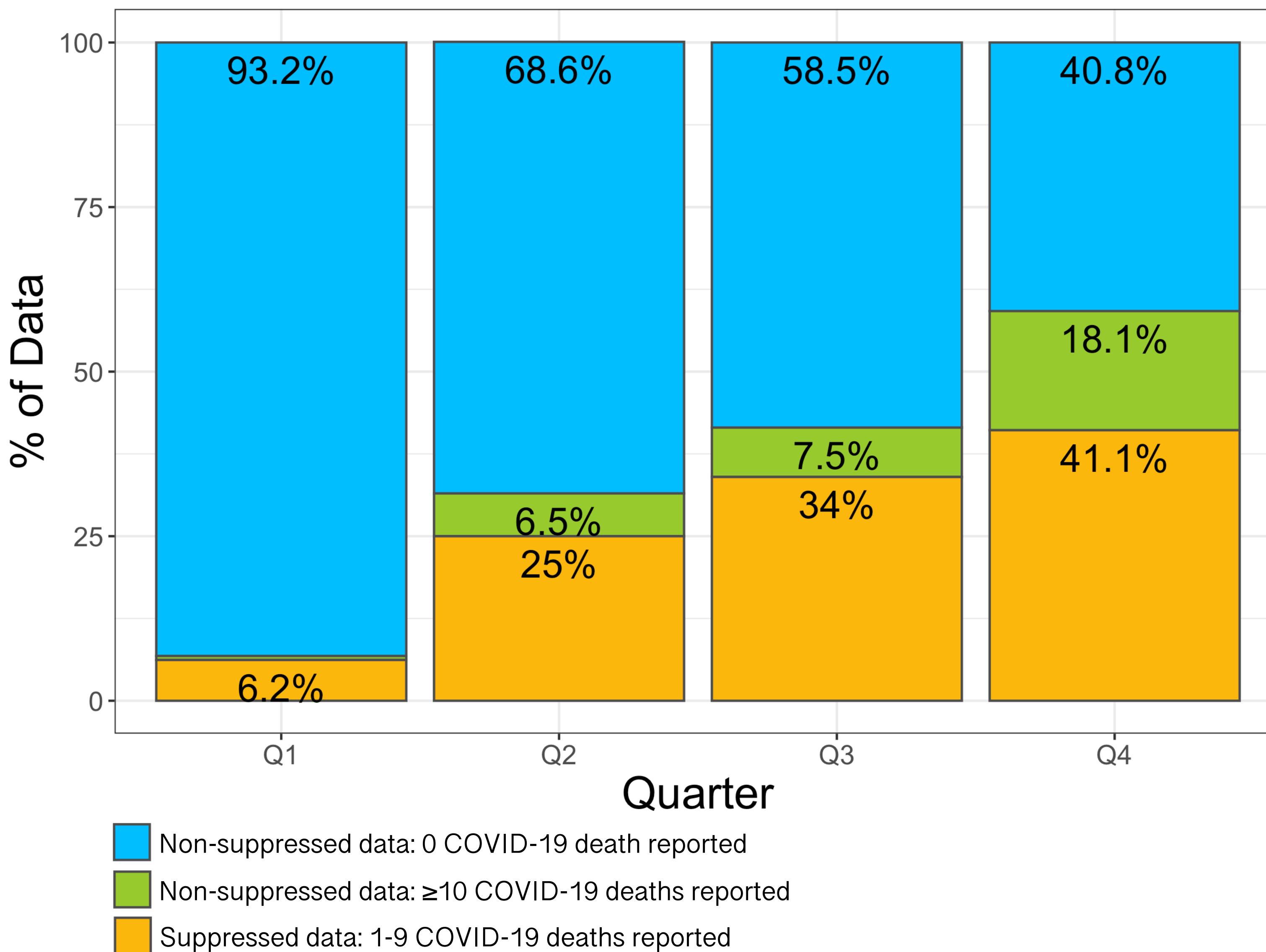


Imputation for Provisional COVID-19 Deaths in 2020

Why imputation?

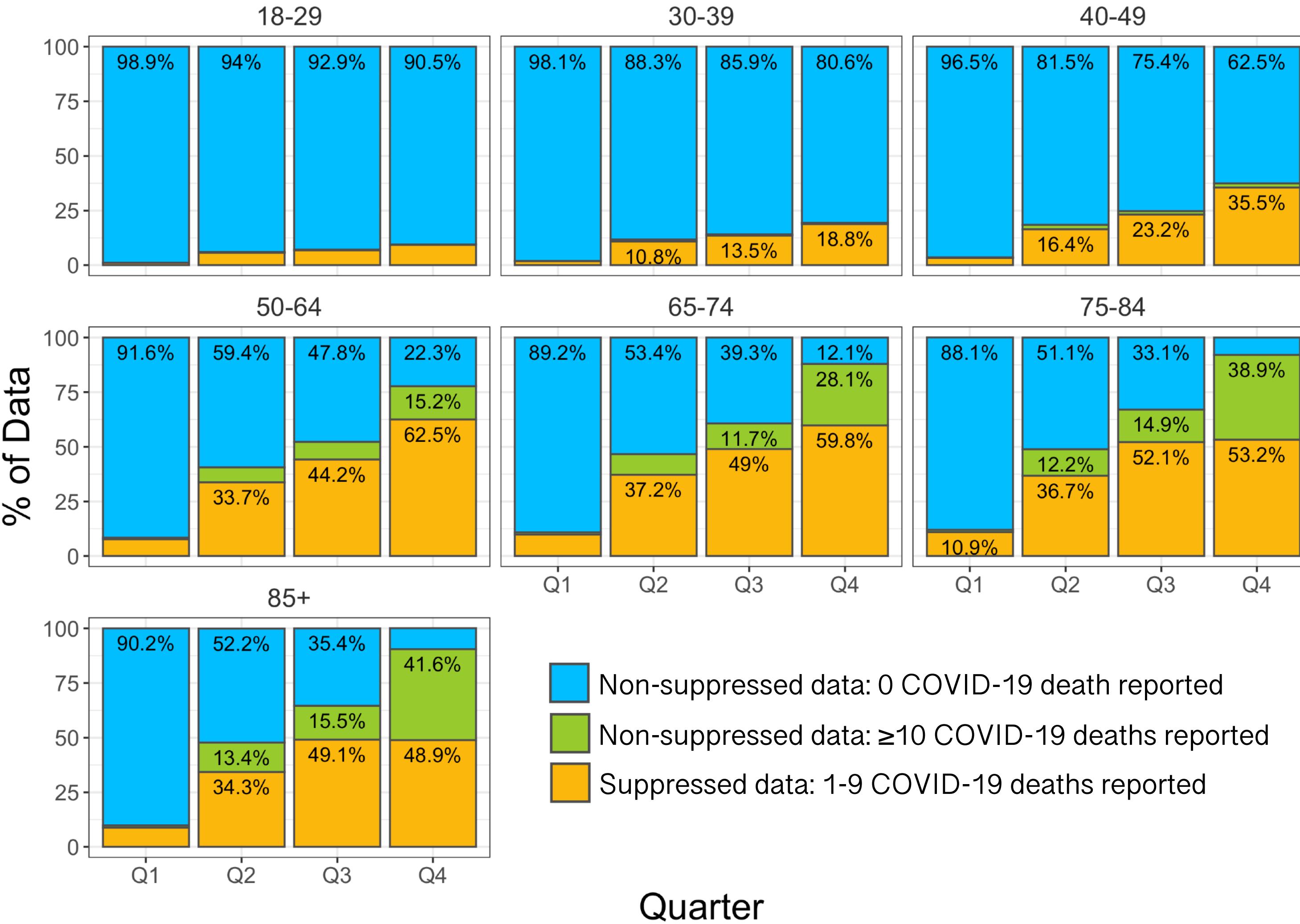
Distribution of suppressed and non-suppressed COVID-19 deaths



- County-level COVID-19 death counts from 1 to 9 were suppressed
- The percent of data with suppressed COVID-19 death counts increased from Q1 to Q4
- Could bias the estimates of the years of potential life loss (YPLL) due to COVID-19

