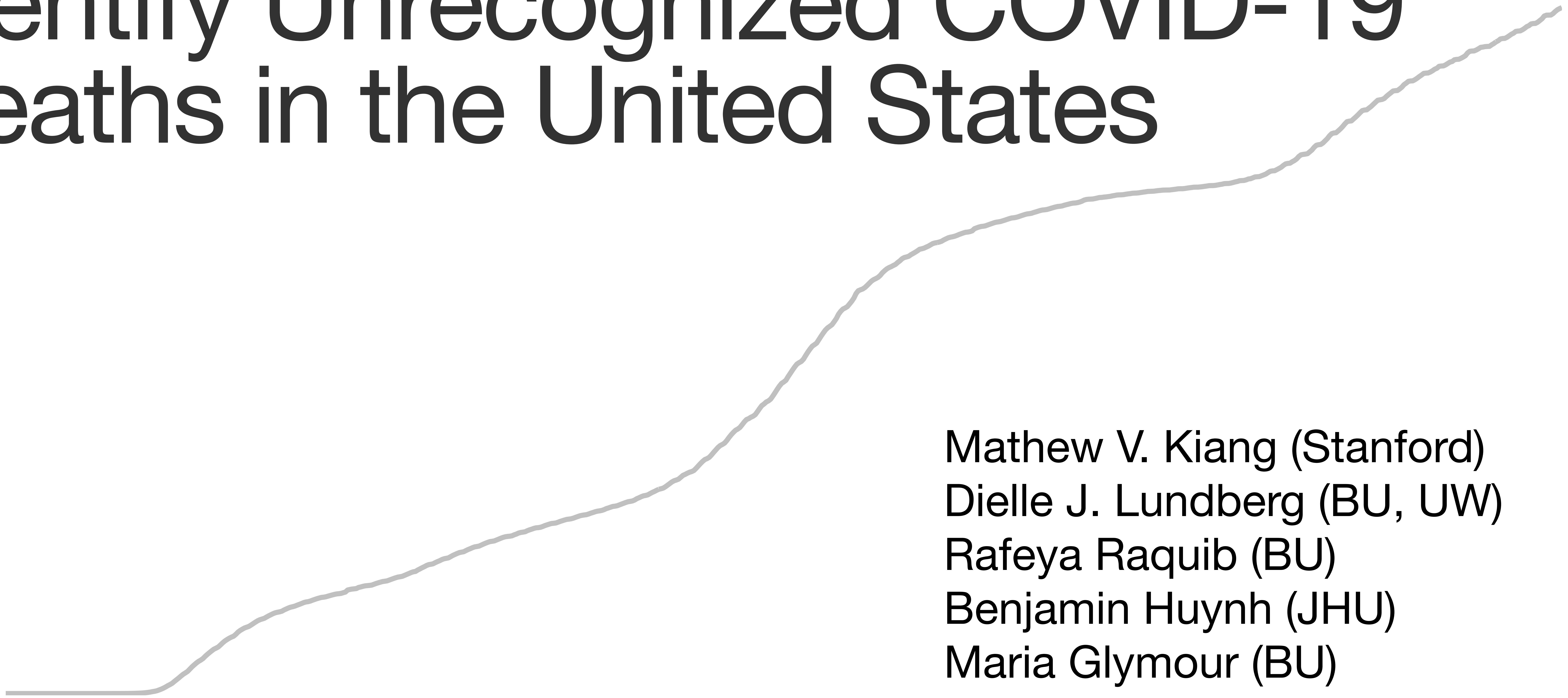
Applying Machine Learning to Identify Unrecognized COVID-19 Deaths in the United States



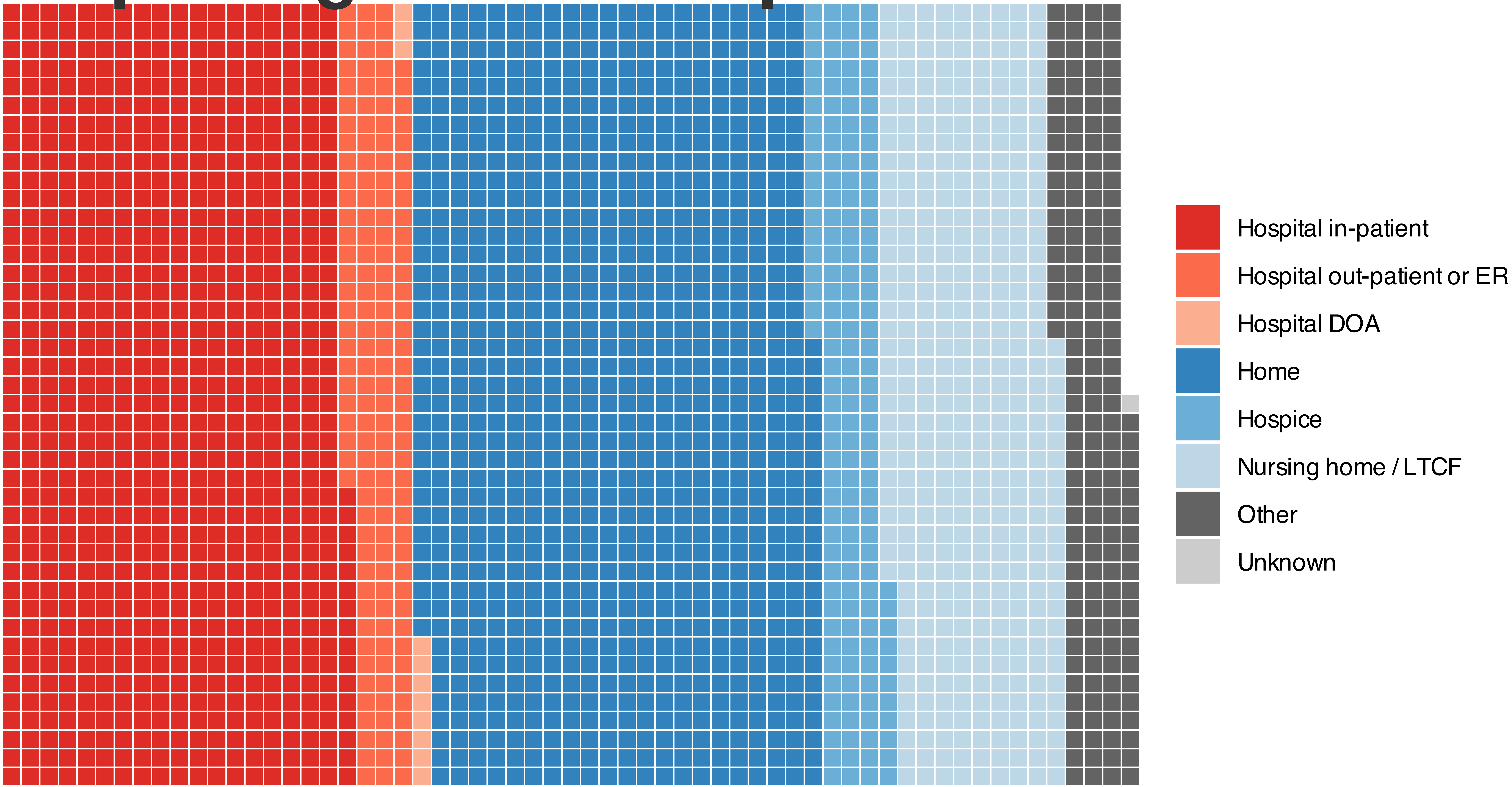
Mathew V. Kiang (Stanford)
Dielle J. Lundberg (BU, UW)
Rafeya Raquib (BU)
Benjamin Huynh (JHU)
Maria Glymour (BU)
Andrew C. Stokes (BU)

Over 824,000 COVID-19 deaths by Dec 2021

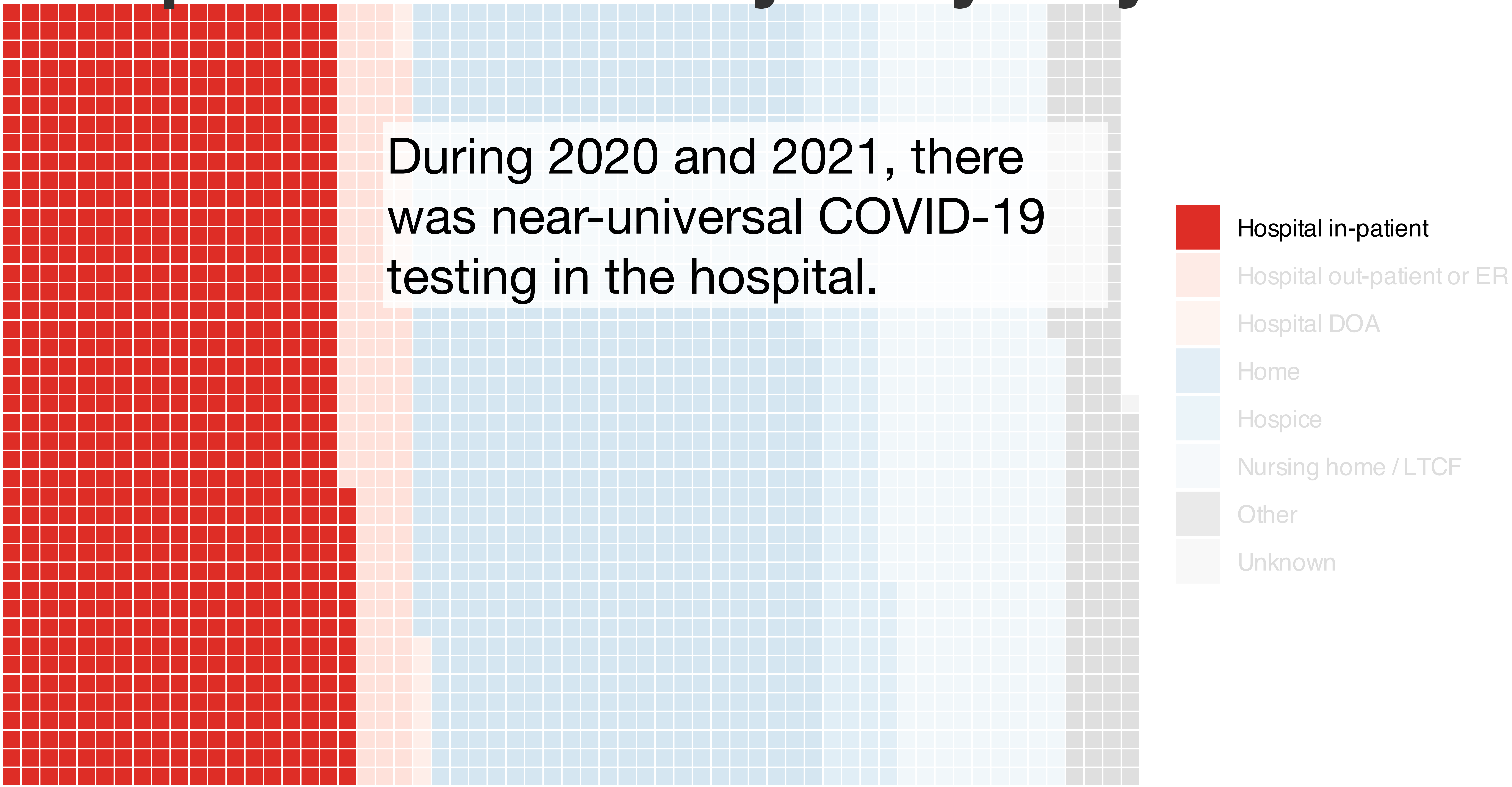
But these are just *official* COVID-19 deaths

Can we use machine learning to get more accurate estimates of COVID-19 mortality?

