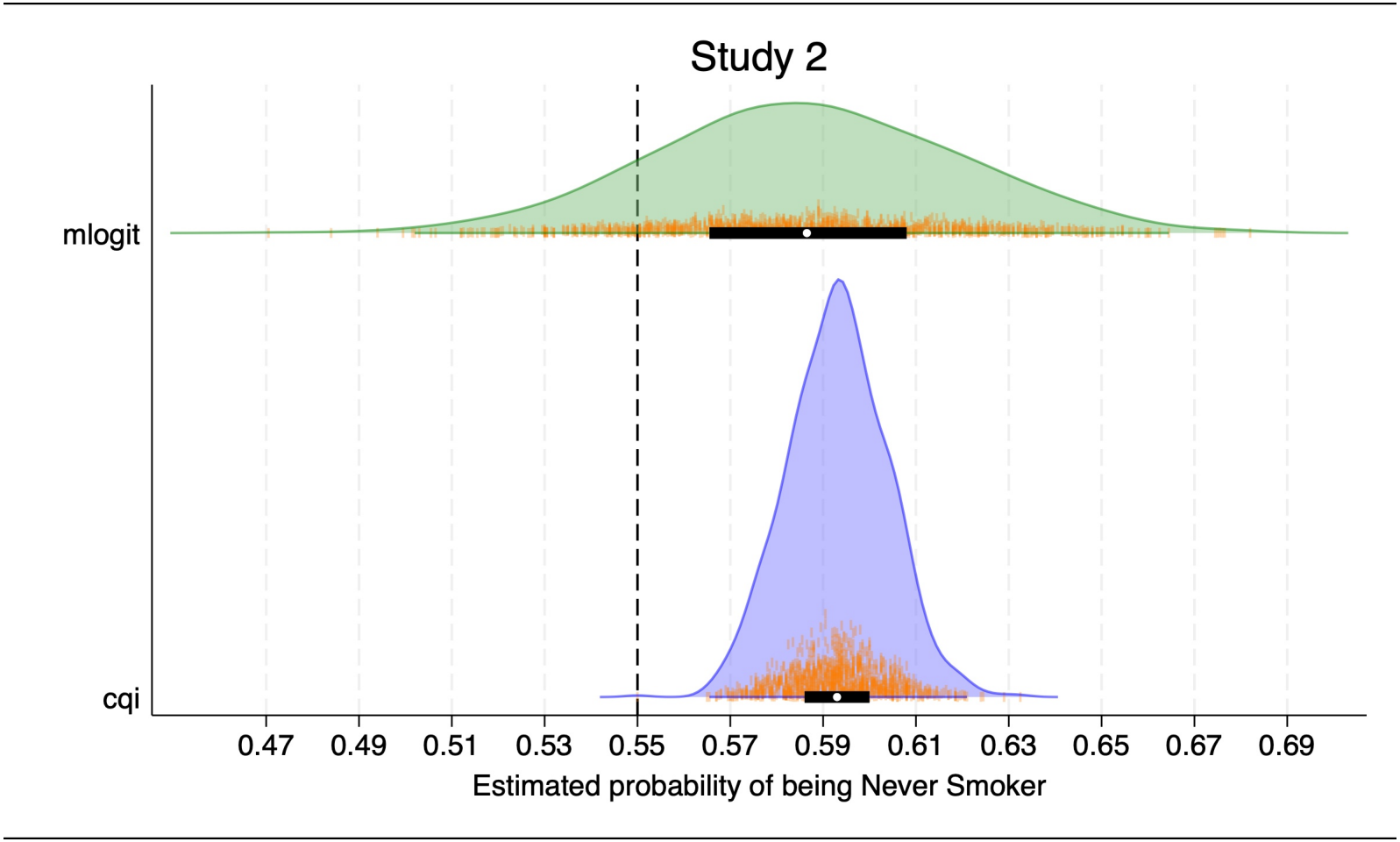
Study 5: 20% missing data

In Study 5, sample of n=10,000 and 20% missing data, the central tendencies of the imputed cumulative probabilities of smoking pack-years are, overall, close to the corresponding parameter values (0.55 , 0.70 , 0.75 , 0.77 , 0.90 , 0.98).