Data type by age group

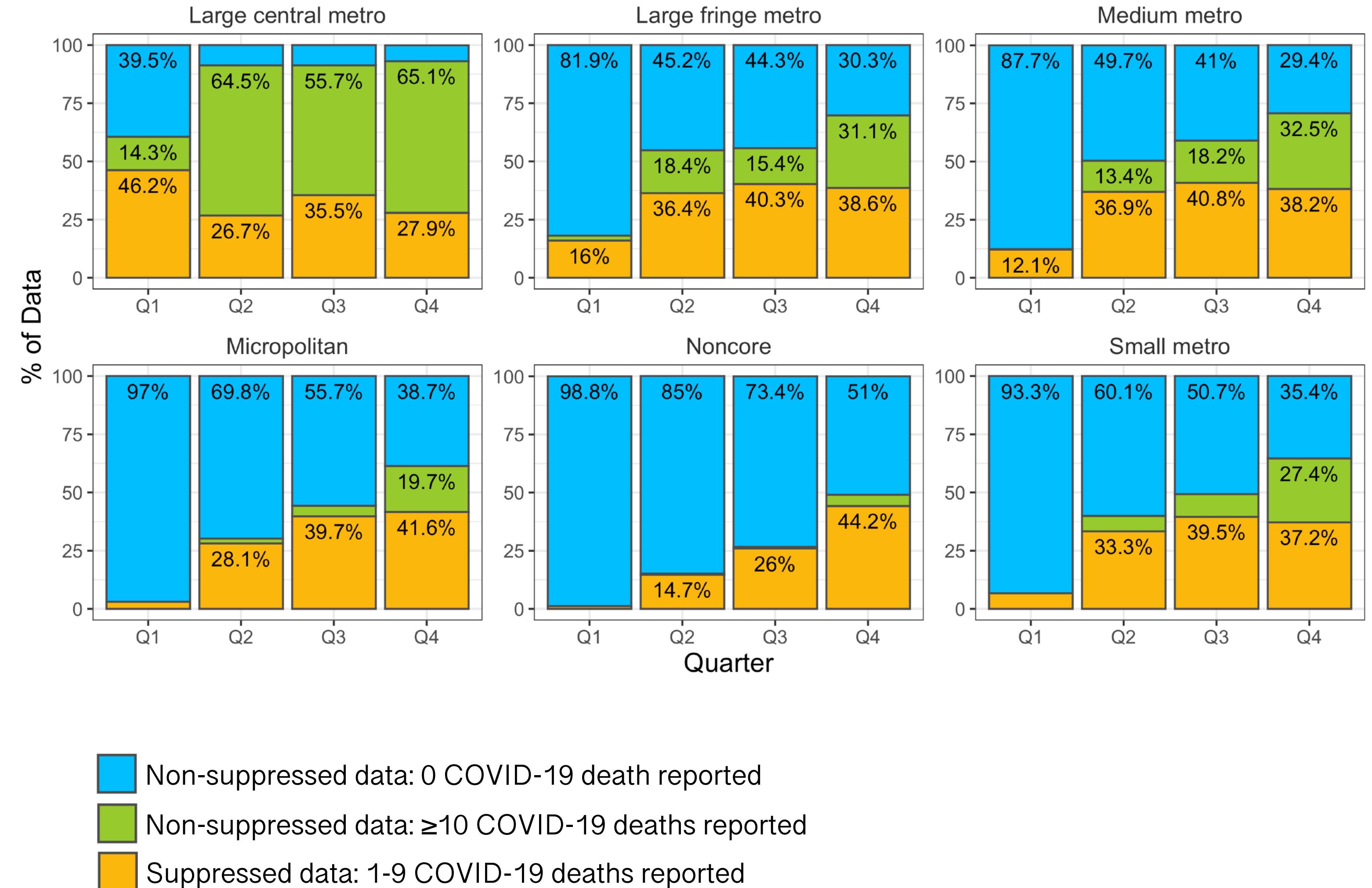
Age group: distribution of suppressed and non-suppressed COVID-19 deaths



- More suppressed data among older age groups
- More zeros among younger age groups (18-29, 30-39, 40-49)

Data type by county type (urban/rural)

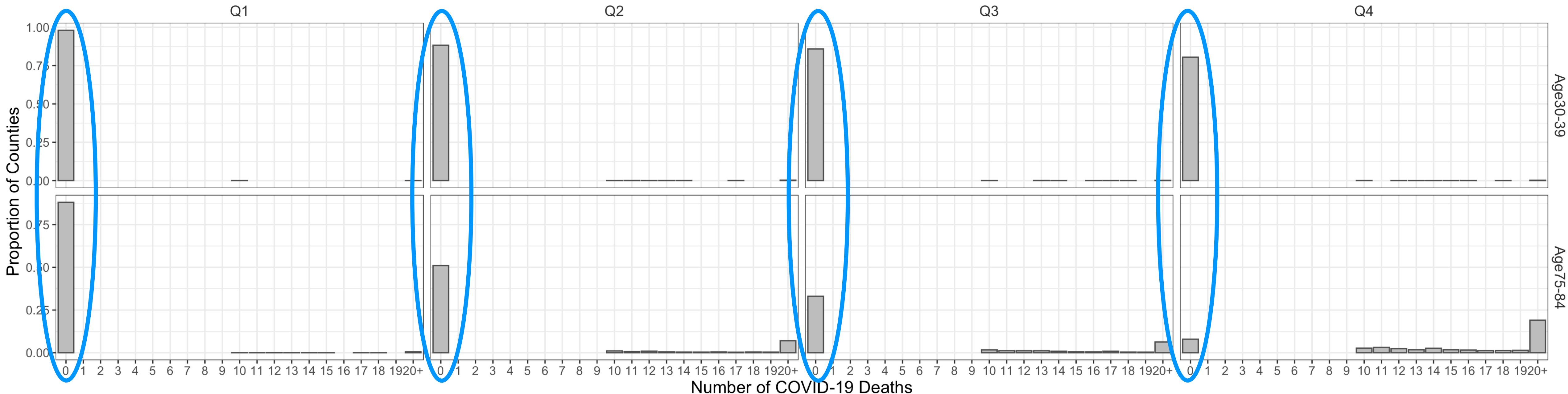
Urban-rural code: distribution of suppressed and non-suppressed COVID-19 deaths



- The pattern of counties with zeros and suppressed data varied with the types of counties
- Large central metro had a very different pattern than the other types of counties
- Large fringe, medium, and small metro had a similar pattern of zeros and suppressed data

Distribution of COVID-19 deaths

Distribution of county-level COVID-19 deaths (selected age groups)

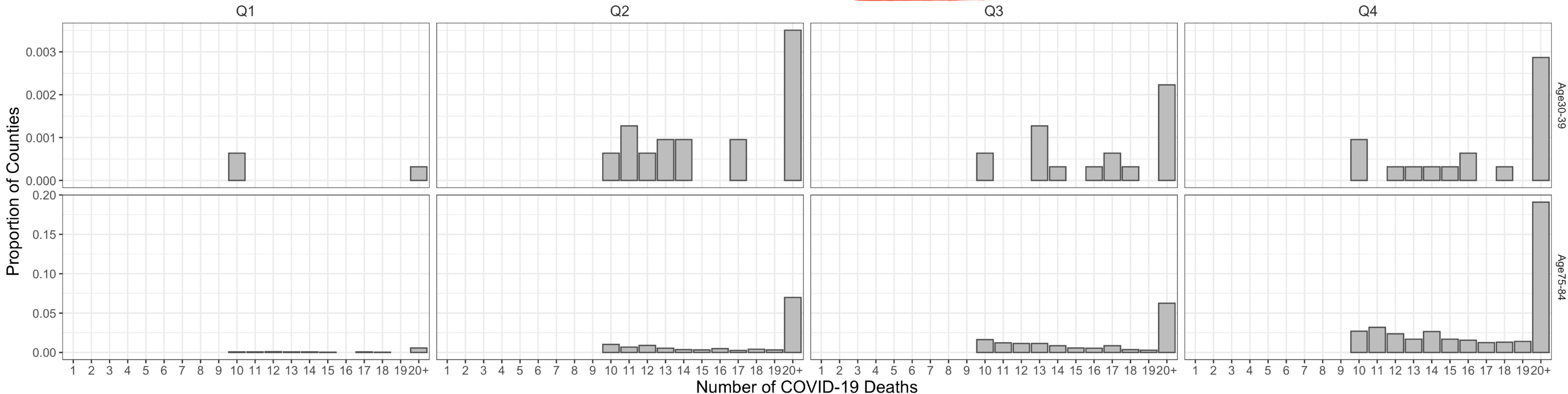


Note: Number of COVID-19 deaths was top-coded at 20. All death counts ≥ 20 were collapsed into “20+”

- A large mass of zero COVID-19 deaths
 - In early pandemic (e.g., Q1 and Q2)
 - Among younger age groups

Distribution of COVID-19 deaths

Distribution of county-level COVID-19 deaths (positive only; selected age groups)

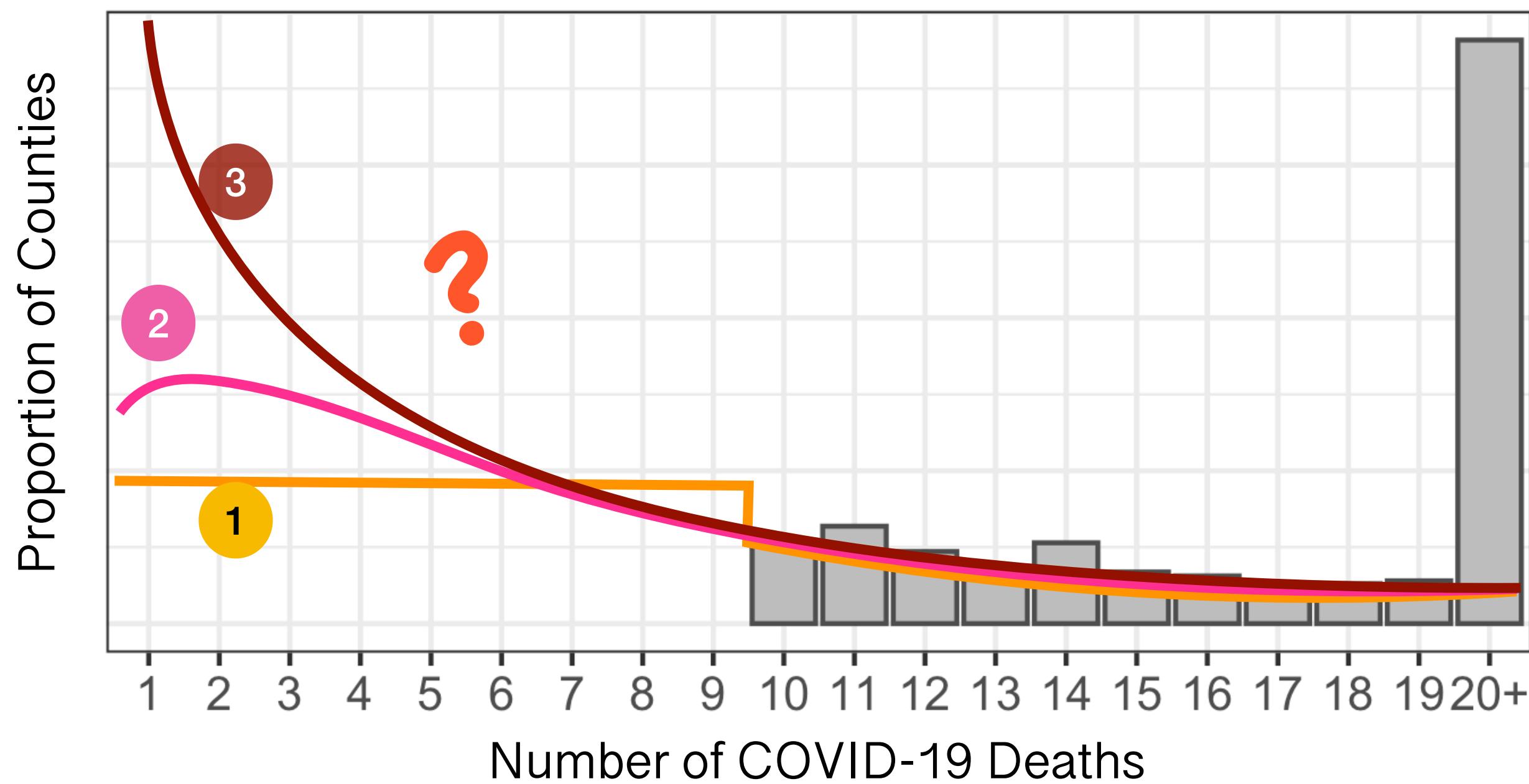


Note: Number of COVID-19 deaths was top-coded at 20. All death counts ≥ 20 were collapsed into “20+”