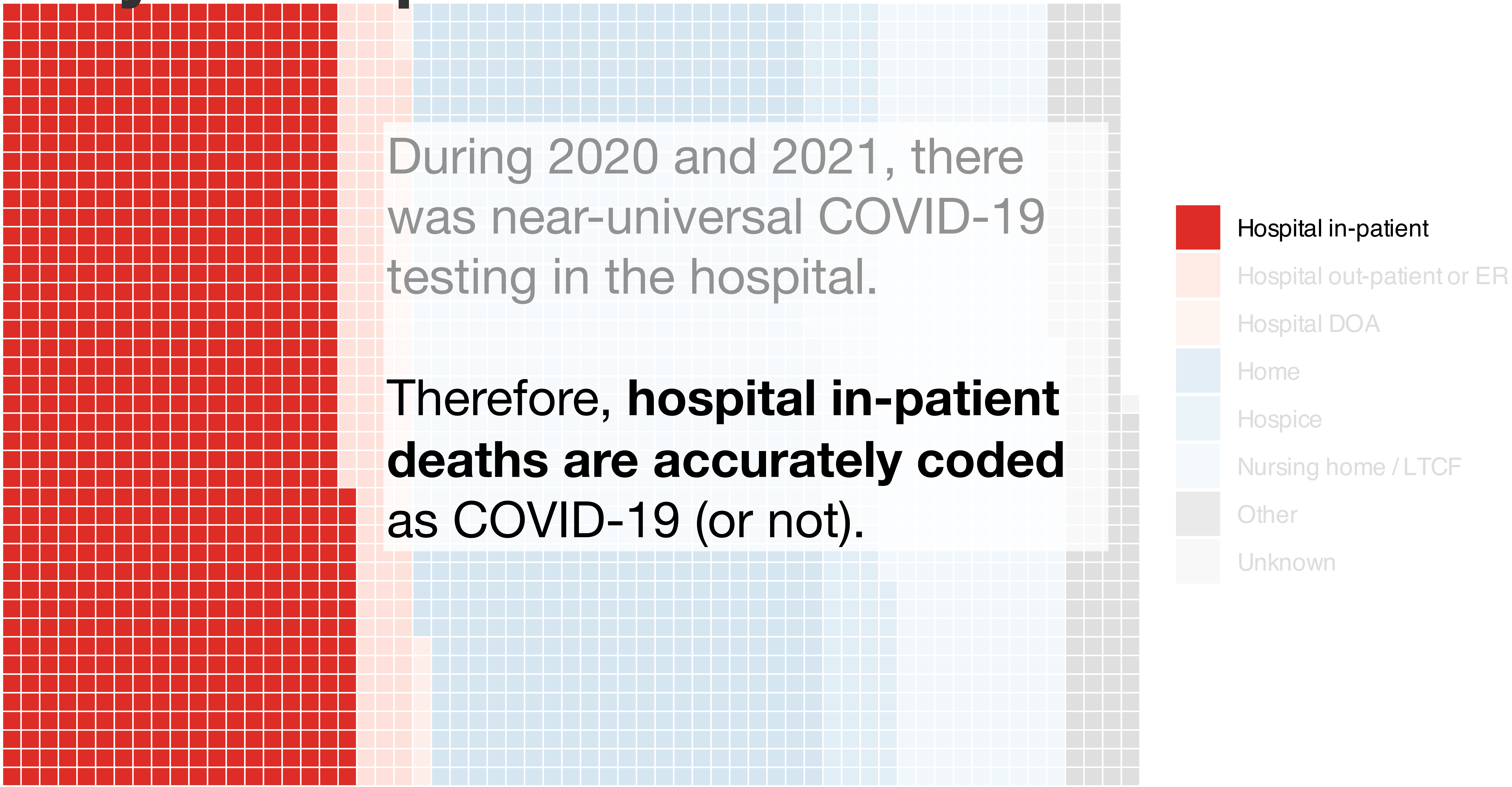
Exploiting variation in place of death



Hospitals tested nearly everybody



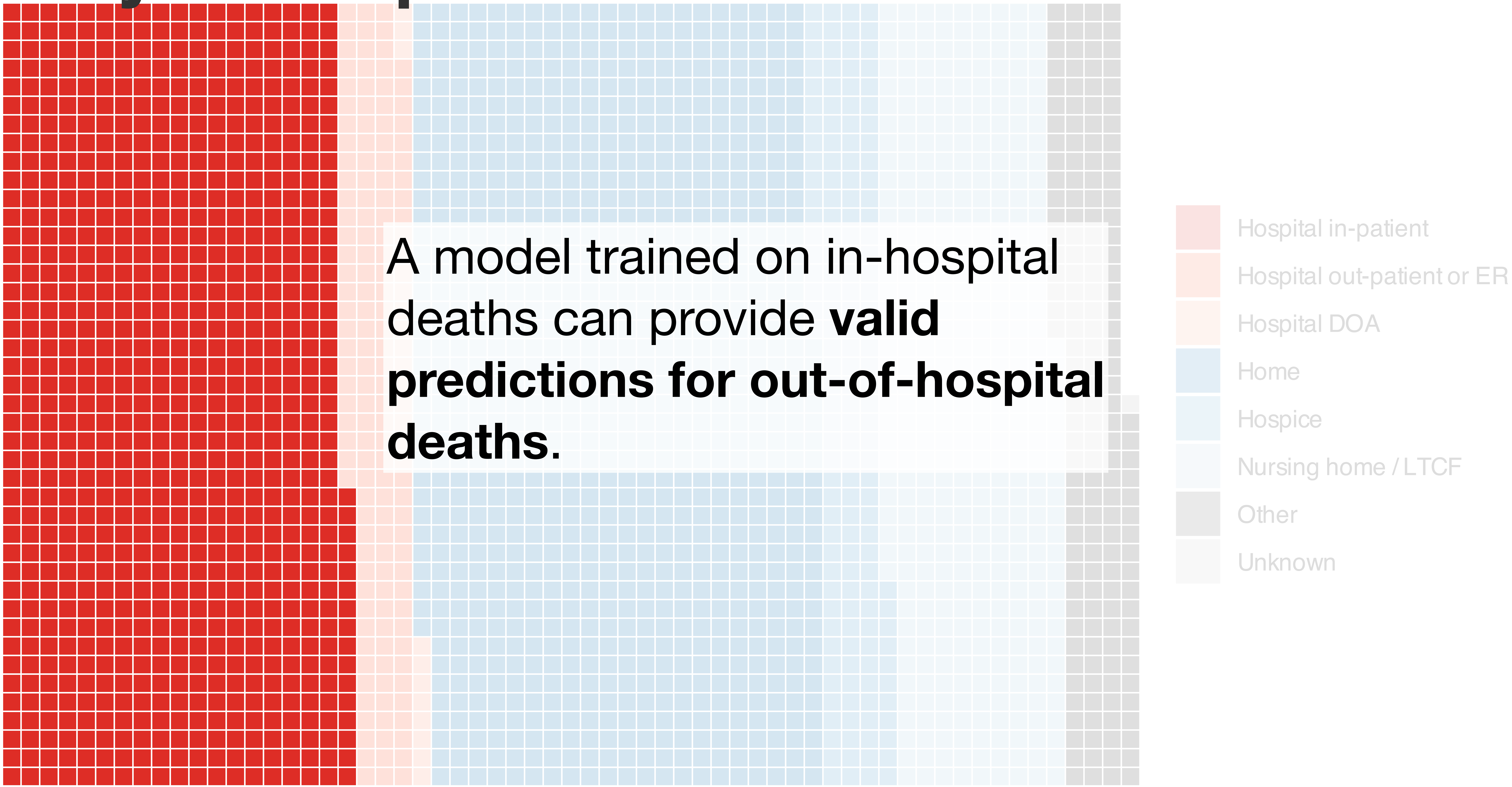
Key assumption 1



During 2020 and 2021, there was near-universal COVID-19 testing in the hospital.

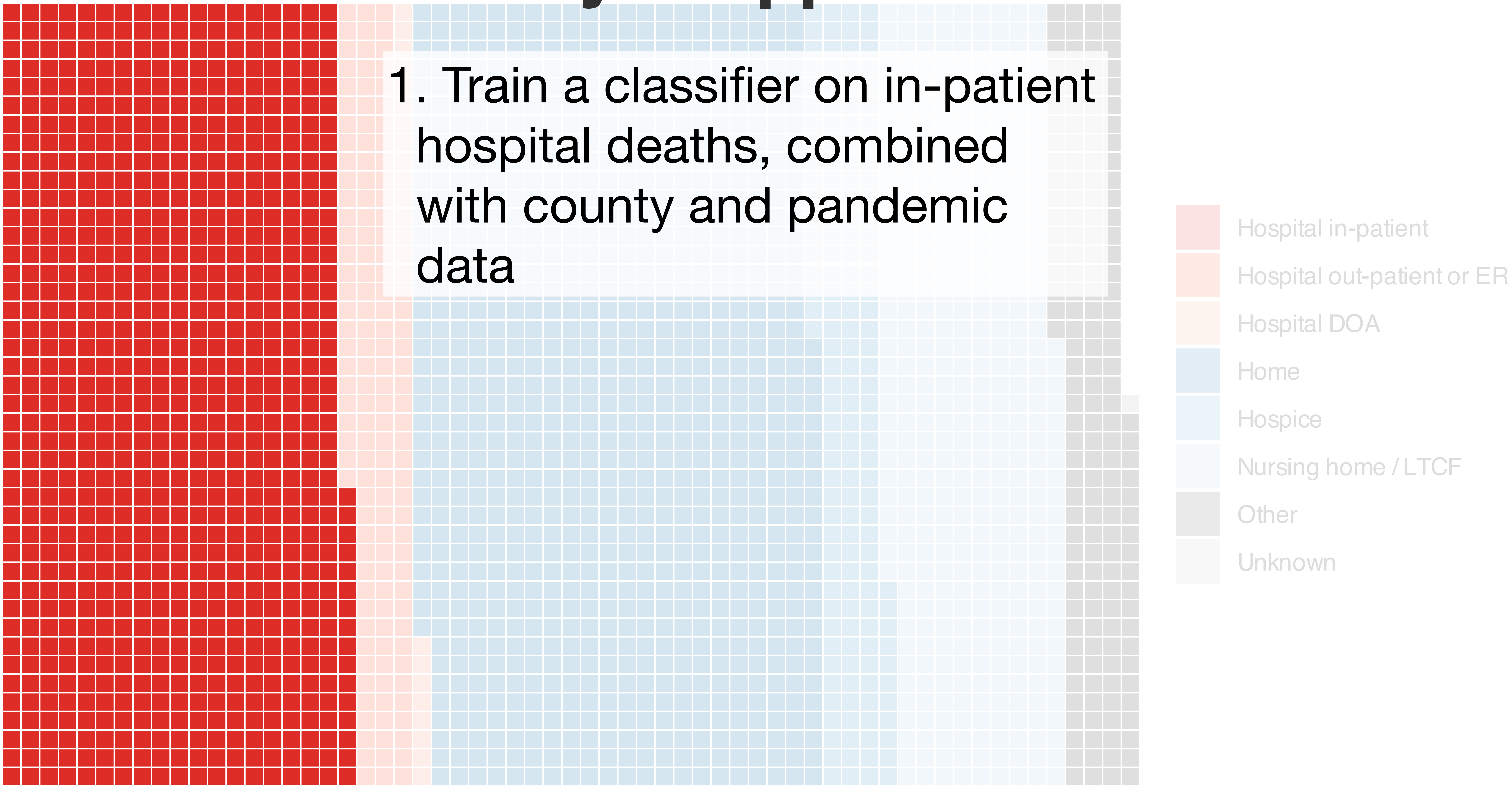
Therefore, **hospital in-patient deaths are accurately coded** as COVID-19 (or not).