Difference mi impute mlogit vs. mi impute cqi

```
mi impute mlogit packyc , add(100) by(study)
mi impute    cqi packyc , add(100) id(study)
```



Sample size and fraction of missing data vary from 10% to 100% across 5 studies

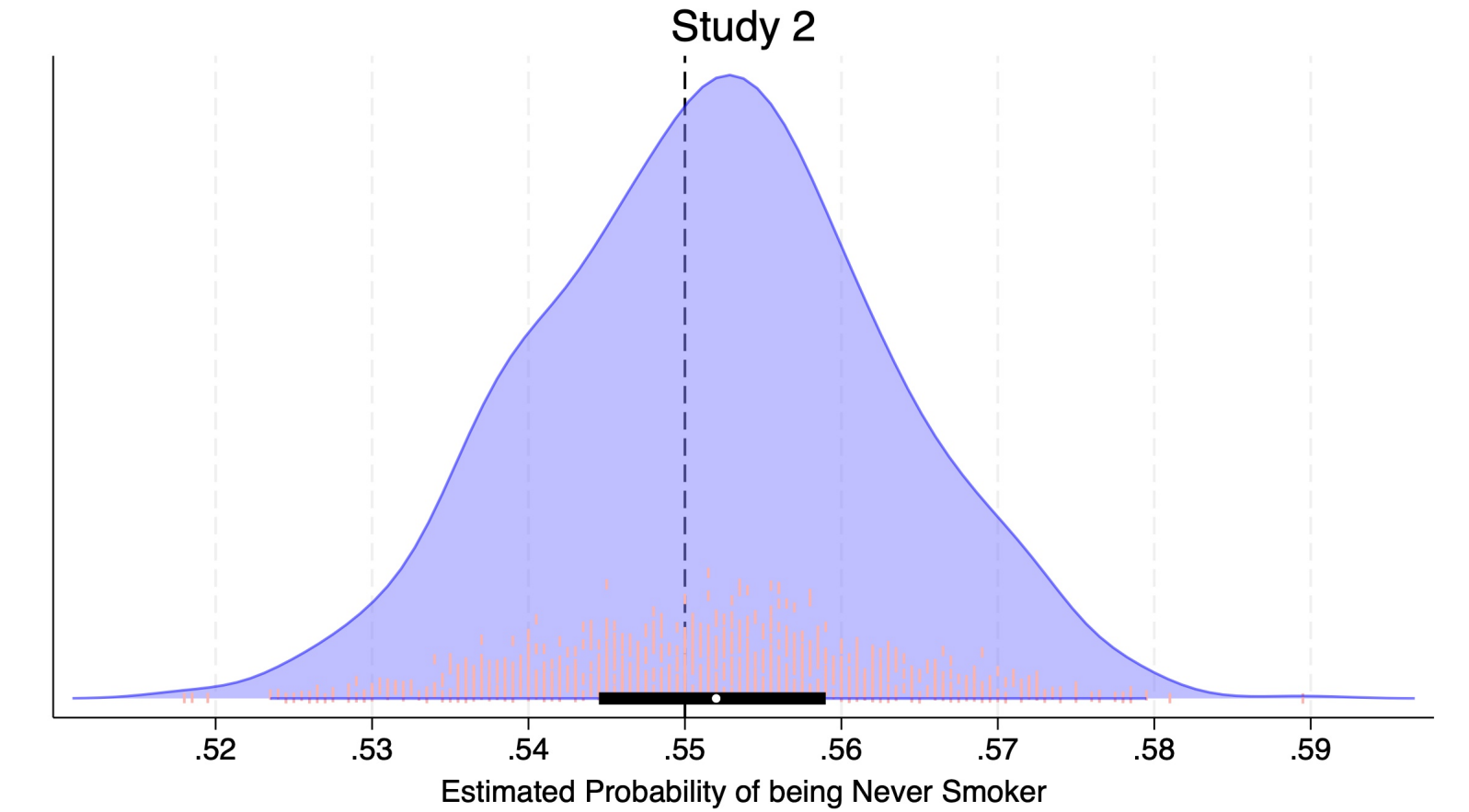
packyc	study 1	2	3	4	5	Total
0	277	0	1,480	647	4,453	6,857
	27.70	0.00	49.33	16.18	44.53	34.28
1	74	0	400	193	1,200	1,867
	7.40	0.00	13.33	4.83	12.00	9.34
2	21	0	130	50	394	595
	2.10	0.00	4.33	1.25	3.94	2.97
3	6	0	54	21	158	239
	0.60	0.00	1.80	0.53	1.58	1.20
4	73	0	333	161	1,021	1,588
	7.30	0.00	11.10	4.03	10.21	7.94
5	42	0	240	83	649	1,014
	4.20	0.00	8.00	2.08	6.49	5.07
6	8	0	64	23	158	253
	0.80	0.00	2.13	0.57	1.58	1.26
.	499	2,000	299	2,822	1,967	7,587
	49.90	100.00	9.97	70.55	19.67	37.94
Total	1,000	2,000	3,000	4,000	10,000	20,000
	100.00	100.00	100.00	100.00	100.00	100.00