- A large mass of zero COVID-19 deaths among counties
- Positive death counts follow a right-skewed distribution
 - The largest value is 1,505 COVID-19 deaths among the population aged 75-84 in Queens County, NY in quarter 2

Distribution of COVID-19 deaths

Distribution of county-level COVID-19 deaths
(positive only): Age 75-84; Quarter 4

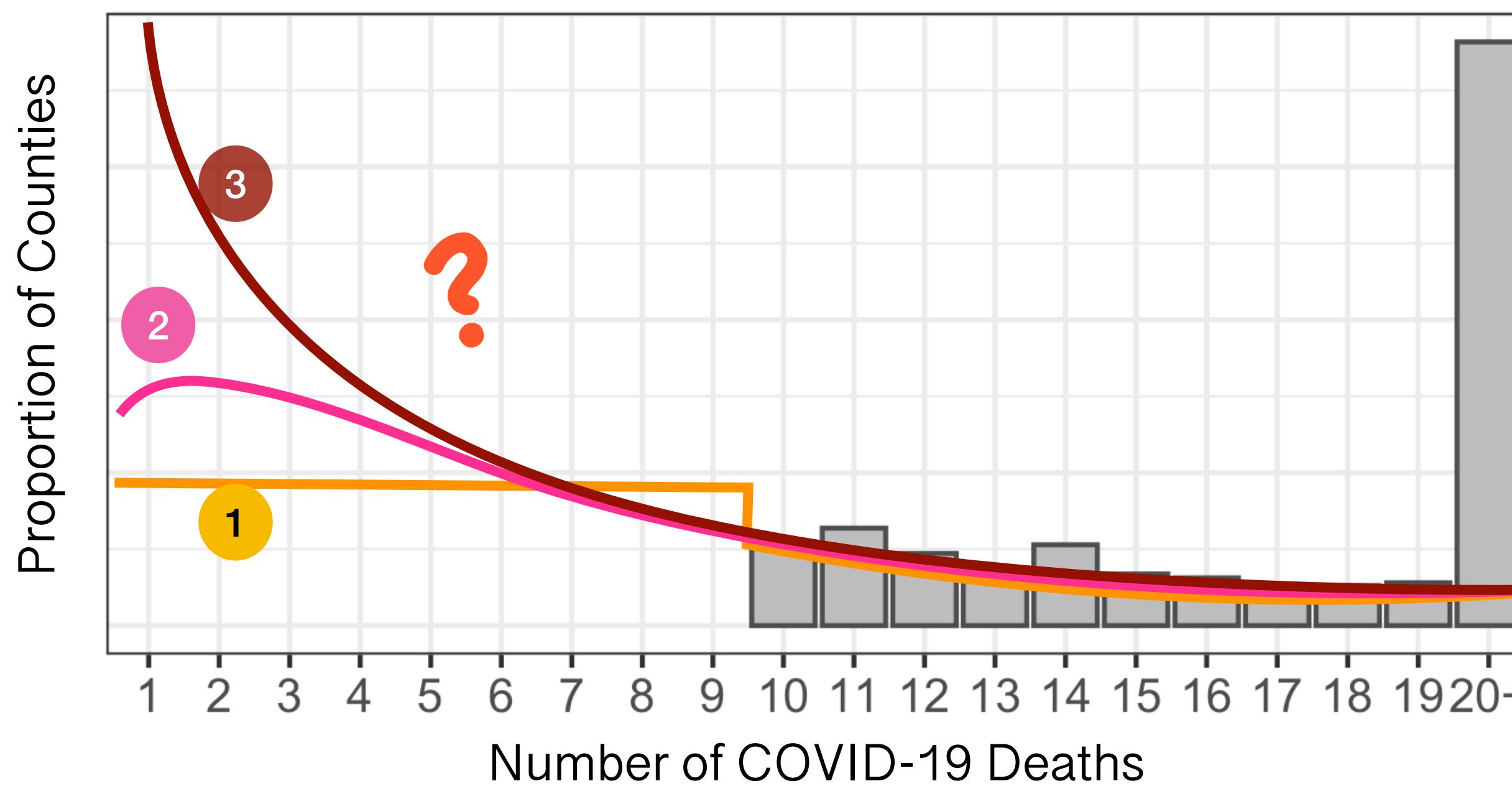


Note: Number of COVID-19 deaths was top-coded at 20.

All death counts ≥ 20 were collapsed into “20+”

Distribution of COVID-19 deaths

Distribution of county-level COVID-19 deaths
(positive only): Age 75-84; Quarter 4



Note: Number of COVID-19 deaths was top-coded at 20.

All death counts ≥ 20 were collapsed into "20+"

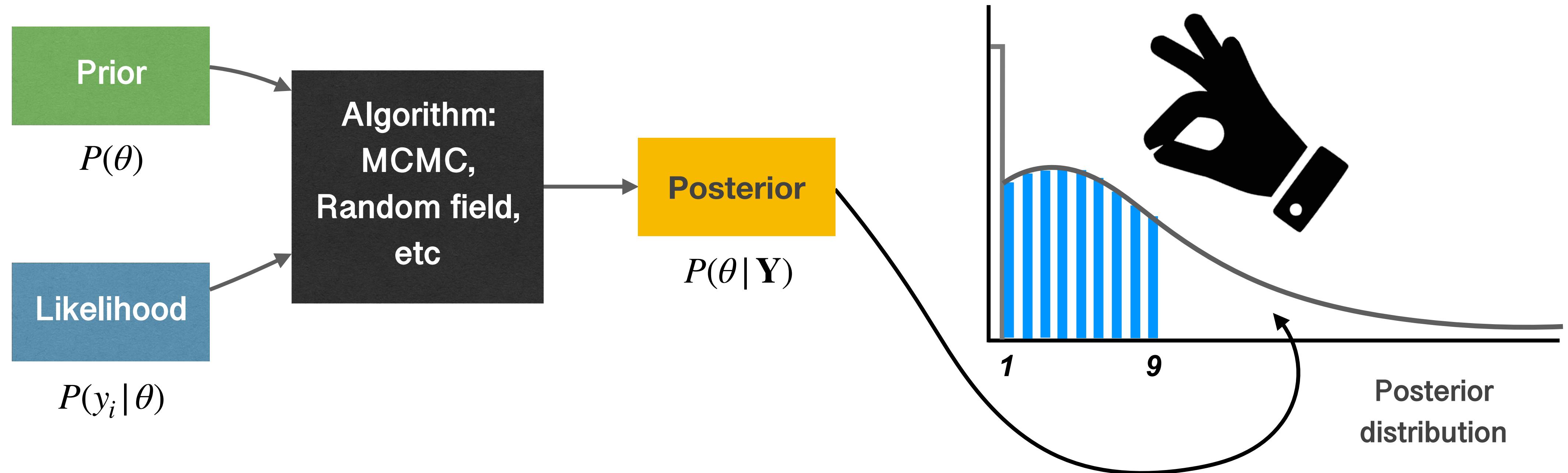
The suppressed COVID-19 death counts have to follow some hypothetical distribution

- Have a large mass of zero
- Right-skewed distribution
- Other determinants
 - Temporal trend
 - Age group
 - Type of county (urban/rural)
 - Population size
 - State trend

Methods

- Population:
 - 3,138 counties (removed 2 counties due to no population estimates)
- Data:
 - AH Provisional COVID-19 Deaths by Quarter, County and Age for 2020
 - American Community Survey 5-Year Estimates
- Methods: Bayesian multilevel gamma hurdle model
 1. Estimate the probability of having 0 deaths and positive deaths separately
 2. Assume a right-skewed distribution among the positive deaths
 3. 3 levels: time trend (quarters), county level, and state level
 - ▶ DC and NYC are accounted as state level in the analysis
 4. Factors included: age groups, type of county, population size
 5. Aggregate annual COVID-19 deaths by age group at state level is part of likelihood