Key assumption 2



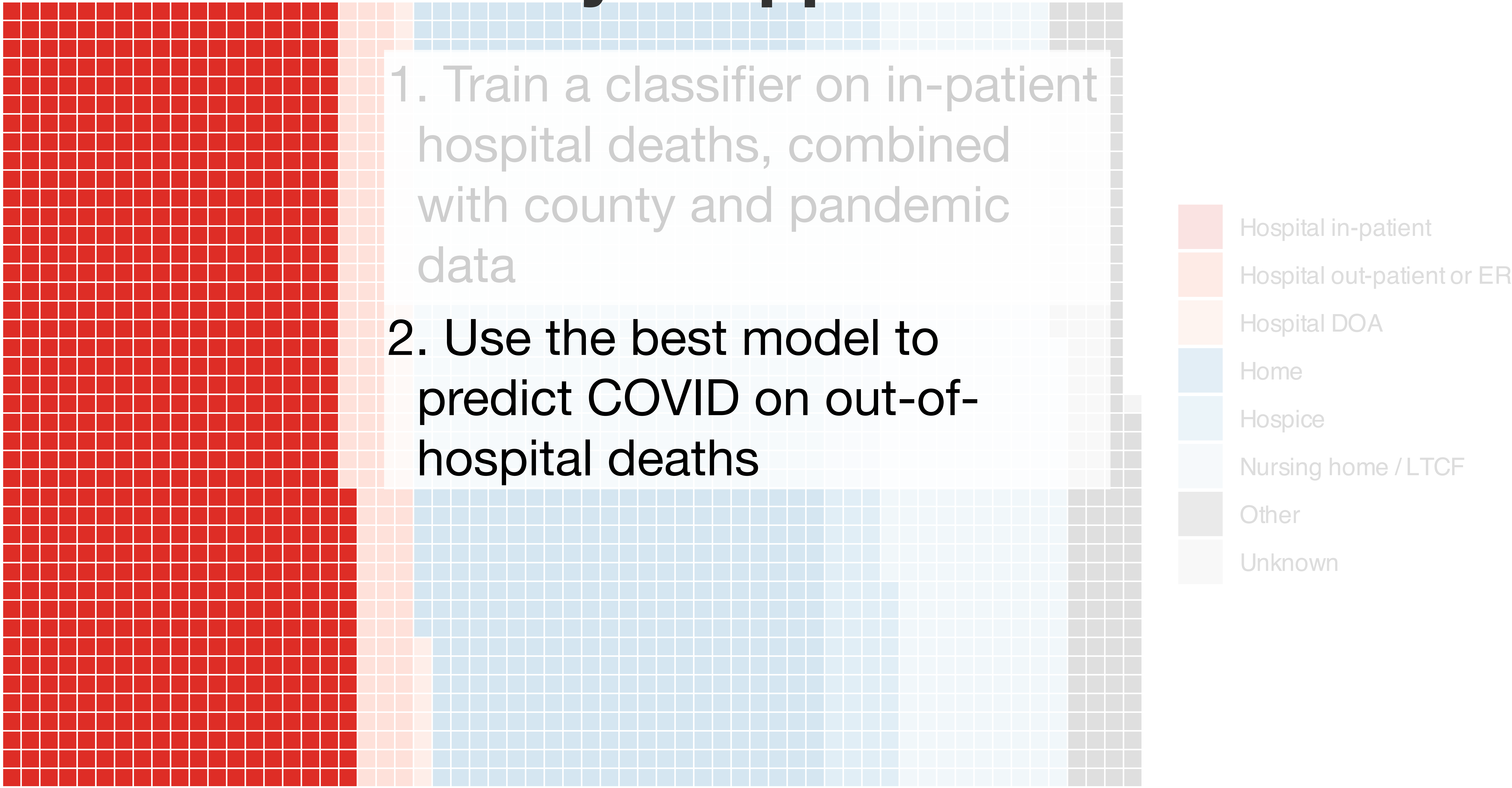
A model trained on in-hospital deaths can provide **valid predictions for out-of-hospital deaths.**

Overview of analytic approach



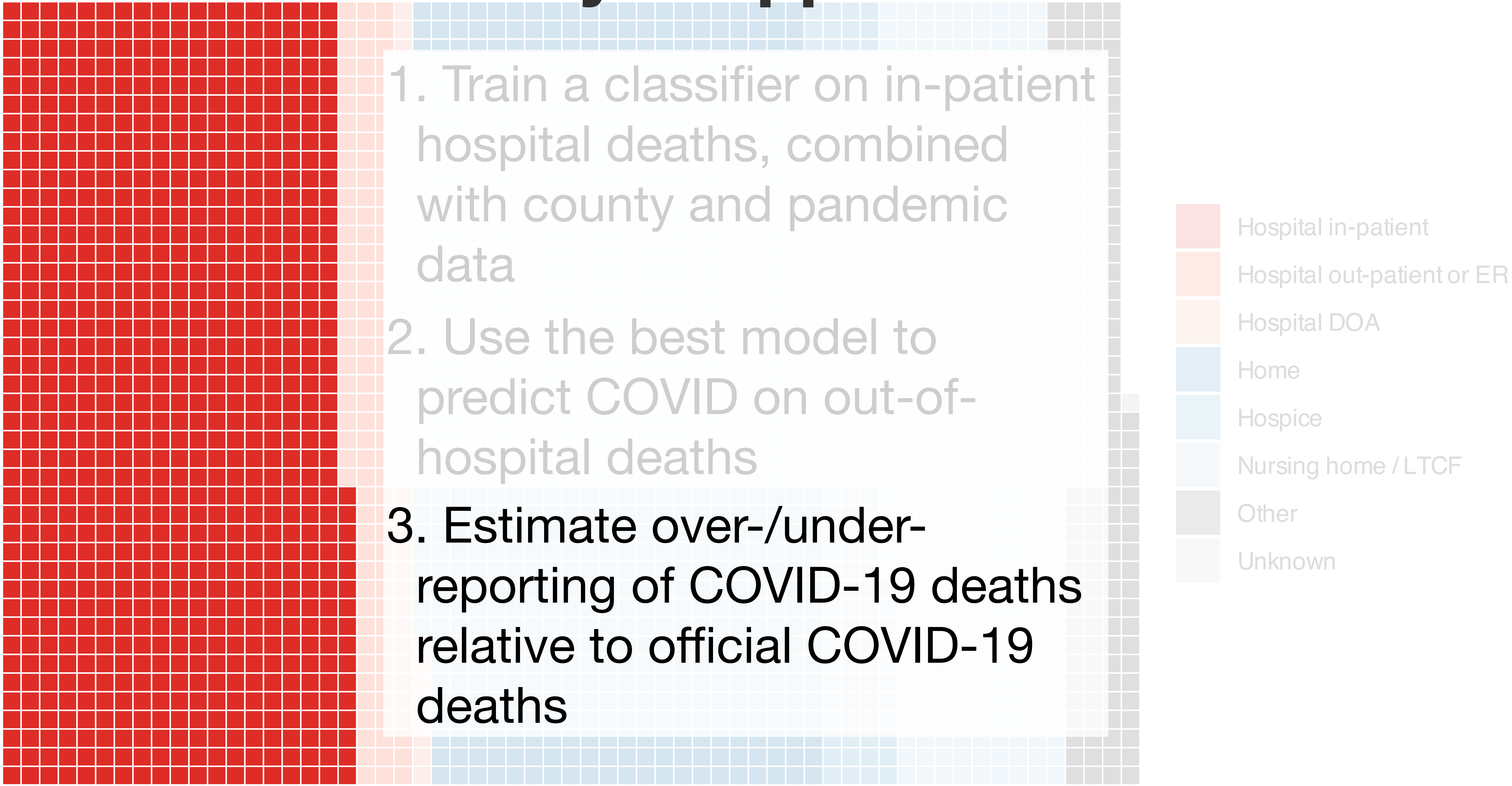
1. Train a classifier on in-patient hospital deaths, combined with county and pandemic data

Overview of analytic approach



1. Train a classifier on in-patient hospital deaths, combined with county and pandemic data
2. Use the best model to predict COVID on out-of-hospital deaths

Overview of analytic approach



1. Train a classifier on in-patient hospital deaths, combined with county and pandemic data
2. Use the best model to predict COVID on out-of-hospital deaths
3. Estimate over-/under-reporting of COVID-19 deaths relative to official COVID-19 deaths

Analytic data sets

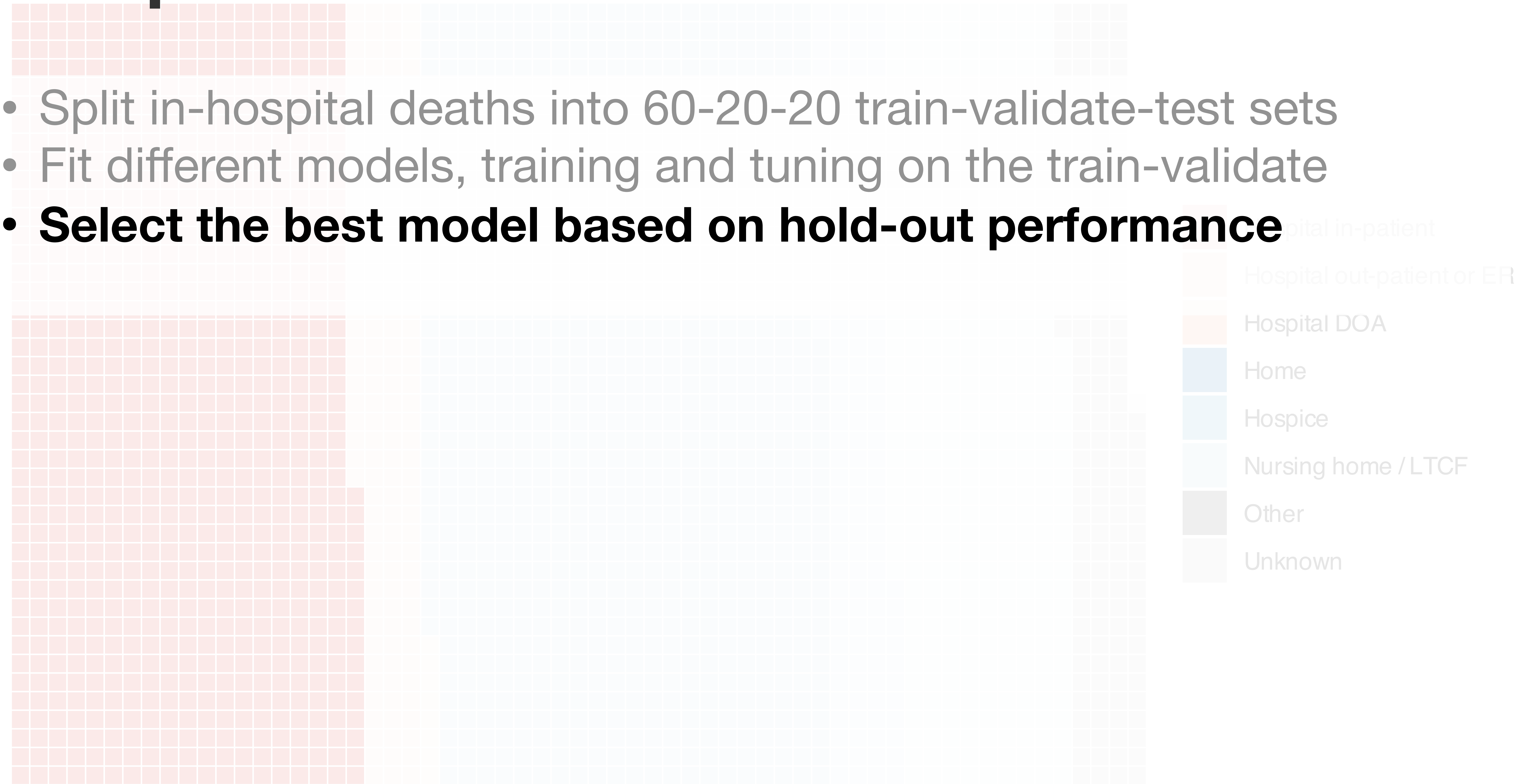
- Individual-level death certificates
- County-level characteristics (static, pre-2020)
- COVID-19 covariates (time-varying)
- **Total N = 5.7 million deaths** (> 25 y; natural causes only):
  - ~1.9 million in-hospital deaths ("gold standard")
  - ~3.8 million out-of-hospital deaths (prediction)