No comparison can be made for a study with systematic missing

```
.  tab packyc if study == 2
no observations

.  tab _1_packyc if study == 2

  _1_packyc |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      1,120       56.00       56.00
          1 |        296       14.80       70.80
          2 |         85        4.25       75.05
          3 |         39        1.95       77.00
          4 |        258       12.90       89.90
          5 |        156        7.80       97.70
          6 |         46        2.30      100.00
------------+-----------------------------------
      Total |      2,000      100.00
```

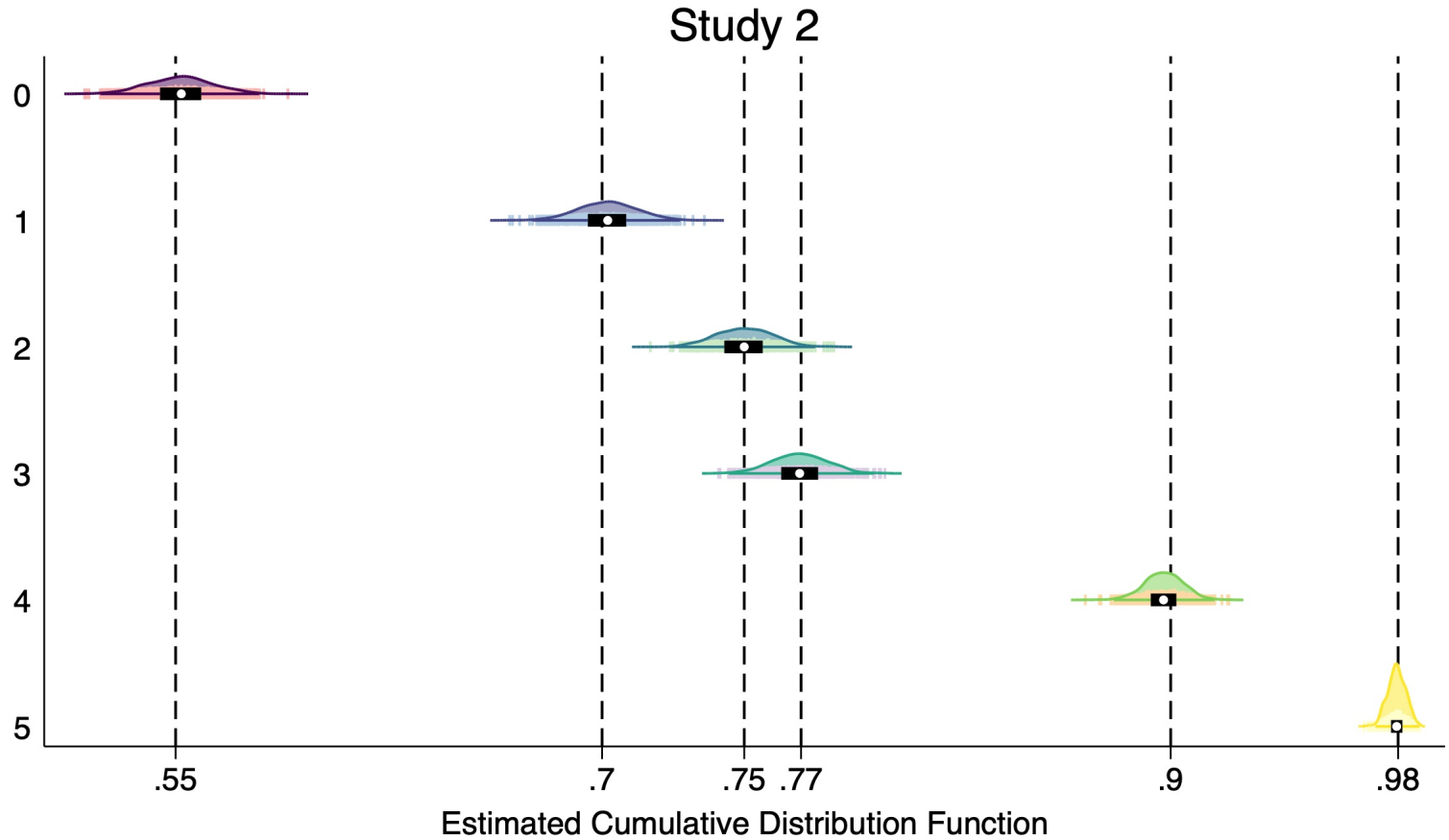
Study 2: 100% missing data

In Study 2, with 100% missing data, the imputed probability (average of 0.552) of being never smoker is very close to the corresponding parameter value (0.550).



Study 2: 100% missing data

The imputed cumulative probabilities of smoking pack-years are, overall, centered about their corresponding parameter values (0.55 , 0.70 , 0.75 , 0.77 , 0.90 , 0.98). The key was to learn the parameters of the CDF underlying smoking pack-years from other studies.



Smoking distribution depends on sex

Consider a cumulative distribution function of smoking that vary with sex. Among men ($x = 1$), the percent of "Never smoker", is much smaller than women ($x = 0$).

Levels	y	$F_{Y\|x=0}(y)$	$F_{Y\|x=1}(y)$
Never	0	0.55	0.35
Ex $<$ 20	1	0.70	0.60
Ex 20-39	2	0.75	0.70
Ex $\geq$ 40	3	0.77	0.75
Cur $<$ 20	4	0.90	0.85
Cur 20-39	5	0.98	0.95
Cur $\geq$ 40	6	1.00	1.00