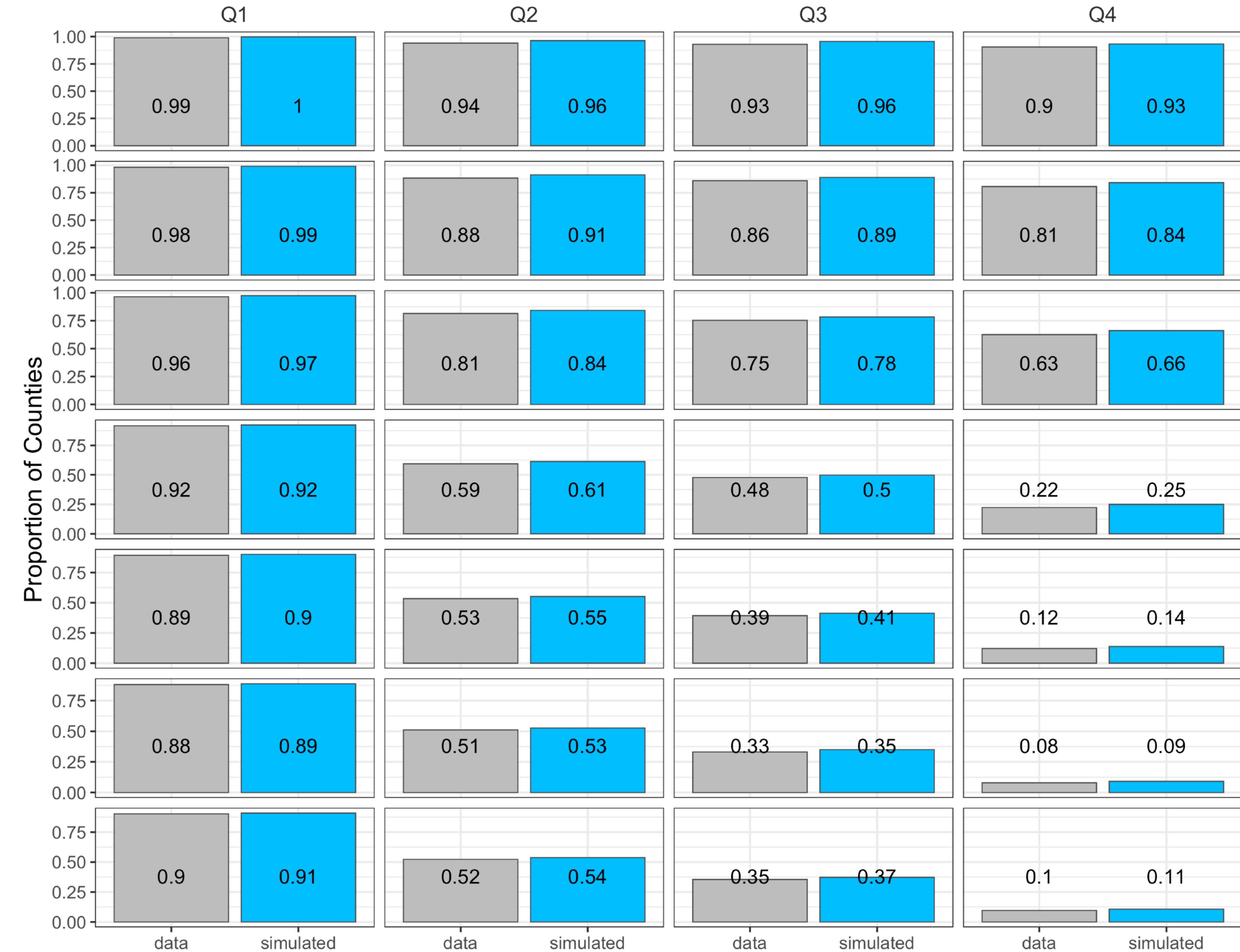
Why Bayesian method?



- If we can reconstruct the posterior distribution, we can get the credible values for the suppressed data by sampling from the posterior distribution.
- In the next two slides, we show whether the Bayesian method creates a plausible distribution

Simulated data vs observed data

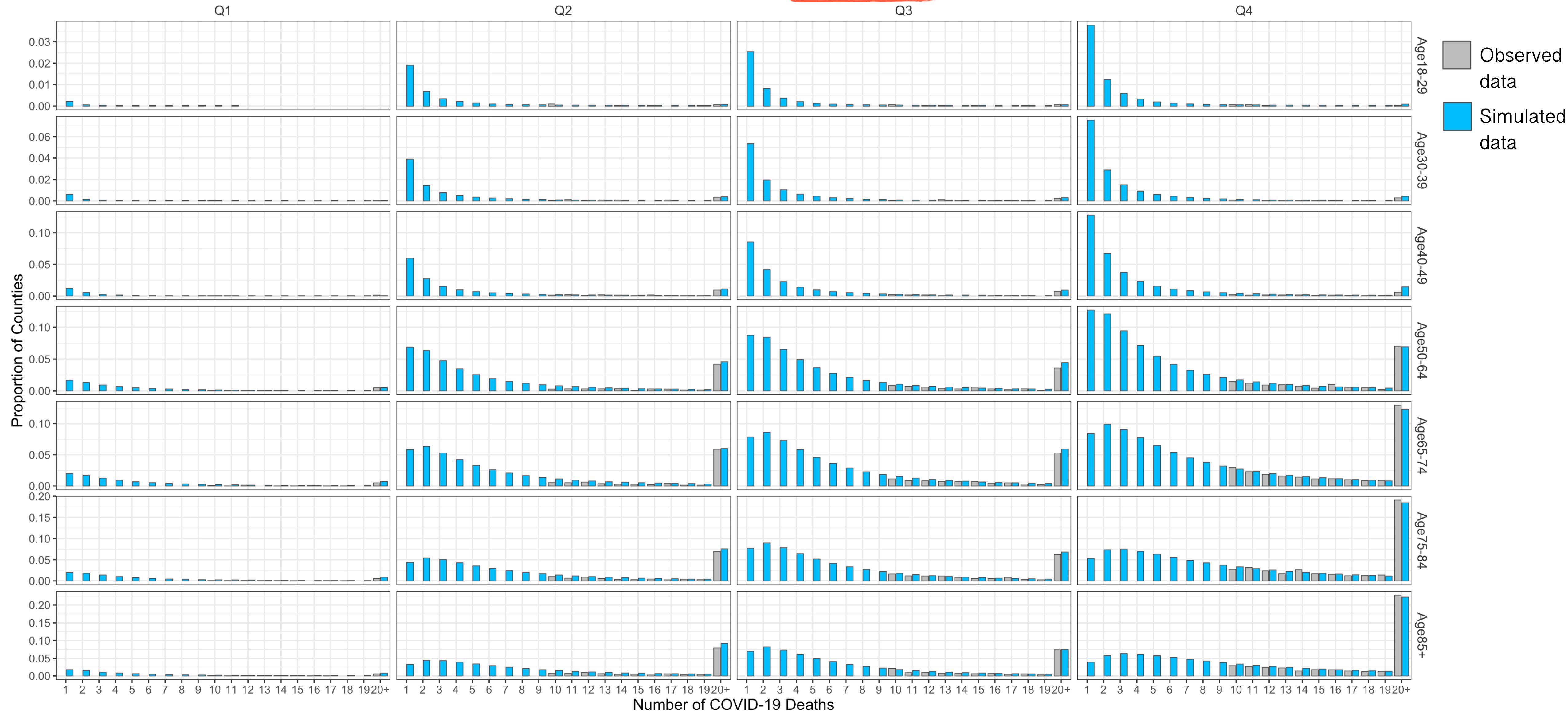
Proportion of counties with zero COVID-19 death (simulated vs data)



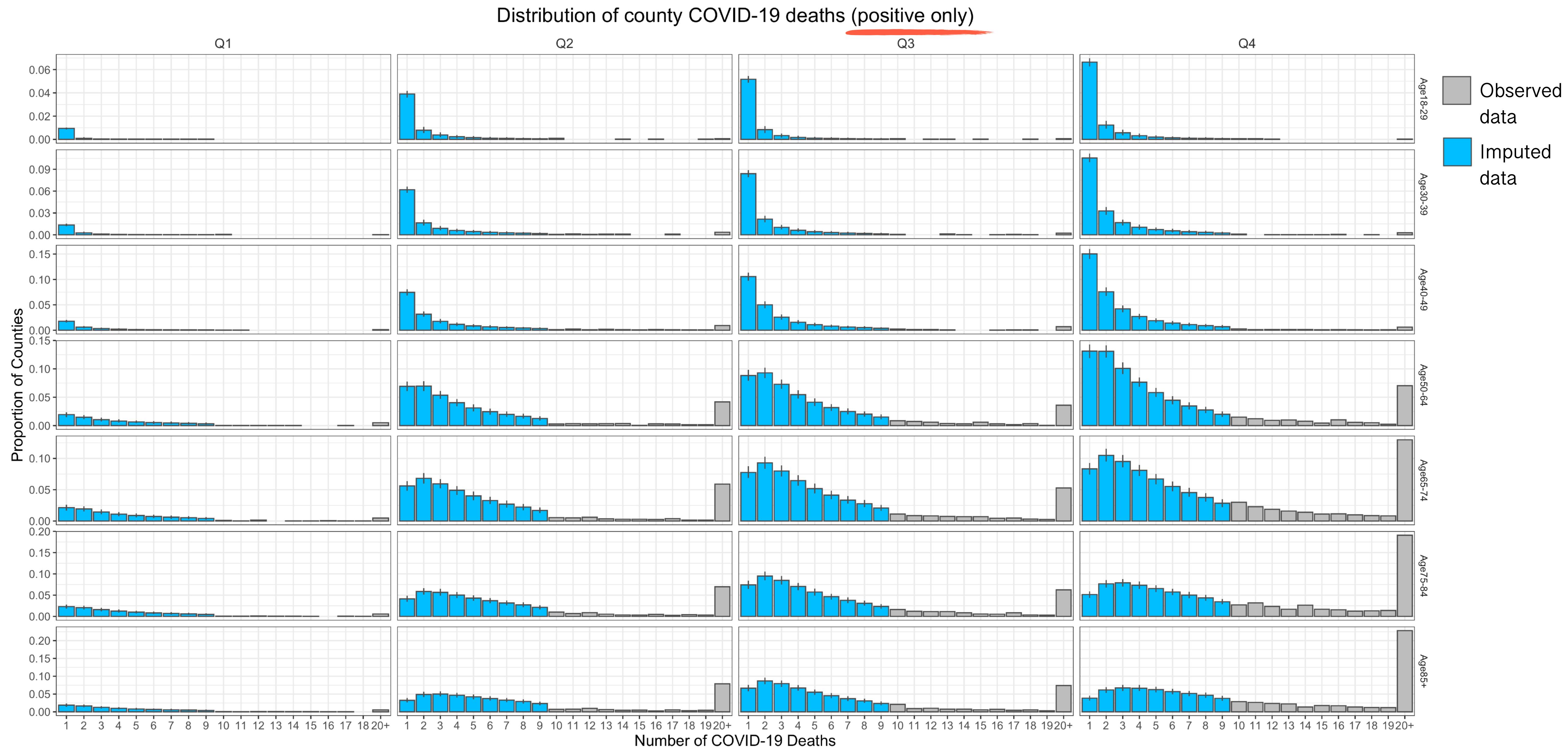
- This is to check whether the bayesian method created a plausible distribution
- The proportions of counties with zero COVID-19 deaths are similar between the simulation from the posterior distribution and observed data