Step 1: Train a classifier

- Split in-hospital deaths into 60-20-20 train-validate-test sets
- **Fit different models, training and tuning on the train-validate:**
  - Total of 22 model-covariate set combinations
    - Logistic regression with LASSO penalty
    - Logistic regression using Elastic Net penalty
    - Random forests
    - Bayesian Additive Regression Trees (BART)
    - LightGBM
    - XGBoost
  - Used Bayesian optimization to find best hyperparameters

Step 1: Train a classifier

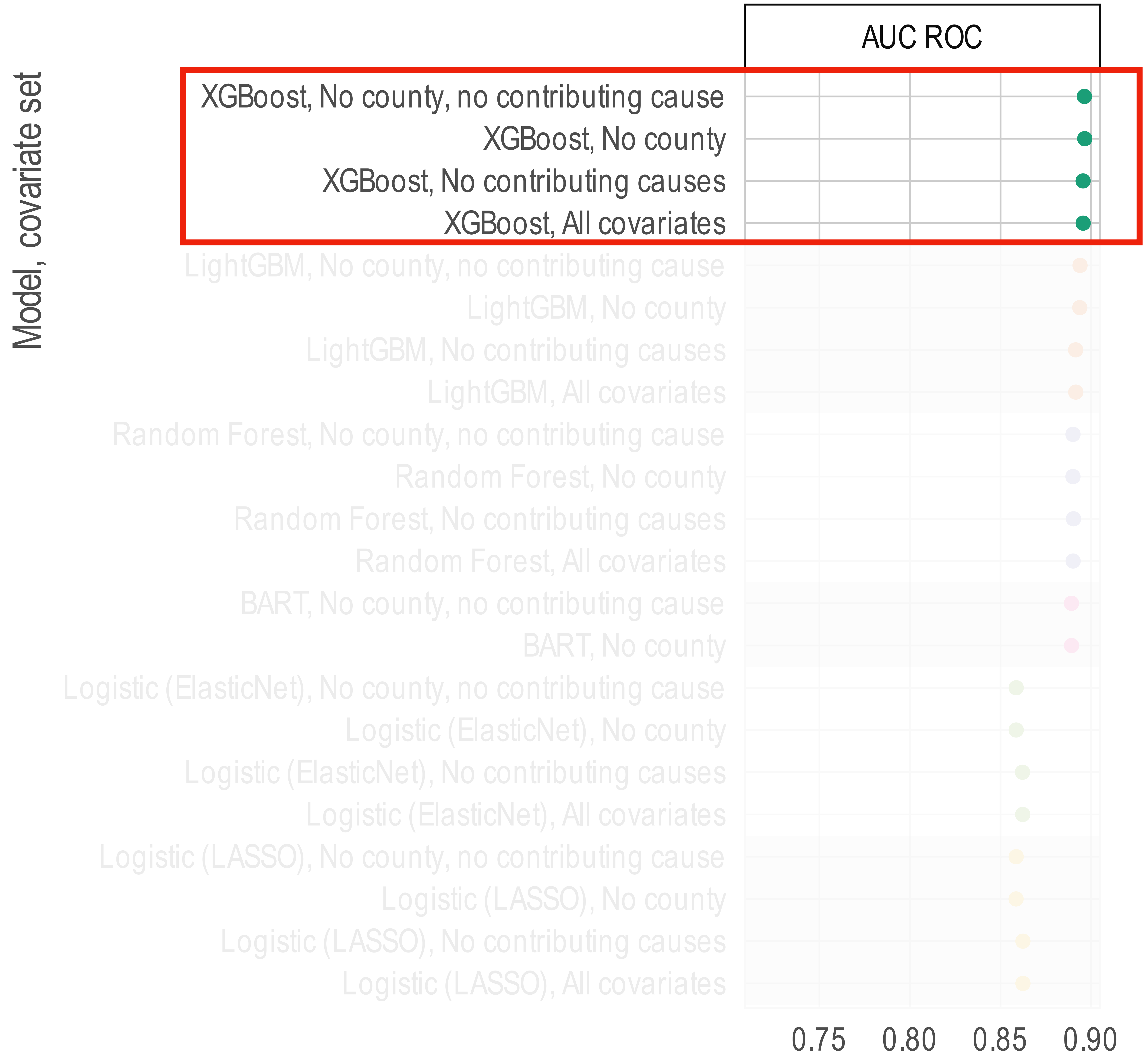
- Split in-hospital deaths into 60-20-20 train-validate-test sets
- Fit different models, training and tuning on the train-validate
- **Select the best model based on hold-out performance**



Good out-of-sample model performance

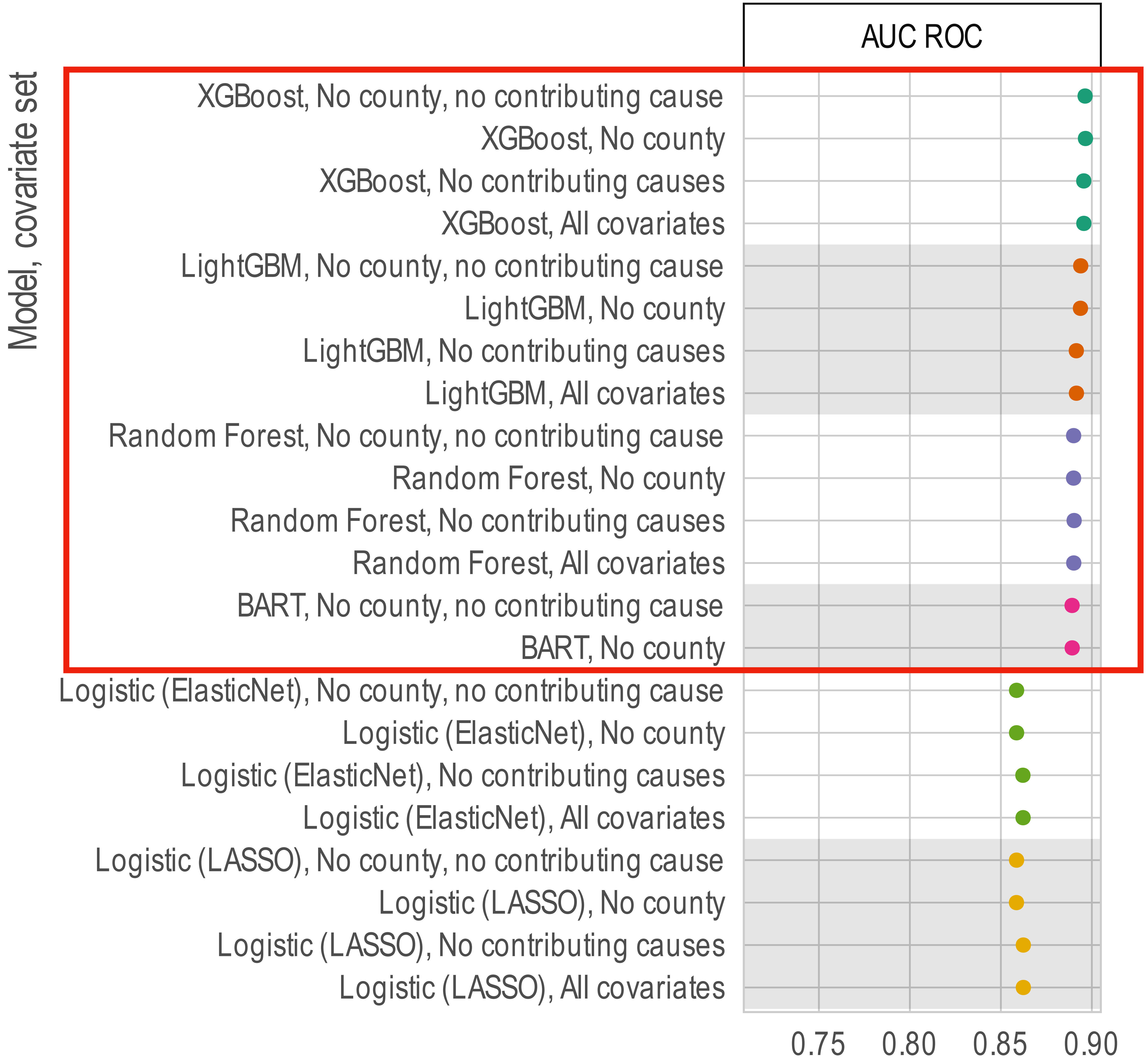
AUC ROC = 0.896

Choice of covariate set does not matter



AUC ROC > 0.890

All tree-based methods perform well



Consistent across performance metrics

Step 2: Predict out of hospital deaths

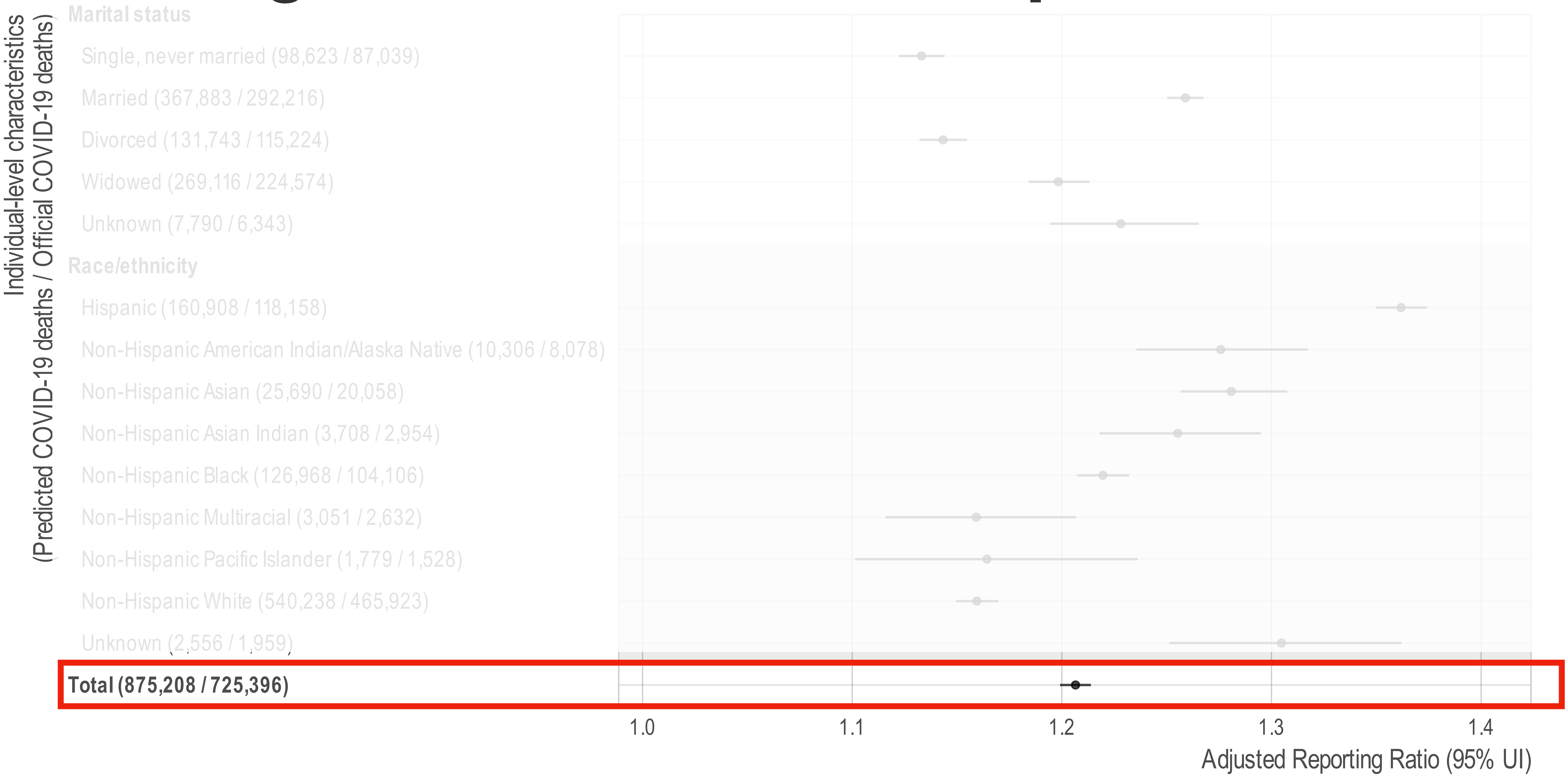
- Split in-hospital deaths into 60-20-20 train-validate-test sets
- Fit different models, training and tuning on the train-validate
- Select the best model based on performance on the test set
- Use SHAP values to understand what the model is doing
- Refit main model with all in-hospital death data
- **Use refitted model to classify out-of-hospital deaths**