Missing data in smoking depends on sex

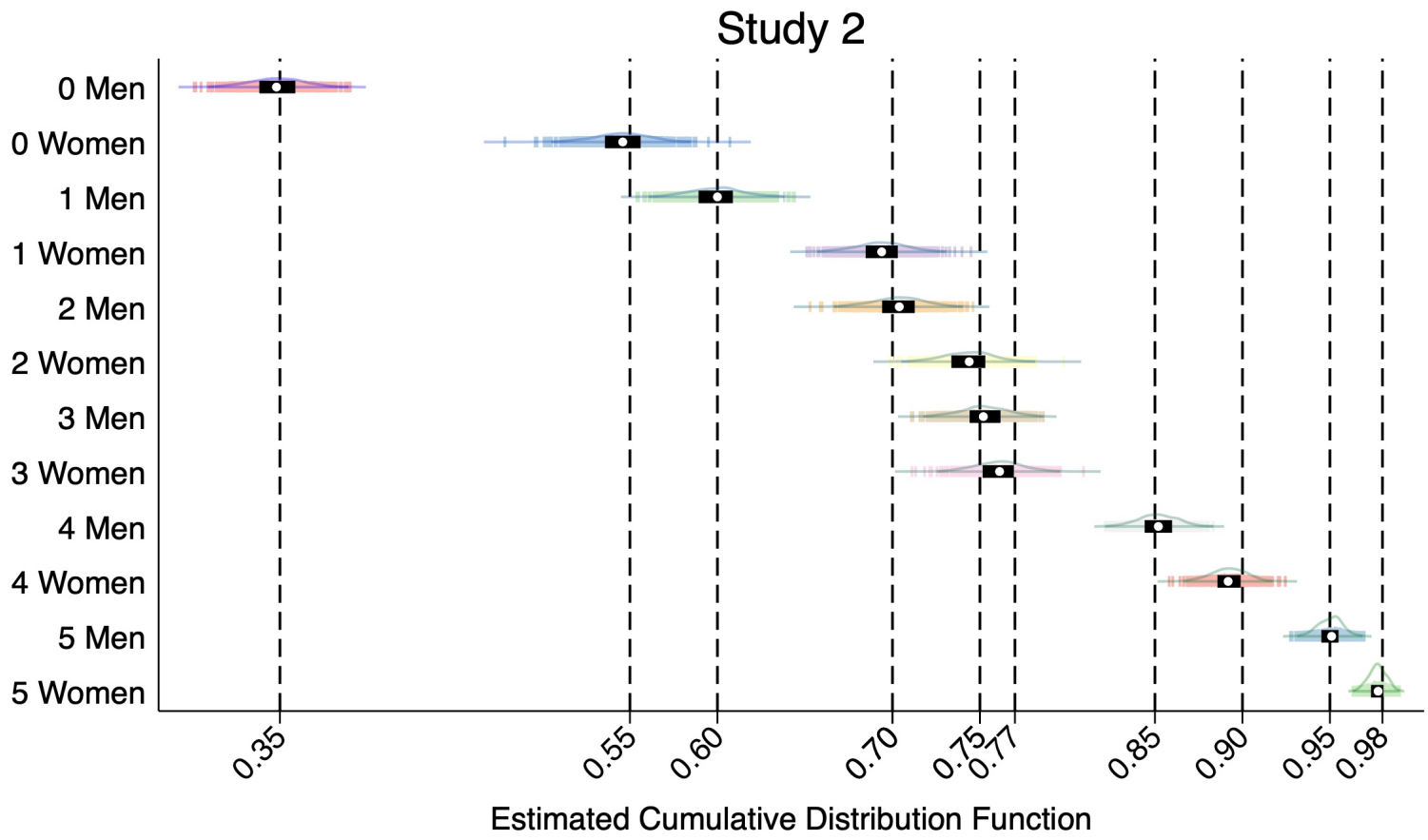
- Men are more prone to smoking than women
- Men are twice as likely as women to have missing data on smoking
- 5 studies
- Sample size $= \{1000, 2000, 3000, 4000, 10000\}$
- Fraction of Missing Data Men $= \{0.10, 1, 0.50, 0.70, 0.90\}$
- Fraction of Missing Data Women $= \{0.05, 1, 0.25, 0.35, 0.45\}$
- Study 2 has data on sex but not smoking

Imputation of smoking conditionally on sex

```
use data_packyc_sex_sm, clear
mi set wide
mi register imputed packyc
mi impute cqi packyc men , add(1) id(study)
```

```
. tabulate _1_packyc men if study == 2, col nofreq
           |          men
 _1_packyc |         0          1 |     Total
-----------+----------------------+----------
         0 |     54.80      34.90 |     44.65
         1 |     13.88      26.96 |     20.55
         2 |      5.31       9.31 |      7.35
         3 |      1.63       4.02 |      2.85
         4 |     13.57       8.92 |     11.20
         5 |      8.57      11.08 |      9.85
         6 |      2.24       4.80 |      3.55
-----------+----------------------+----------
     Total |    100.00     100.00 |    100.00
```

Imputed CDFs in Study 2 with no data



IPD Meta-analysis with a question about effect modification

- The hypothesis is that the treatment effect on mortality rate is varying according to the level of a prognostic factor, called effect modifier.
- The treatment has been randomly allocated in 5 randomized trials
- The effect modifier, however, has been measured at baseline in just 3 out of 5 trials
- The endpoint is the time elapsed from treatment randomization until death (or end of follow-up)
- The statistical power to detect an interaction effect combining the 3 studies with no missing data is low (44%)