Simulated data vs observed data

Distribution of county-level COVID-19 deaths (positive only; simulated vs data)

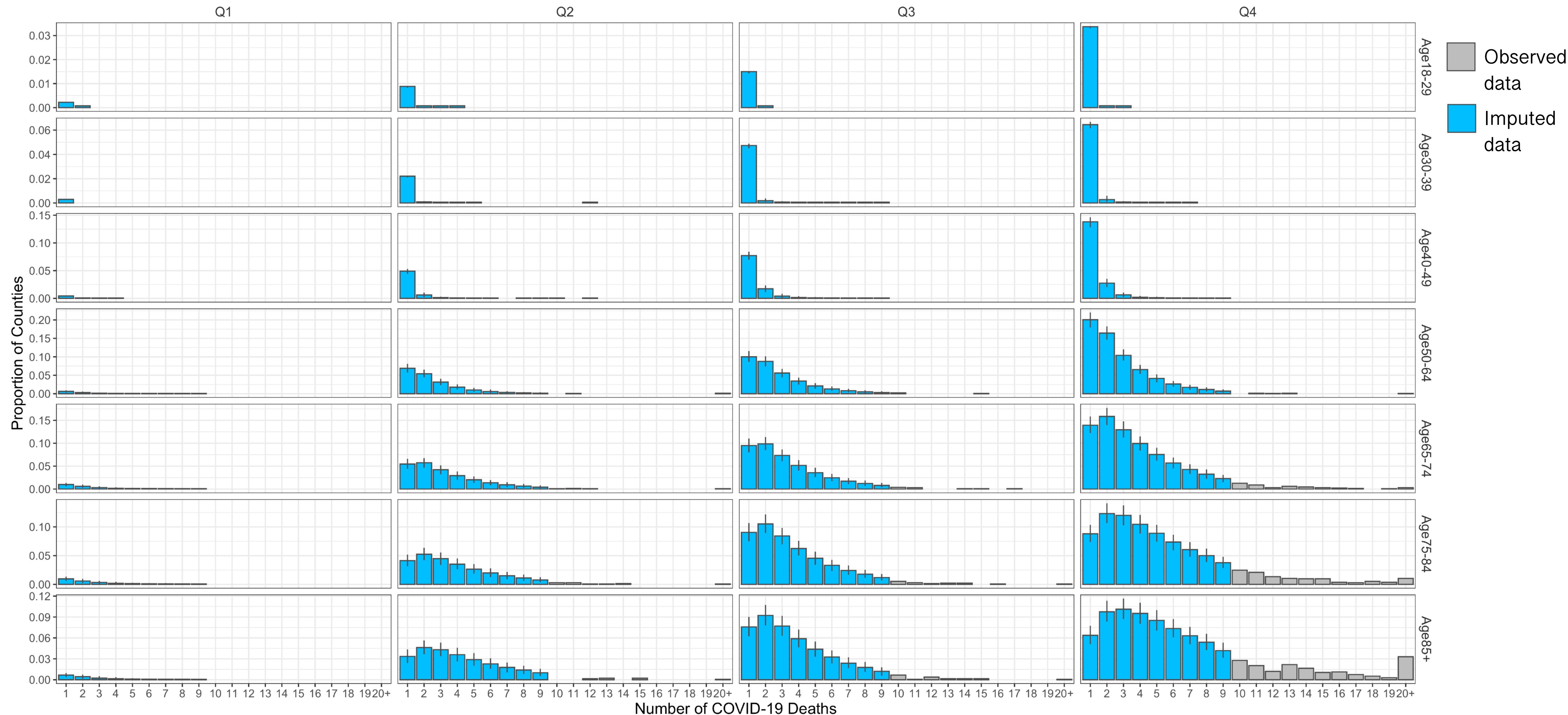


Combine imputed and observed data



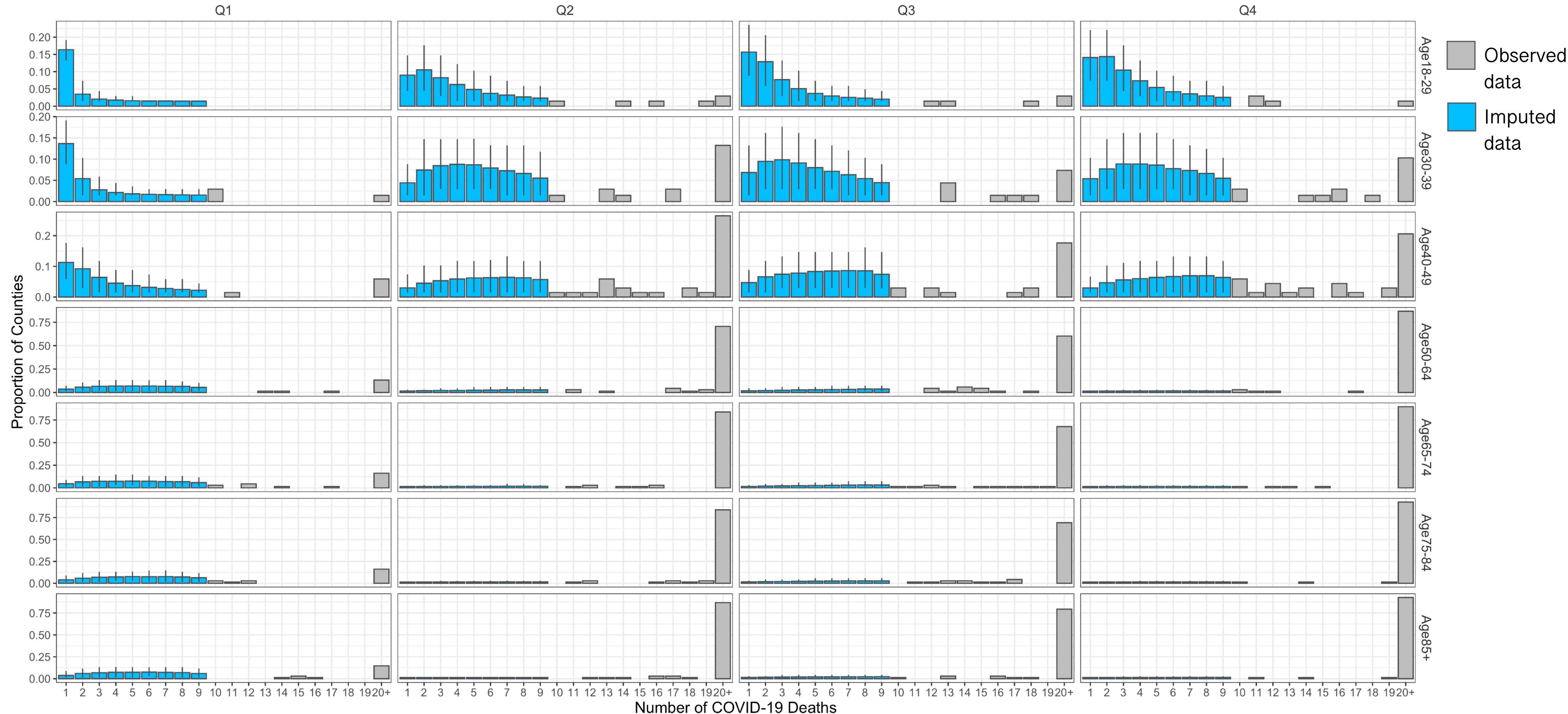
Imputation: Noncore counties

Noncore (positive only): number of counties = 1332

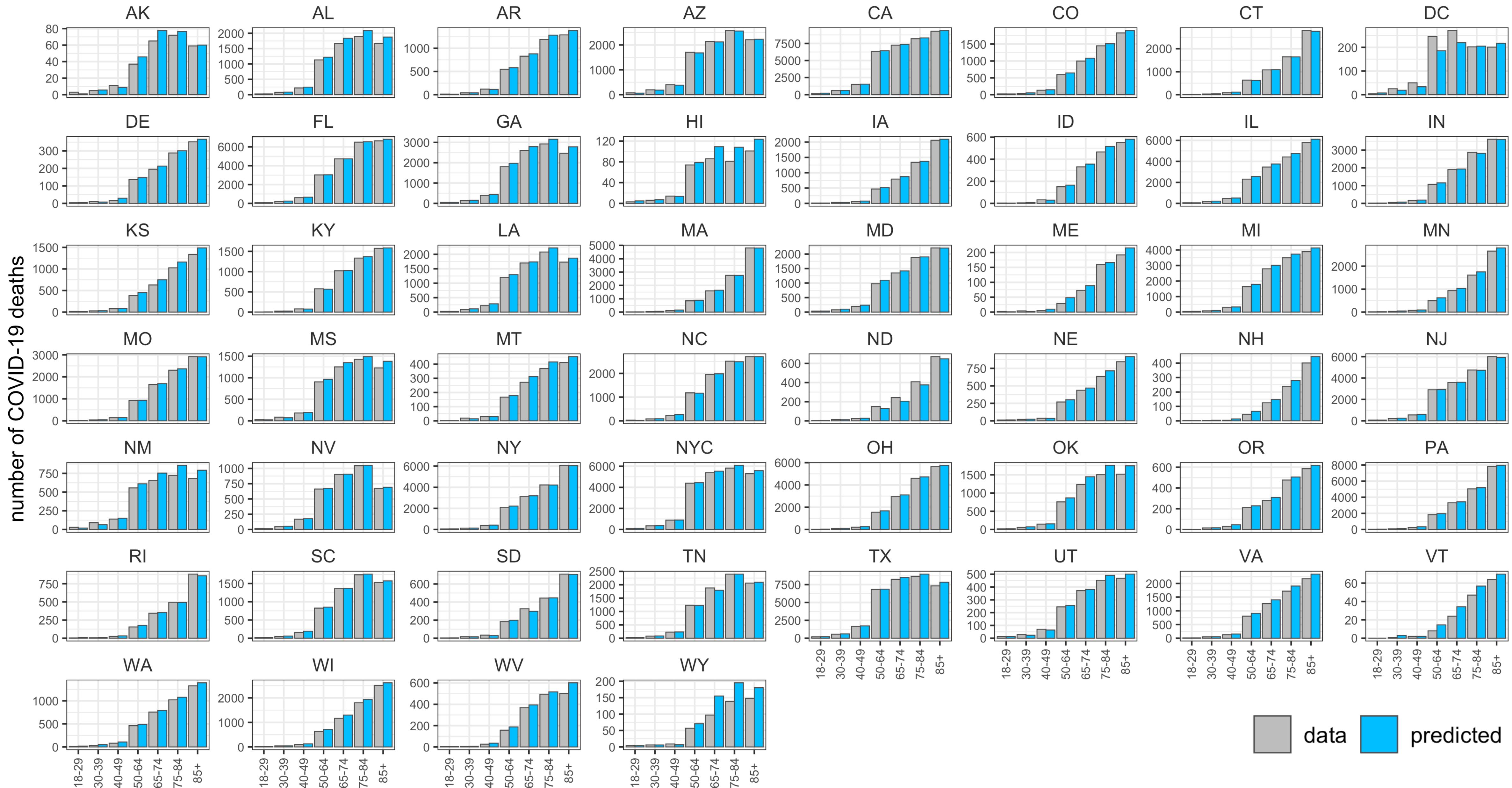


Imputation: Large central metro

Large central metro (positive only): number of counties = 68

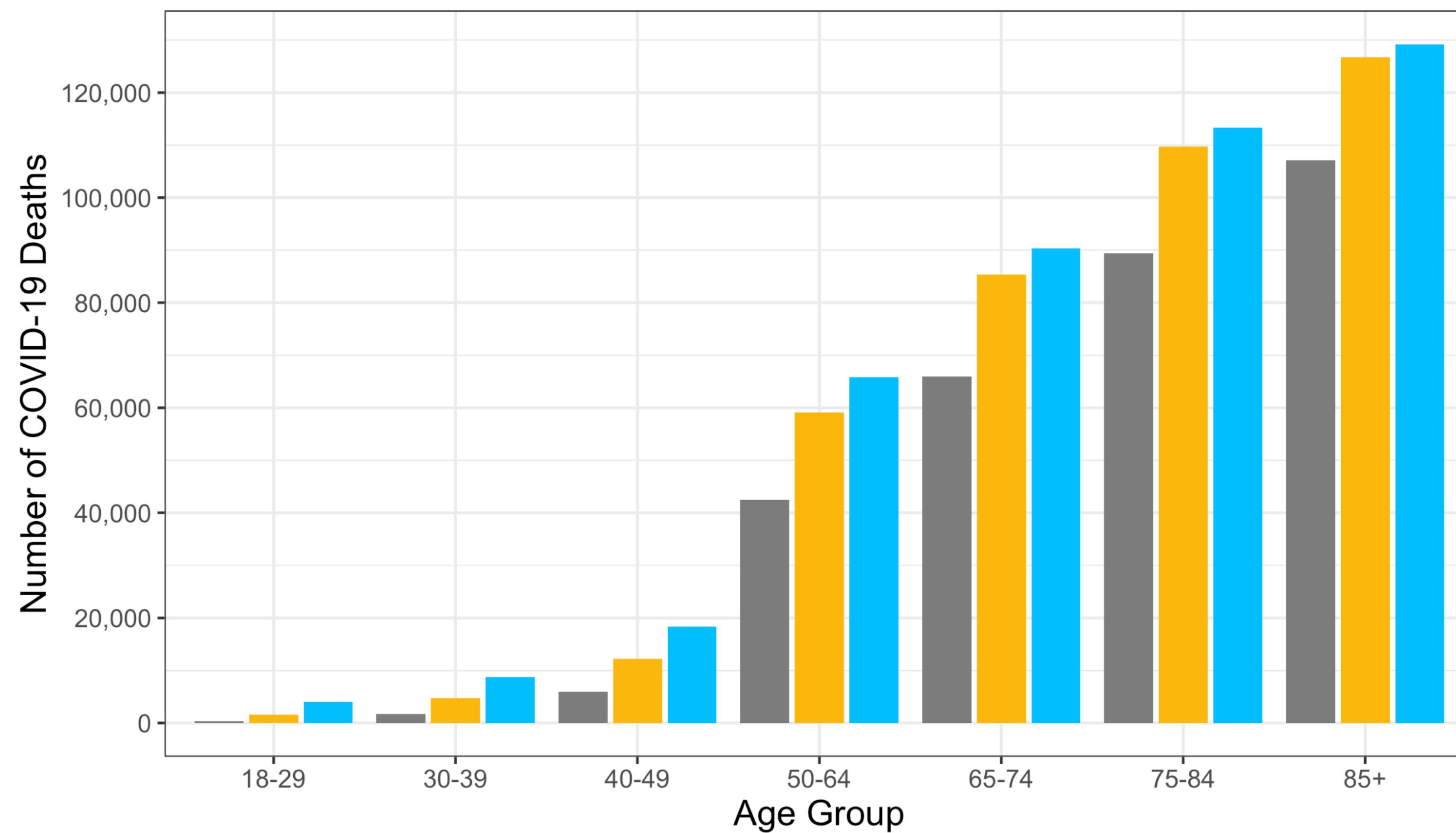


State distribution of COVID-19 deaths