Step 3: Estimate over-/under-reporting

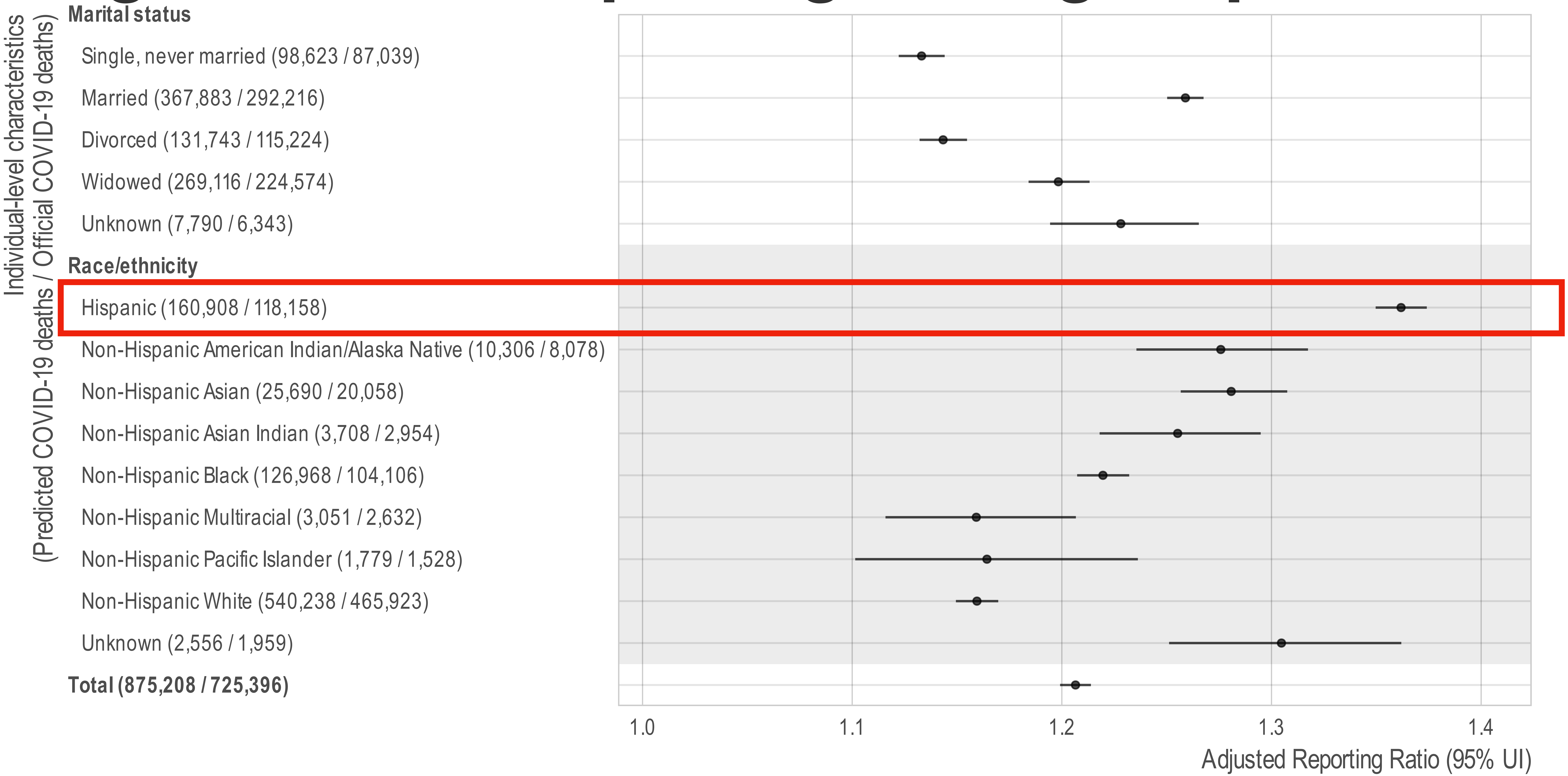
$$ARR = \frac{D_{in} + \widehat{D_{out}}}{D_{in} + D_{out}}$$

- Assumes in-hospital deaths are accurate (Key Assumption #1)
- Accounts for differences in proportion of in- vs out-hospital deaths
- Bootstrap (B=10,000) for 95% confidence intervals