Data Generating Mechanism

- 5 studies
- Sample size $= \{1000, 2000, 3000, 10000, 20000\}$
- categorical effect modifier $Z = \{0, 1, 2\}$ with $F(z) = (0.4, 0.8, 1)$
- treatment $X = \{0, 1\}$ with $F(x) = 0.5$
- continuous outcome t is time from baseline to death (years) or 10 years, whichever came first
- the two largest studies (10000 and 20000) have no data on the effect modifier

Outcome statistical model

An interaction effect between treatment and effect modifier in a Weibull survival model with $S(t) = e^{-\lambda t^{\gamma}}$ is determined by the following equation

$$ln(\lambda|x,z) = \beta_0 + \beta_1 x + \beta_2 I(z=1) + \beta_3 I(z=2) + \beta_4 x I(z=1) + \beta_5 x I(z=2)$$

A random draw from a continuous uniform distribution provides a distribution of time to death conditionally on treatment and effect modifier.

$$t = [-\ln(U)/(\lambda|x,z)]^{1/\gamma}$$

Model parameters

The treatment effects for low, medium, and high levels of the effect modifier are

$$HR_{x|z=0} = e^{\beta_1} = 0.8 \quad \text{Beneficial}$$
$$HR_{x|z=1} = e^{\beta_1+\beta_4} = 1.0 \quad \text{Null}$$
$$HR_{x|z=2} = e^{\beta_1+\beta_5} = 1.2 \quad \text{Harmful}$$

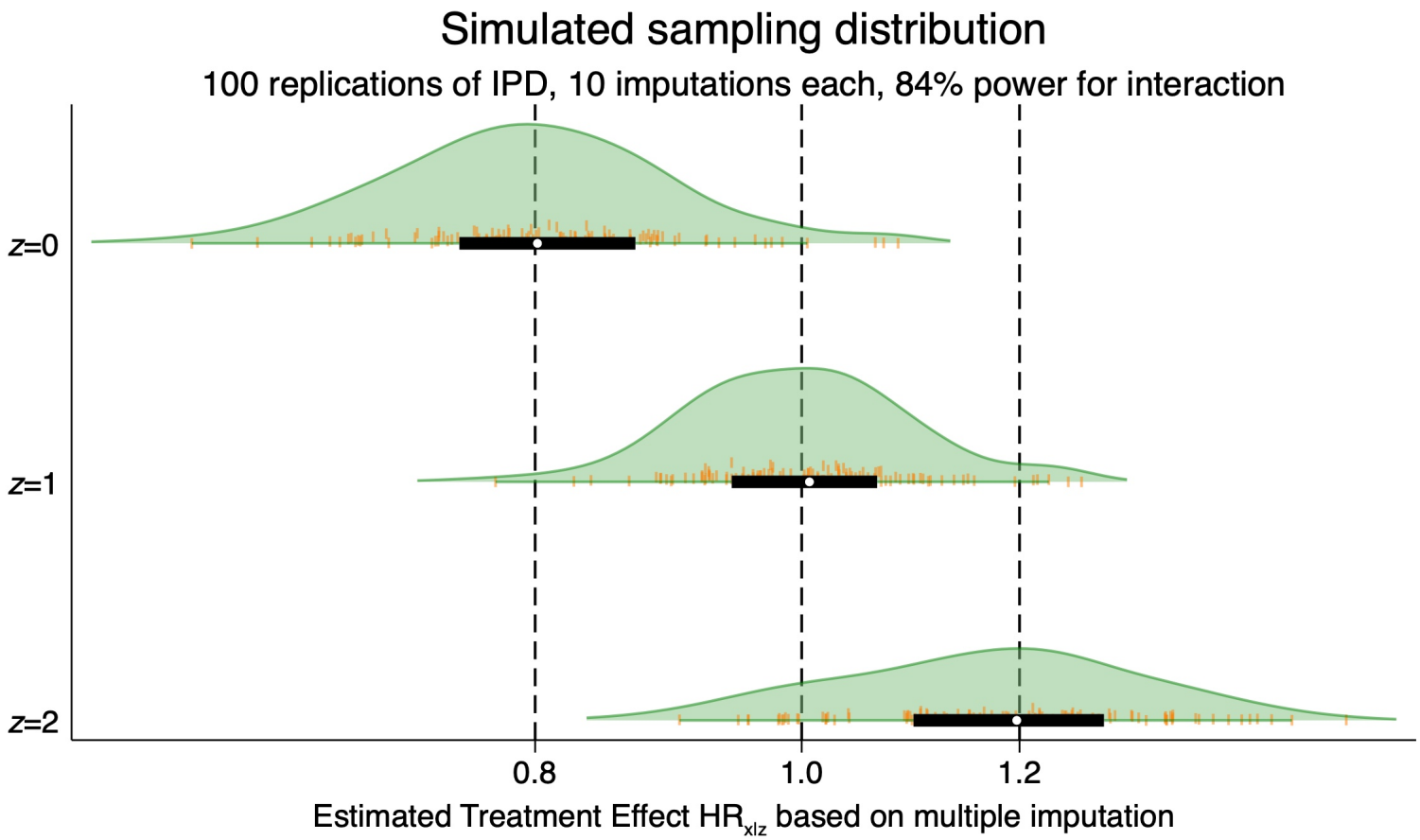
The above parameters are underlying all 5 studies