Estimated COVID-19 deaths in 2020

- Uniform imputation is to random sample values between 1 and 9 from a uniform distribution



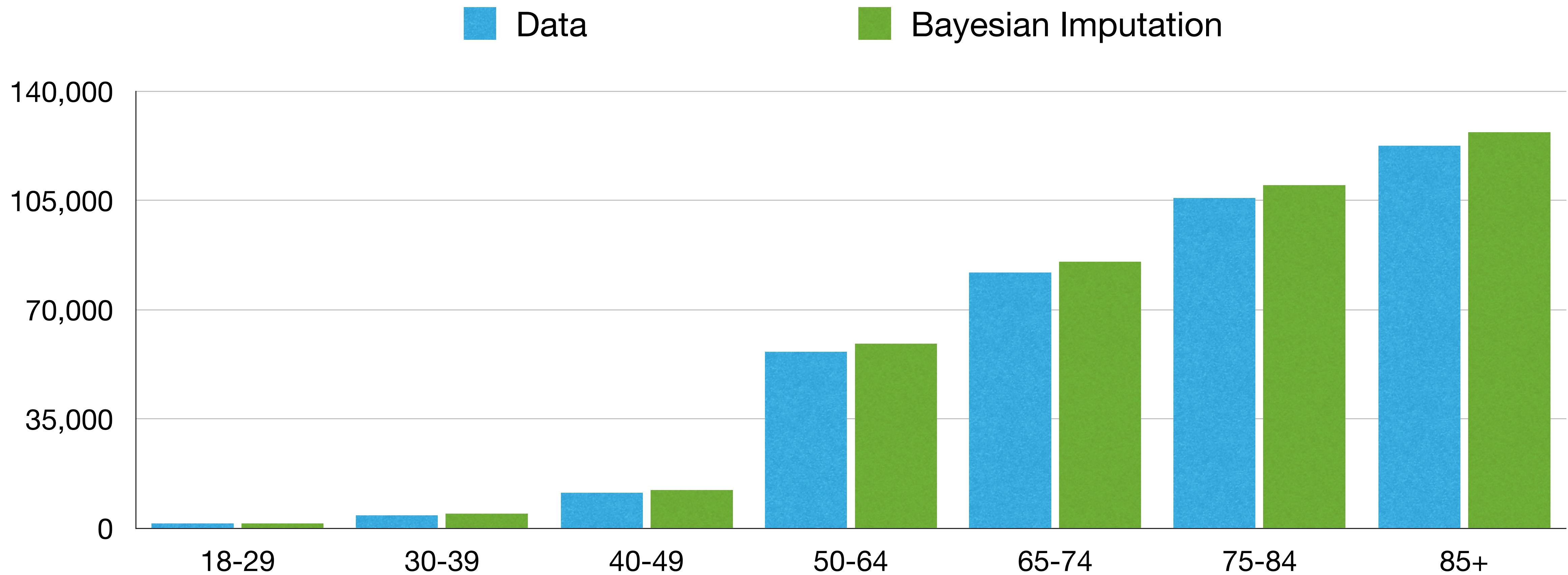
Age group	Estimated COVID-19 deaths in 2020		
	Provisional county-level data (no imputation)	Bayesian imputation (mean)	Uniform imputation (mean)
18-29	347	1,536	3,954
30-39	1,701	4,680	8,756
40-49	5,992	12,259	18,299
50-64	42,538	59,095	65,778
65-74	65,905	85,355	90,354
75-84	89,367	109,802	113,344
≥85	107,056	126,738	129,189