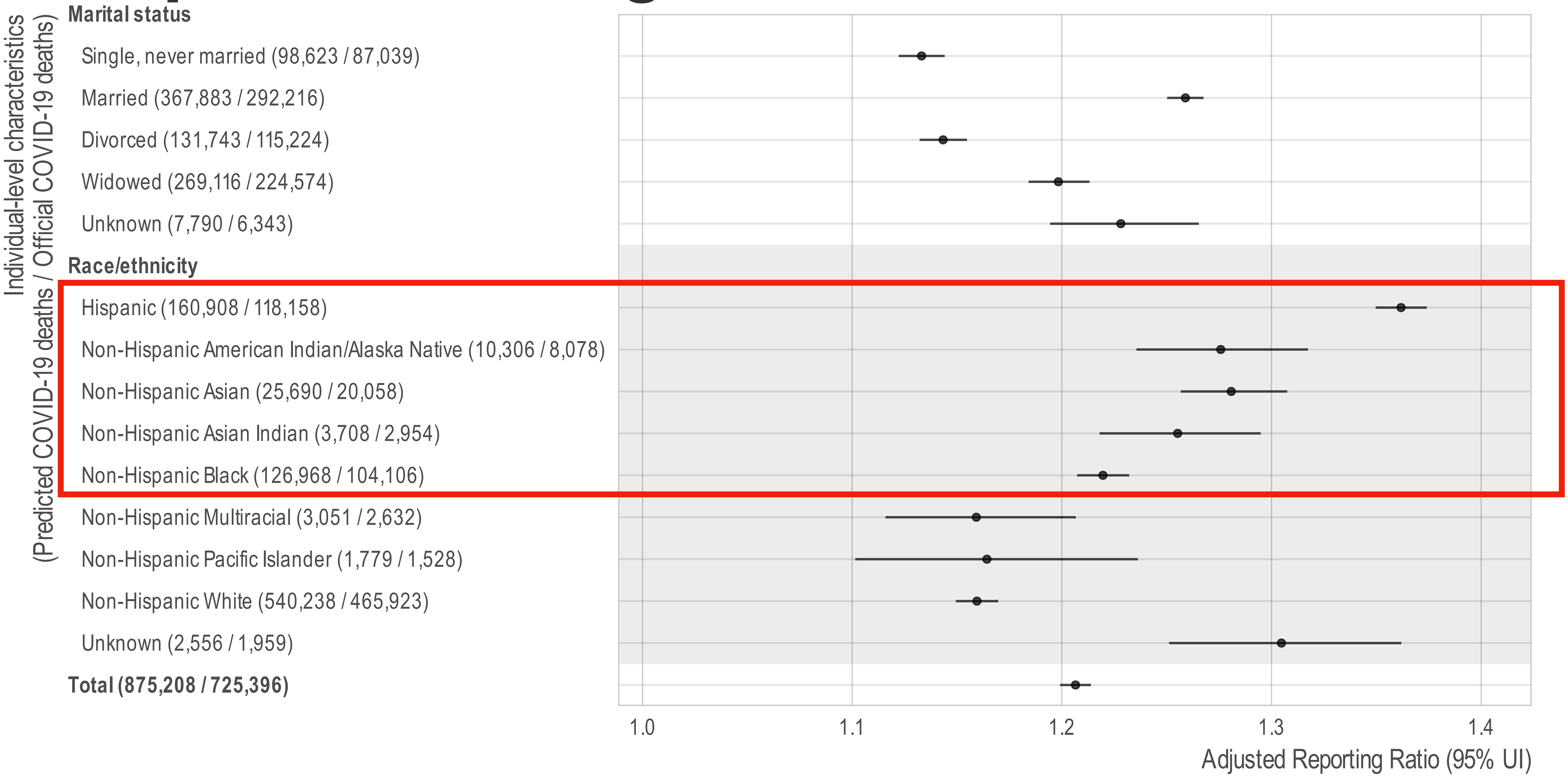
20% higher COVID-19 than reported



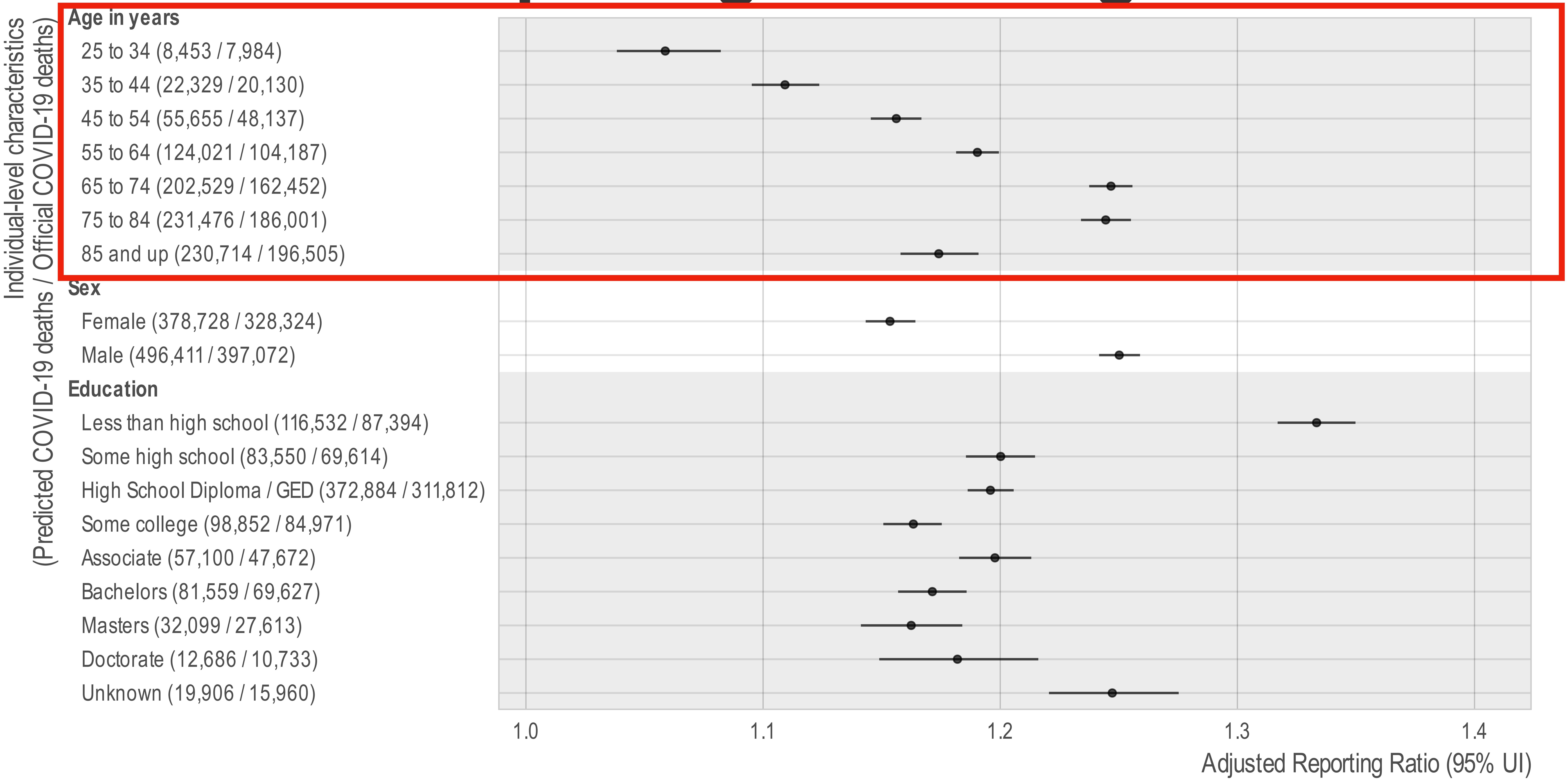
Highest under-reporting among Hispanic



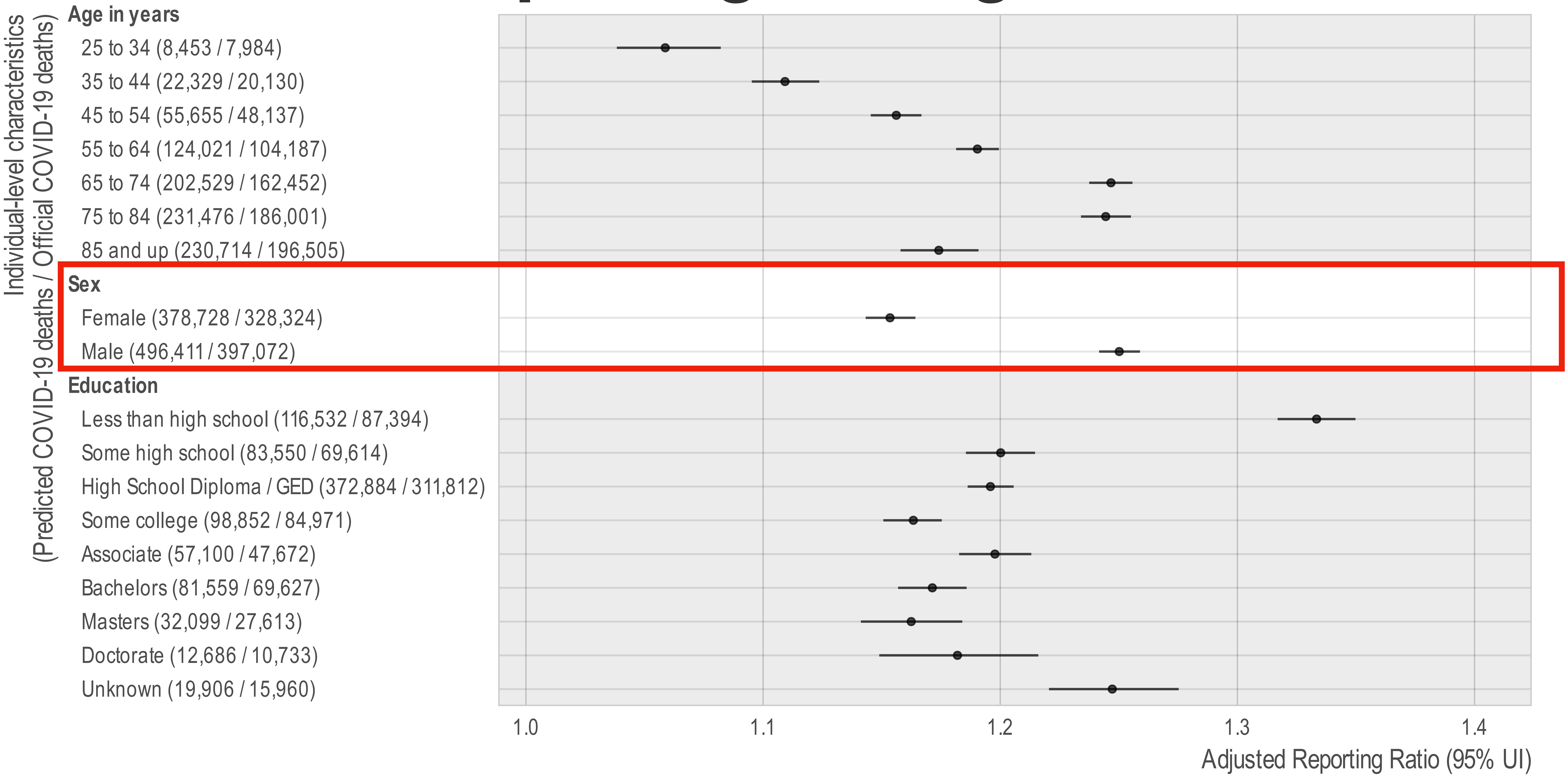
Inequities even higher than documented



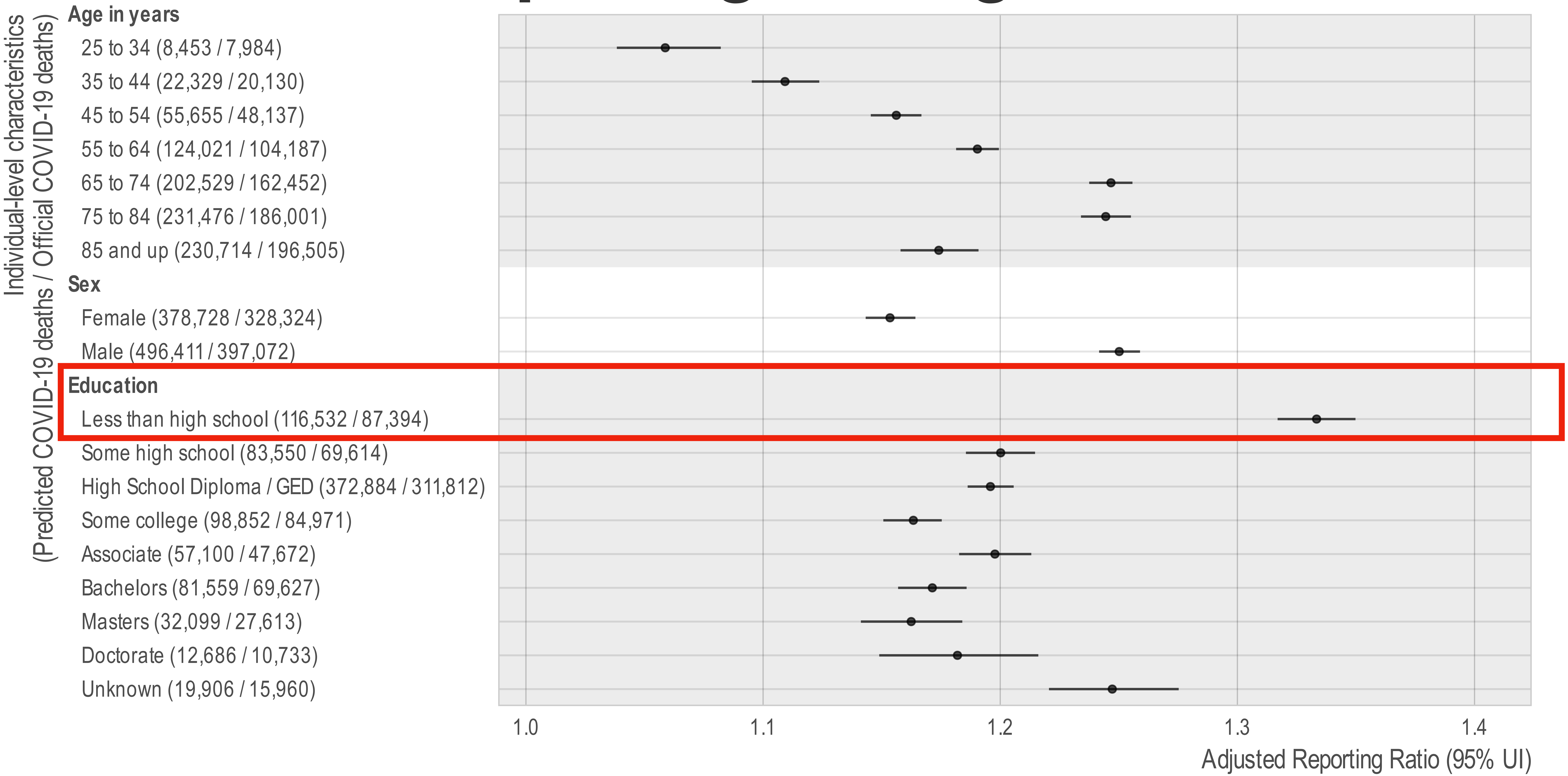
More under-reporting at older ages



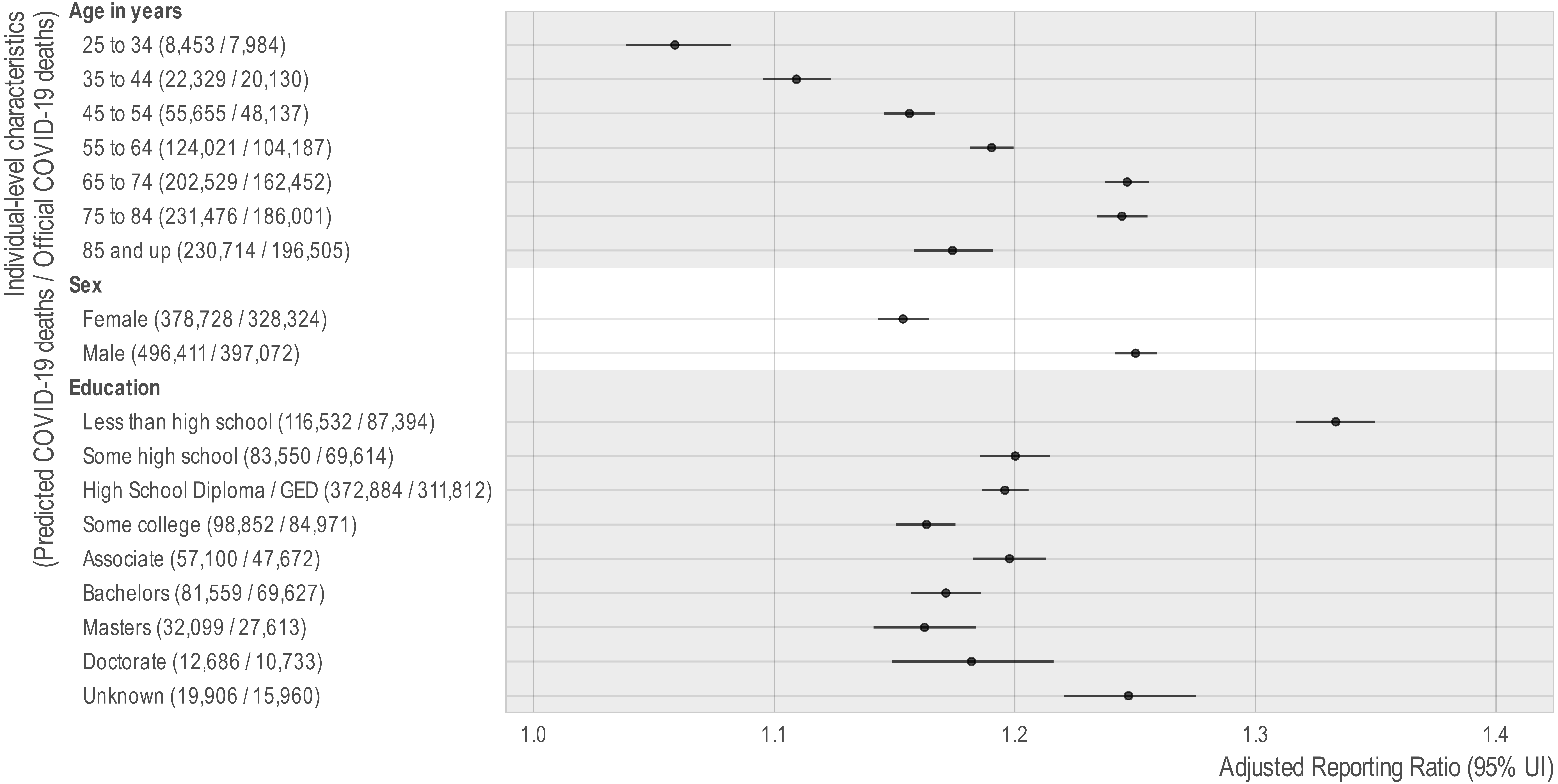
More under-reporting among males



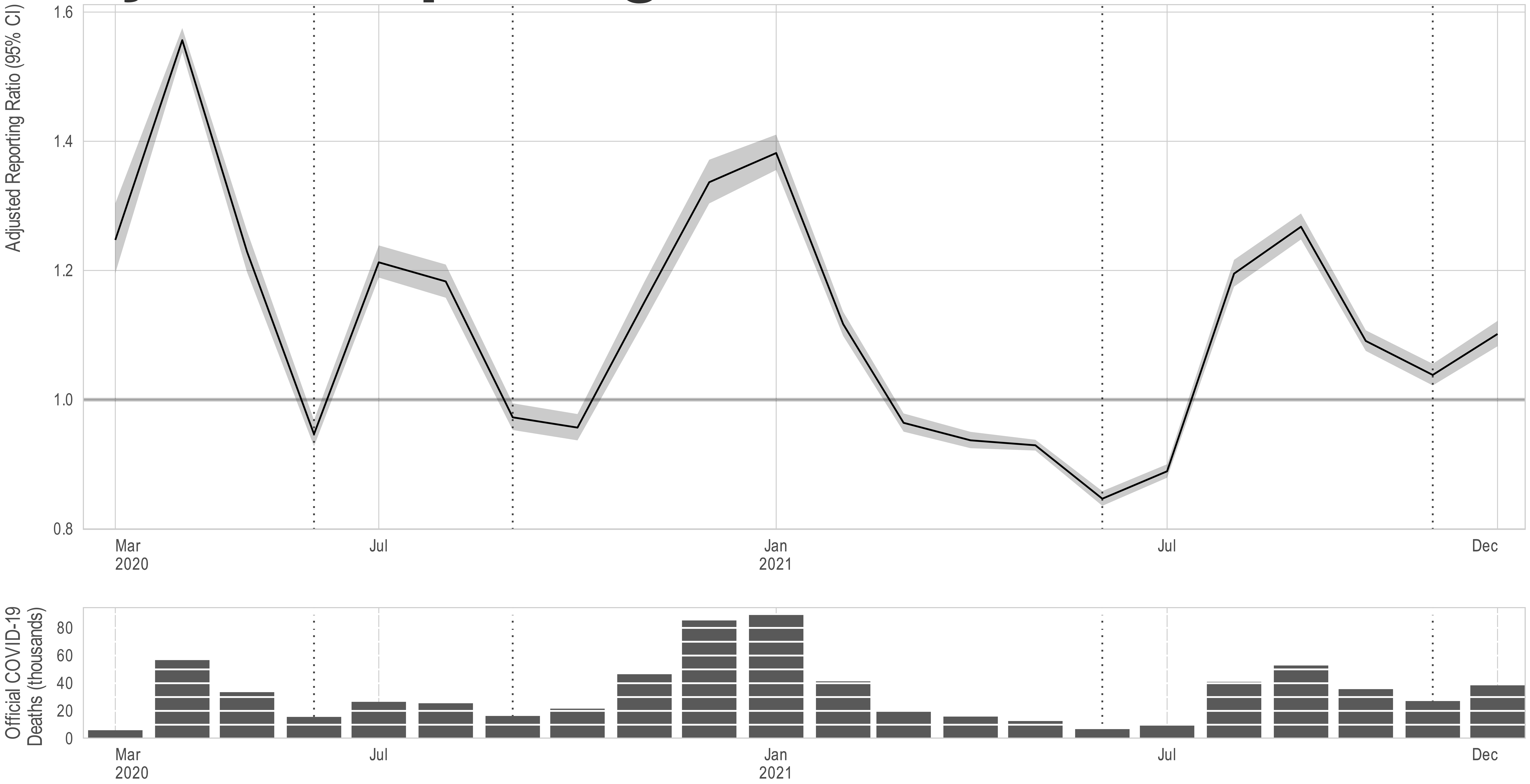
More under-reporting among < HS/GED



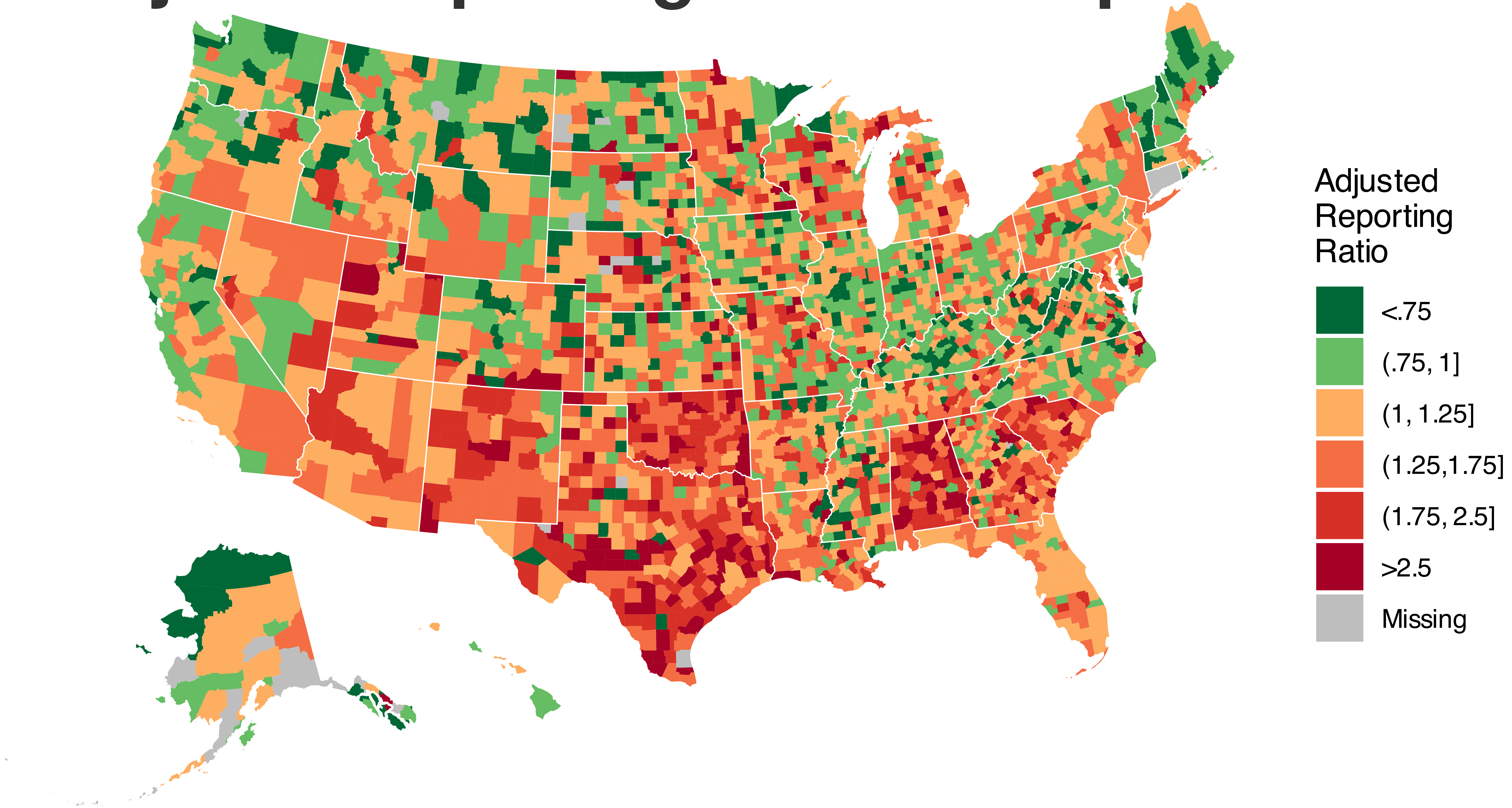
All groups are under-reported



Adjusted reporting ratio over time



Adjusted reporting ratio over space



Key Results

1. COVID-19 deaths ~**20% higher than official estimates**.
2. Race-specific results imply even **greater inequities.**
3. **Male, older,** and **less educated** more under-reported.
4. Variation across both time and space but, in general, COVID-19 was **underreported nearly everywhere** and most of the period of interest.
5. **Robust** to model or covariate choice.

Thank you!

We would like to acknowledge the important intellectual contributions of:

- Richard Li (UCSC)
- Elizabeth Wrigley-Field (UMN)
- The PEARL Group (UCSF)

Limitations

1. All results depend on the two key assumptions!

Limitations

1. All results depend on the two key assumptions!
2. Key Assumption 1 (widespread testing) likely true overall but may have important geographic and subgroup heterogeneity.
3. Key Assumption 2 (model can be applied to out of hospital deaths) is impossible to evaluate without additional data.
4. Future work on accurately coded, out-of-hospital deaths may provide better adjustment (e.g., transfer learning).
5. Normal limitations of using death certificates.

Additional Slides

Death reporting in the US varies widely

Death reporting in the US varies widely

1. **Assigning cause of death is hard.**
   Subjective, complex, and resource-intensive with varying levels of accuracy.

- National Academies of Sciences, Engineering, and Medicine, Strengthening Forensic Science in the United States: A Path Forward (The National Academies Press, 2009).
- Institute of Medicine, Medicolegal Death Investigation System: Workshop Summary (The National Academies Press, Washington, DC, 2003; https://www.nap.edu/catalog/10792/medicolegal-death-investigation-system-workshop-summary).

Death reporting in the US varies widely