Code to impute and analyze an IPD meta-analysis

```
use data_nordic23_sm, clear

mi set wide
mi stset time, fail(death)
gen x_t = x*_t
gen x_d = x*_d

mi register regular x _t _d x_t x_d
mi register imputed z
mi register passive zi1 zi2 x_zi1 x_zi2

mi impute cqi z x _t _d x_t x_d , add(10) id(study)

mi passive: replace zi1 = (z==1)
mi passive: replace zi2 = (z==2)
mi passive: replace x_zi1 = x*zi1
mi passive: replace x_zi2 = x*zi2

mi estimate , post saving(miestfile, replace): twostage streg x zi1 zi2 x_zi1 x_zi2 , dist(weibull) id(study)

mi test x_zi1 x_zi2

mi predictnl  est_bxz0 = _b[x]             using miestfile, se(est_se_bxz0)
mi predictnl  est_bxz1 = _b[x] + _b[x_zi1] using miestfile, se(est_se_bxz1)
mi predictnl  est_bxz2 = _b[x] + _b[x_zi2] using miestfile, se(est_se_bxz2)
```

Simulated sampling distribution of estimates based on `twostage` + `cqi` + congenial imputation model

$$P(z|x,t,d) = x + t + d + xt + xd$$



Ignoring interaction in the imputation model

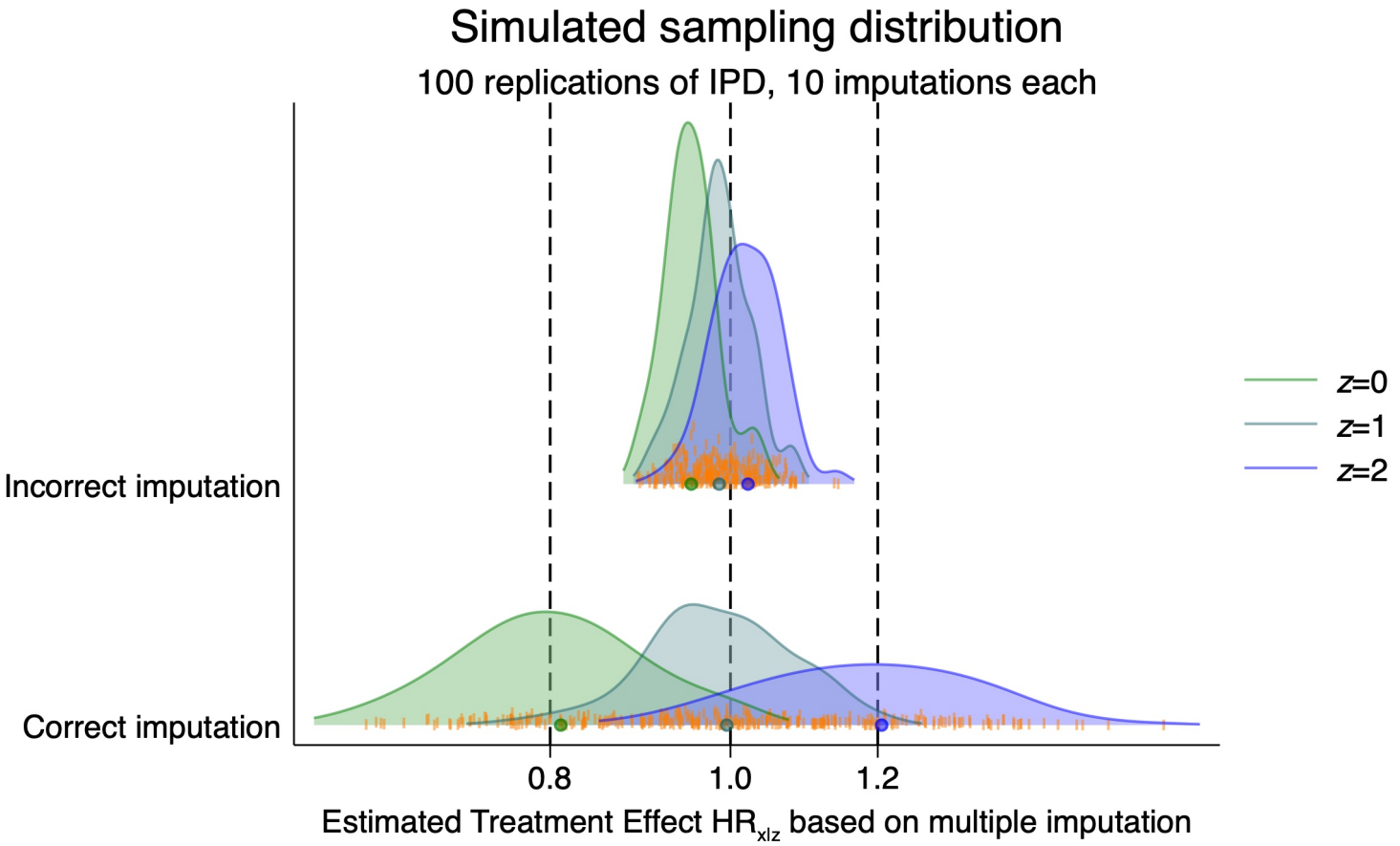
It is quite easy to obtain bias estimates and lose statistical power by apparently minor changes in the specification of the imputation model.

For example, below we (incorrectly) omit the interaction term between the treatment and the two variables defining the survival data.

```
mi impute cqi z x _t _d  , add(10) id(study) // x_t x_d
```

Simulated sampling distribution of estimates based on `twostage` + `cqi` + not congenial imputation model