New Results

Age group	No. of COVID-19 related deaths*	Provisional county-level data: no imputation	% under-estimated (no imputation)	Bayesian imputation (95% credible intervals)	% over-estimated (Bayesian imputation)
18-29	1,476	347	76%	1,536 (1,485, 1,591)	4%
30-39	4,272	1,701	59%	4,680 (4,582, 4,769)	10%
40-49	11,291	5,992	46%	12,259 (12,096, 12,437)	9%
50-64	56,630	42,538	24%	59,095 (58,757, 59,418)	4%
65-74	82,059	65,905	18%	85,355 (85,020, 85,707)	4%
75-84	105,964	89,367	14%	109,802 (109,470, 110,141)	4%
≥85	122,488	107,056	11%	126,738 (126,401, 127,081)	3%
Total	384,180	312,906	17%	399,465	4%

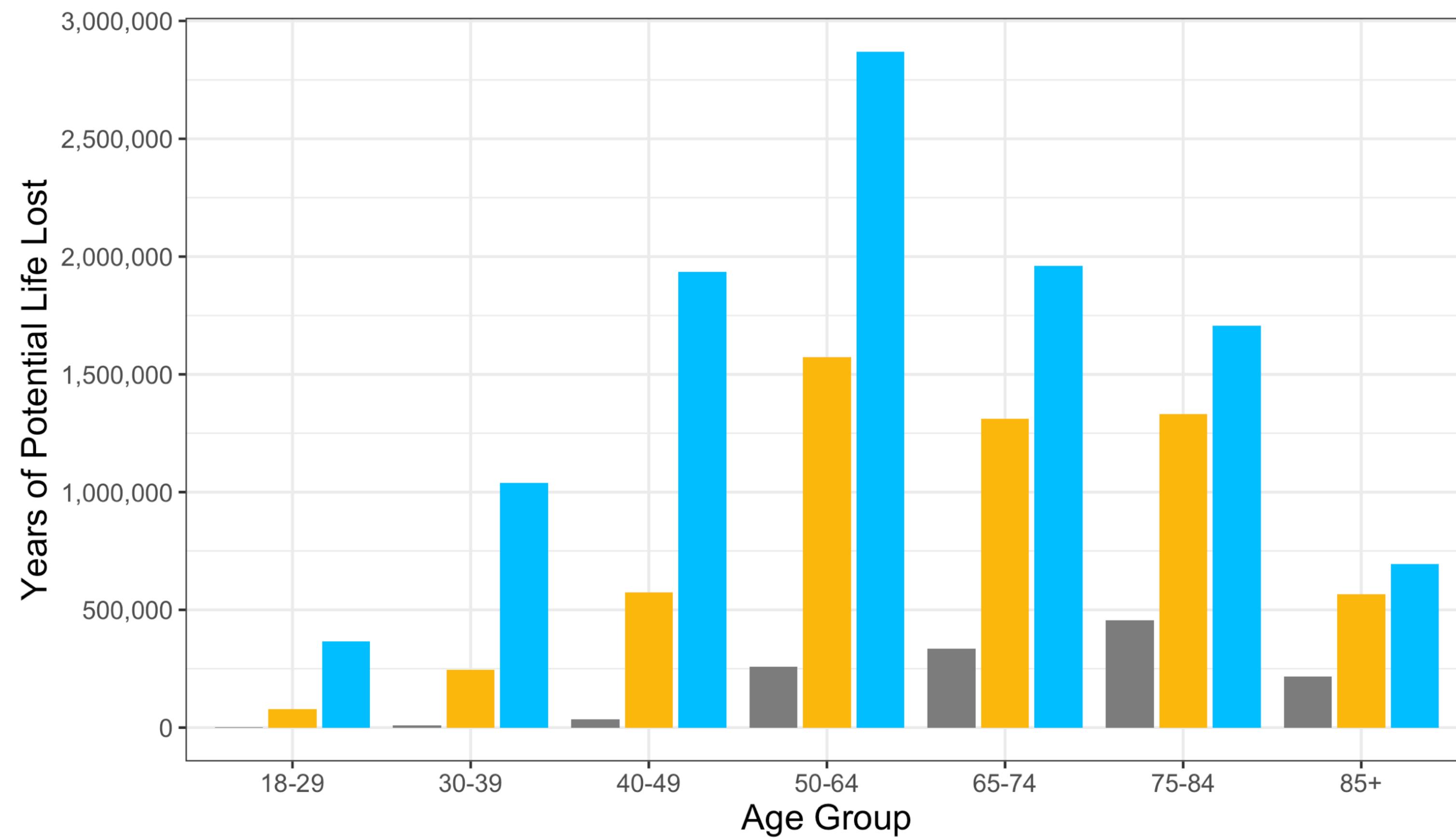
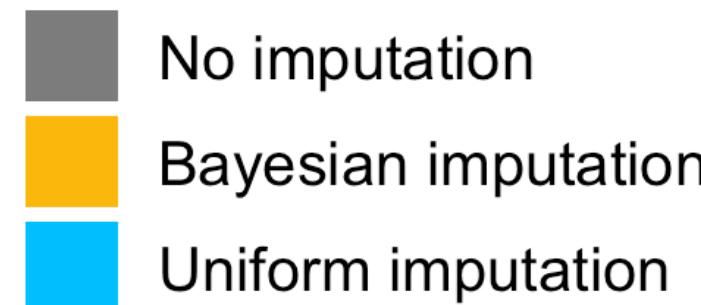
* Data from: <https://www.cdc.gov/nchs/nvss/vsrr/covid19/index.htm>

New Results



Appendix

Total years of potential life loss



Age group	Total years of potential life loss		
	Provisional county-level data (no imputation)	Bayesian imputation (mean)	Uniform imputation (mean)
18-29	1,013	77,982	365,238
30-39	8,198	245,040	1,039,841
40-49	33,872	574,731	1,933,730
50-64	258,457	1,571,465	2,869,634
65-74	334,698	1,310,738	1,961,067
75-84	456,314	1,331,086	1,707,382
≥85	218,466	566,891	694,254

Mortality data

- Life expectancy: <https://www.cdc.gov/nchs/data/vsrr/VSRR10-508.pdf>
- Death data: <https://www.cdc.gov/mmwr/volumes/70/wr/mm7014e1.htm>
- COVID-19 death certificates: <https://www.cdc.gov/mmwr/volumes/70/wr/mm7014e2.htm>
- Total deaths: <https://www.cdc.gov/nchs/nvss/vsrr/covid19/index.htm>

Not updated yet

Model specification

Likelihood

$$P(Y_{ijt} = 0) = 1 - p_{ijt}$$

$$P(Y_{ijt} > 0) = p_{ijt} \cdot \text{Gamma}\left(\text{shape}, \frac{\text{shape}}{\exp(\mu_{ijt})}\right)$$

Estimation of zero part

Time trend: $\text{logit}(p_{ijt}) = \tilde{\alpha}_{ij} + \widetilde{\mathbf{X}}_{ijt}^T \gamma$