1. **Assigning cause of death is hard.**
2. **Death investigators may have personal biases.**
   Conflicts of interest or political roles (e.g., Sheriff-Coroners)

- D. Bergin, R. Keller, Uncounted: Inaccurate death certificates across the country hide the true toll of COVID-19, Missouri Independent (2021). https://missouriindependent.com/2021/12/22/uncounted-inaccurate-death-certificates-across-the-country-hide-the-true-toll-of-covid-19/.
- J. M. Feldman, S. Gruskin, B. A. Coull, N. Krieger, Quantifying underreporting of law-enforcement-related deaths in United States vital statistics and news-media-based data sources: A capture-recapture analysis. PLoS Med. 14, e1002399 (2017).
- S. Gaby, D. Cunningham, H. Lee, G. Ward, A. N. Jackson, Exculpating Injustice: Coroner Constructions of White Innocence in the Postbellum South. Socius 7, 2378023120983647 (2021).

Death reporting in the US varies widely

1. **Assigning cause of death is hard.**
2. **Death investigators may have personal biases.**
3. **Decedent- and family-level factors**
   Attitudes and stigma about COVID-19, testing behaviors

- A. Cakanlar, R. Trudel, K. White, Political Ideology and the Perceived Impact of Coronavirus Prevention Behaviors for the Self and Others. Journal of the Association for Consumer Research 7, 36–44 (2022).
- S. L. Choi, P. Martin, J. Cho, Y. J. Ryou, M. Heinz, Personality and compliance with COVID-19 protective measures among older Americans: Moderating effects of age, gender, and race/ethnicity. Pers. Individ. Dif. 189, 111499 (2022).

Death reporting in the US varies widely

1. **Assigning cause of death is hard.**
2. **Death investigators may have personal biases.**
3. **Decedent- and family-level factors**
4. **Community-level factors**
   Test availability, distribution of pre-existing conditions

- M. Kanyangarara, V. Daguise, L. Gual-Gonzalez, A. H. Litwin, J. Korte, C. Ross, M. S. Nolan, SC STRONG Project Team, COVID-19 Testing Practices, Preventive Behaviors, and Factors Associated With Test Positivity: Population-Based Statewide Survey Study. JMIR Public Health Surveill 9, e34579 (2023).
- D. Dada, J. N. Djiometio, S. M. McFadden, J. Demeke, D. Vlahov, L. Wilton, M. Wang, L. E. Nelson, Strategies That Promote Equity in COVID-19 Vaccine Uptake for Black Communities: a Review. J. Urban Health 99, 15–27 (2022).