$$P(z|x,t,d) = x + t + d$$



Final remarks

- `mi impute cqi` is based on the principle of inverting the conditional CDF with no additional randomness
- `mi impute cqi` can be used with in IPD meta-analysis with both sporadic and systematic missing data
- options of `mi impute` can be used but, unfortunately, it cannot be called by `mi impute chained`
- `mi impute cqi` can be easily extended to different type of distributions
- similarly to any other imputation approach, using `mi impute cqi` with the wrong imputation model would lead to incorrect statistical inference
- This is an on-going joint work with Robert Thiesmeier

Selected references

Riley RD, Tierney J, Stewart LA (Eds). *Individual Participant Data Meta-Analysis: A Handbook for Healthcare Research*. Chichester: Wiley, 2021.

Bottai, M., Zhen, H. (2013). Multiple imputation based on conditional quantile estimation. Epidemiology, Biostatistics and Public Health, 10(1).

Jolani S, Debray TPA, Koffijberg H, va Buuren S, Moons KGM. *Multiple imputation of systematically missing predictors in an individual participant data meta-analysis: a generalised approach using MICE.* Stat Med 2015; 34(11):1841-1863.

Resche-Rigon M, White IR, Bartlett JW, Peters SA, Thompson SG. *Multiple imputation for handling systematically missing confounders in meta-analysis of individual participant data.* Stat Med 2013; 32(28):4890-4905.