County level: $\tilde{\alpha}_{ij} = \theta_i + \nu_{ij}$

State level: $\tilde{\alpha}_i = \theta_0 + \tilde{\nu}_i$

Estimation of positive value

Time trend: $\mu_{ijt} = \alpha_{ij} + \mathbf{X}_{ijt}^T \beta$

County level: $\alpha_{ij} = \phi_i + u_{ij}$

State level: $\alpha_i = \phi_0 + \tilde{u}_i$

Prior distribution

$$\gamma_k \sim \text{Normal}(0, 10)$$

$$\alpha_{ij} \sim \text{Normal}(0, 10)$$

$$\beta_l \sim \text{Normal}(0, 10)$$

$$\text{shape} \sim \text{gamma}(0.1, 0.1)$$

$$\theta_i \sim \text{Normal}(0, 1)$$

$$\theta_0 \sim \text{Normal}(0, 1)$$

$$\nu_{ij} \sim \text{half-T}(3, 0, 2.5), \nu_{ij} > 0$$

$$\tilde{\nu}_i \sim \text{half-T}(3, 0, 2.5), \tilde{\nu}_i > 0$$

$$\phi_i \sim \text{Normal}(0, 1)$$

$$\phi_0 \sim \text{Normal}(0, 1)$$

$$u_{ij} \sim \text{half-T}(3, 0, 2.5), u_{ij} > 0$$

$$\tilde{u}_i \sim \text{half-T}(3, 0, 2.5), \tilde{u}_i > 0$$

Constraint of the missing values

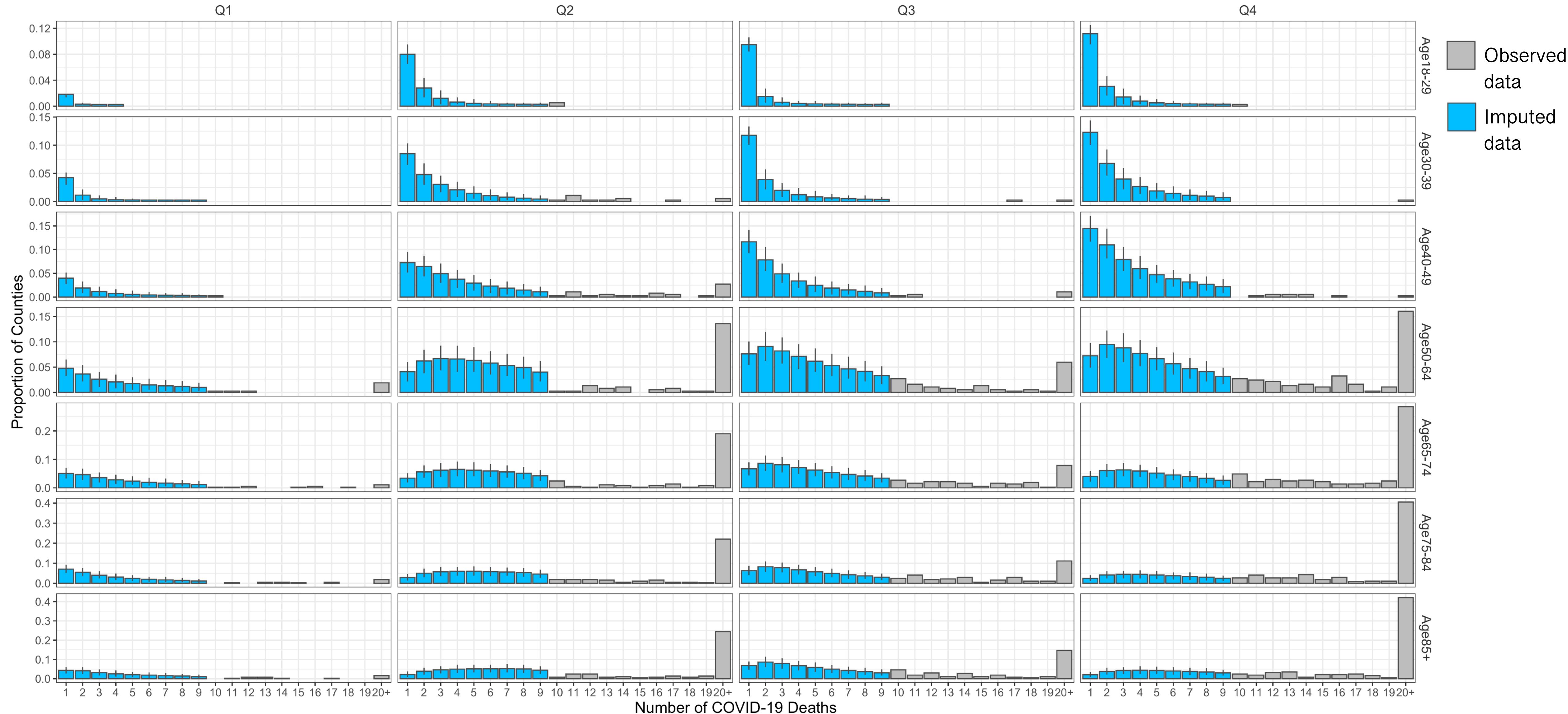
$$y_{miss_{ijt}} \sim \text{Uniform}(0.6, 9.4)$$

Notation

- i : state index
- j : county index
- t : quarter index
- Y_{ijt} : COVID-19 death count at county j state i at quarter t
- $\widetilde{\mathbf{X}}_{ijt}^T$: the design matrix for **the zero part**, including the intercept and quarters
- X_{ijt}^T : the design matrix for **the estimation of positive values**, including the intercept, age groups, types of counties, quarters, and log of population size
- k : index of the fixed effect in zero part
- l : index of the fixed effect in the estimate of positive value

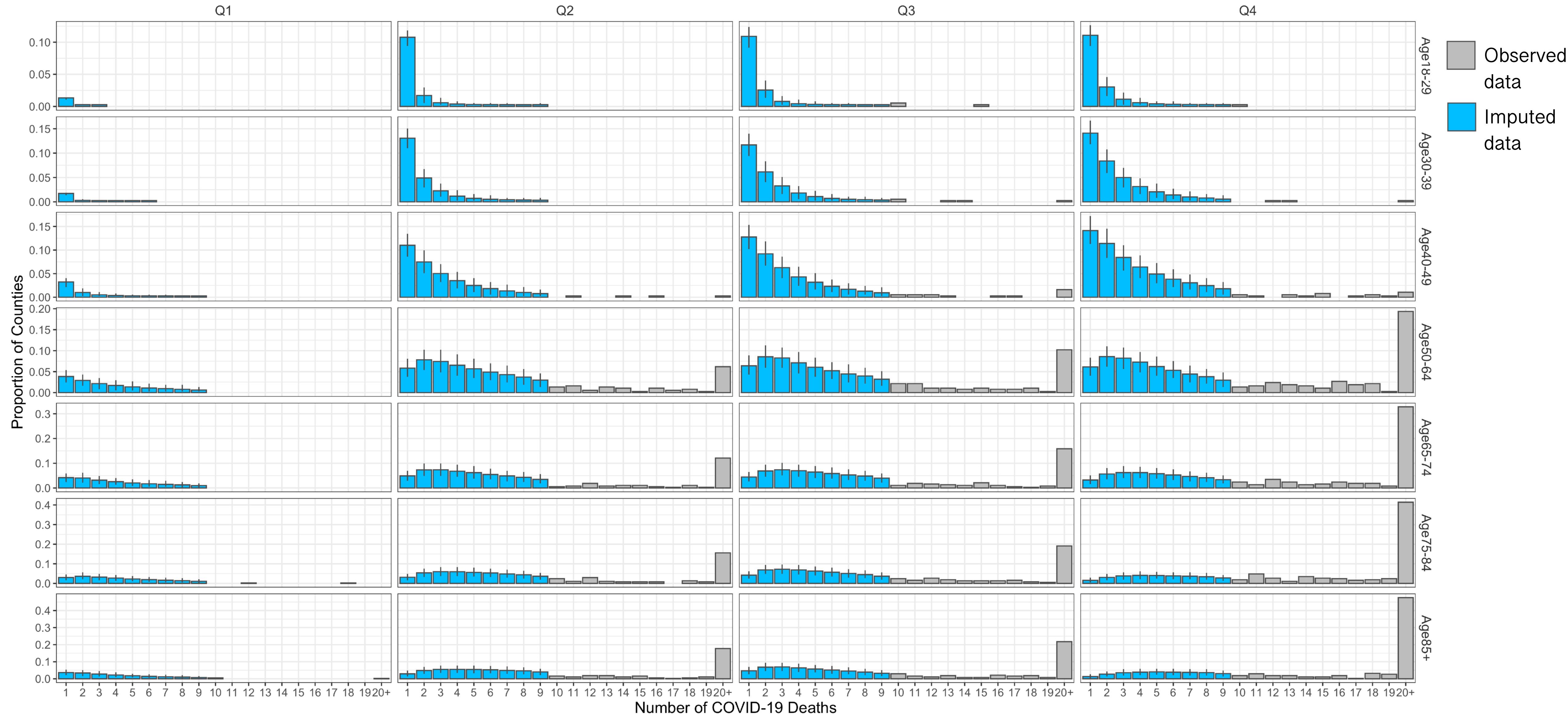
Imputation: Large fringe metro

Large fringe metro (positive only): number of counties = 368



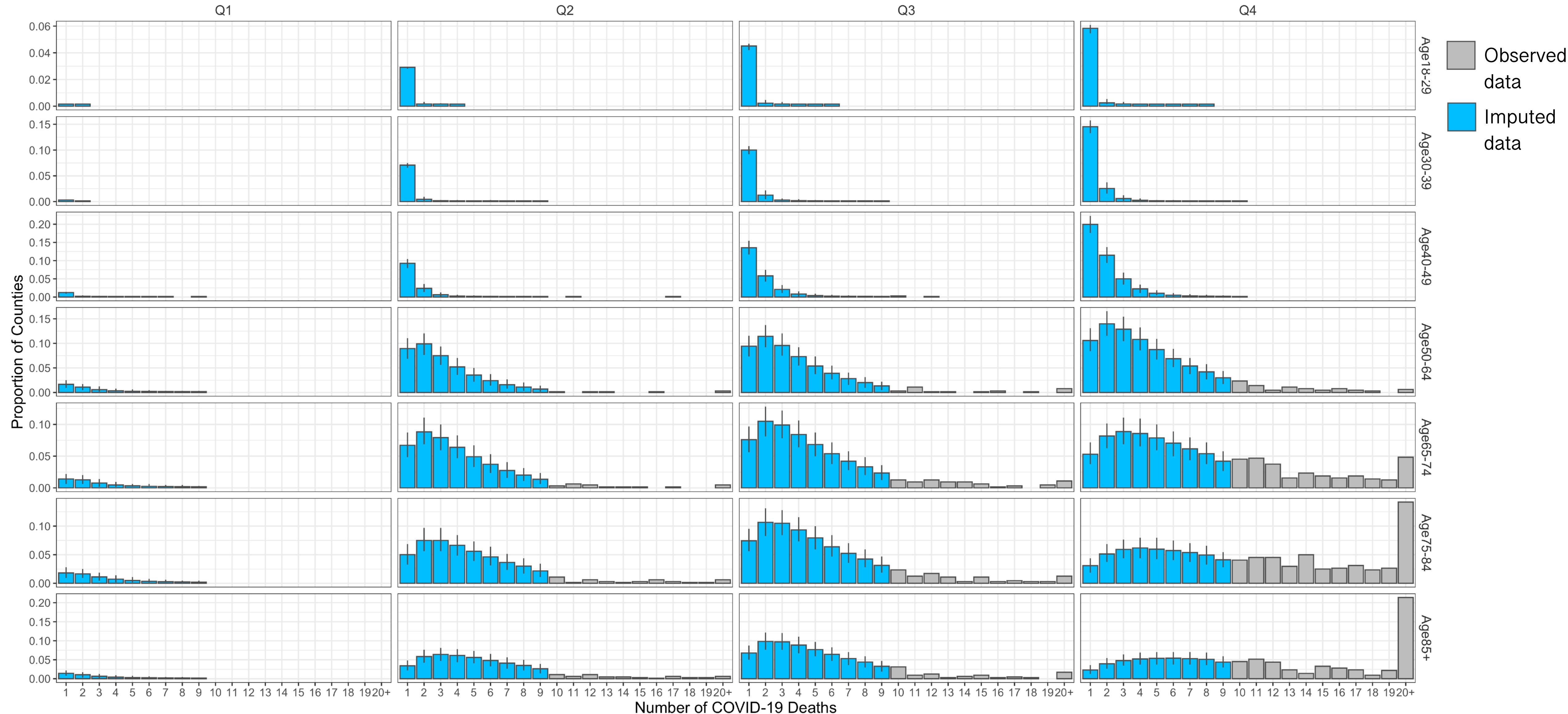
Imputation: Medium metro

Medium metro (positive only): number of counties = 372



Imputation: Micropolitan

Micropolitan (positive only): number of counties = 641



Imputation: Small metro

Small metro (positive only): number of counties = 357