Death reporting in the US varies widely

1. **Assigning cause of death is hard.**
2. **Death investigators may have personal biases.**
3. **Decedent- and family-level factors**
4. **Community-level factors**
5. **Health systems factors**
   Funding for death investigation offices, local policies

- A. C. Stokes, D. J. Lundberg, J. Bor, I. T. Elo, K. Hempstead, S. H. Preston, Association of Health Care Factors With Excess Deaths Not Assigned to COVID-19 in the US. JAMA Netw Open 4, e2125287 (2021).
- A. B. Le, E. G. Brooks, L. A. McNulty, J. R. Gill, J. J. Herstein, J. Rios, S. J. Patlovich, K. C. Jelden, K. K. Schmid, J. J. Lowe, S. G. Gibbs, U.S. Medical Examiner/Coroner capability to handle highly infectious decedents. Forensic Sci. Med. Pathol. 15, 31–40 (2019).
- L. Ruiz, B. M. Posey, M.-A. Neuilly, M. K. Stohr, C. Hemmens, Certifying Death in the United States. J. Forensic Sci. 63, 1138–1145 (2018).

Death reporting in the US varies widely

1. **Assigning cause of death is hard.**
2. **Death investigators may have personal biases.**
3. **Decedent- and family-level factors**
4. **Community-level factors**
5. **Health systems factors**
6. **Structural factors**

   Racism, classism, ableism, etc.

- D. J. Lundberg, J. A. Chen, Structural ableism in public health and healthcare: a definition and conceptual framework. The Lancet Regional Health – Americas 30 (2024).
- M. A. Garcia, P. A. Homan, C. García, T. H. Brown, The Color of COVID-19: Structural Racism and the Disproportionate Impact of the Pandemic on Older Black and Latinx Adults. J. Gerontol. B Psychol. Sci. Soc. Sci. 76, e75–e80 (2021).
- L. E. Egede, R. J. Walker, Structural Racism, Social Risk Factors, and Covid-19 - A Dangerous Convergence for Black Americans. N. Engl. J. Med. 383, e77 (2020).
- E. E. Andrews, K. B. Ayers, K. S. Brown, D. S. Dunn, C. R. Pilarski, No body is expendable: Medical rationing and disability justice during the COVID-19 pandemic. Am. Psychol. 76, 451–461 (2021).
- L. V. Estrada, J. L. Levasseur, A. Maxim, G. A. Benavidez, K. M. Pollack Porter, Structural Racism, Place, and COVID-19: A Narrative Review Describing How We Prepare for an Endemic COVID-19 Future. Health Equity 6, 356–366 (2022).

Accurate COVID-19 death count is critical

Accurate COVID-19 death count is critical

1. **Responsive and equitable health system**
   Understanding inequalities and targeting interventions

Accurate COVID-19 death count is critical

1. **Responsive and equitable health system**
2. **Personal and institutional decision-making**
   Individuals, states, counties, hospitals, schools, or jails made decisions based on these data

Accurate COVID-19 death count is critical

1. **Responsive and equitable health system**
2. **Personal and institutional decision-making**
3. **Epidemiological understanding**
   Transmission dynamics, model parameterization, disease characterization

Accurate COVID-19 death count is critical

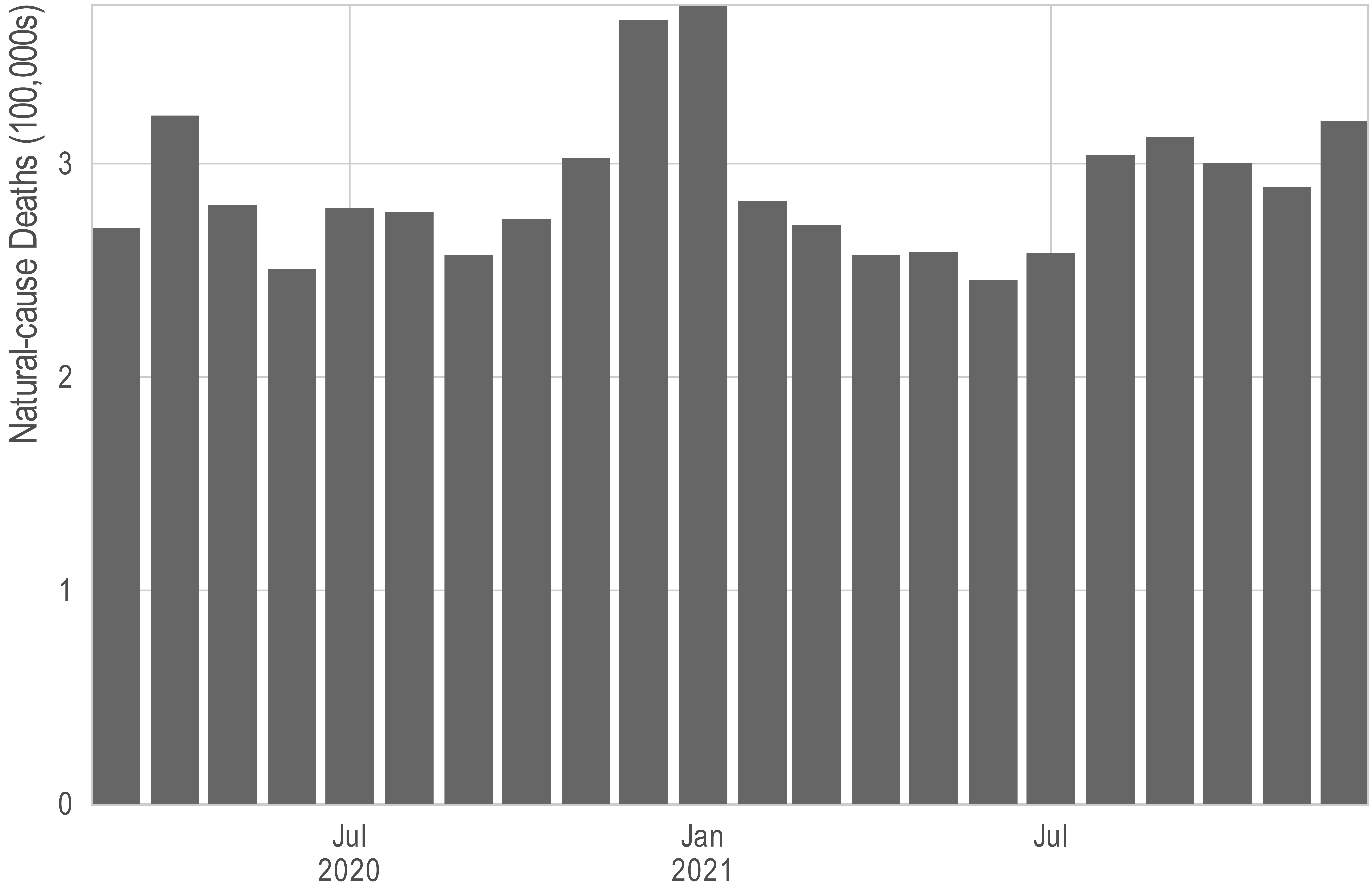
1. **Responsive and equitable health system**
2. **Personal and institutional decision-making**
3. **Epidemiological understanding**
4. **Financially relevant**

   FEMA funeral assistance program (up to $9,000)

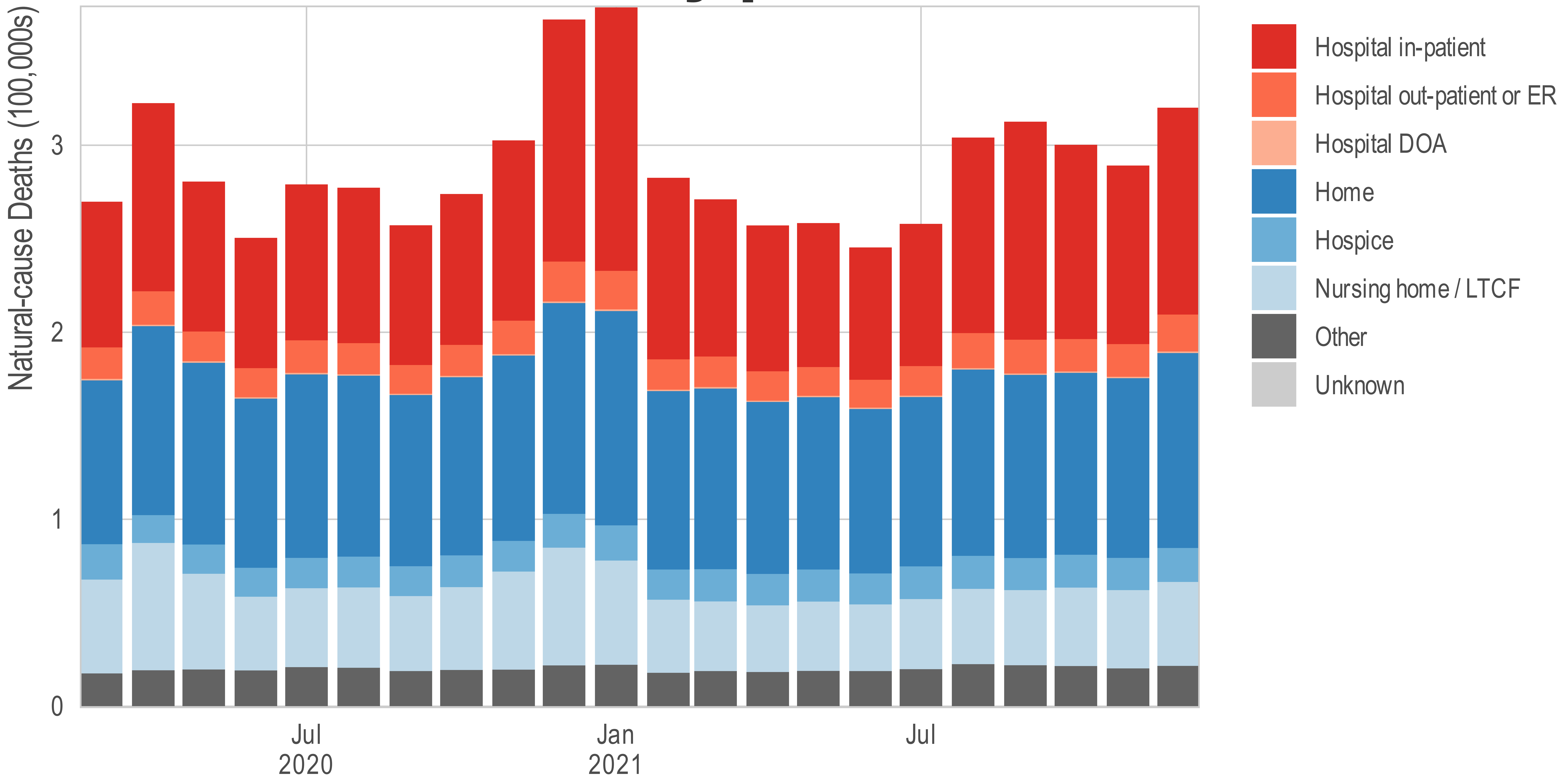
Accurate COVID-19 death count is critical

1. Responsive and equitable health system
2. Personal and institutional decision-making
3. Epidemiological understanding
4. Financially relevant
5. **Just fundamentally important**
   Family members should know what their loved ones died from

Deaths in the US



Deaths in the US by place of death



Create covariate sets

- Individual-level death certificates
- County-level characteristics (static, pre-2020)
- COVID-19 covariates (time-varying)
- Total N = 5.7 million deaths (> 25 y; natural causes only)
- **Create 4 covariate sets**:
  1. County-fixed effects only (P = 3193)
  2. Contributing causes fixed effects only (P = 113)
  3. Naive (everything + fixed effects)
  4. No fixed effects

Data sources

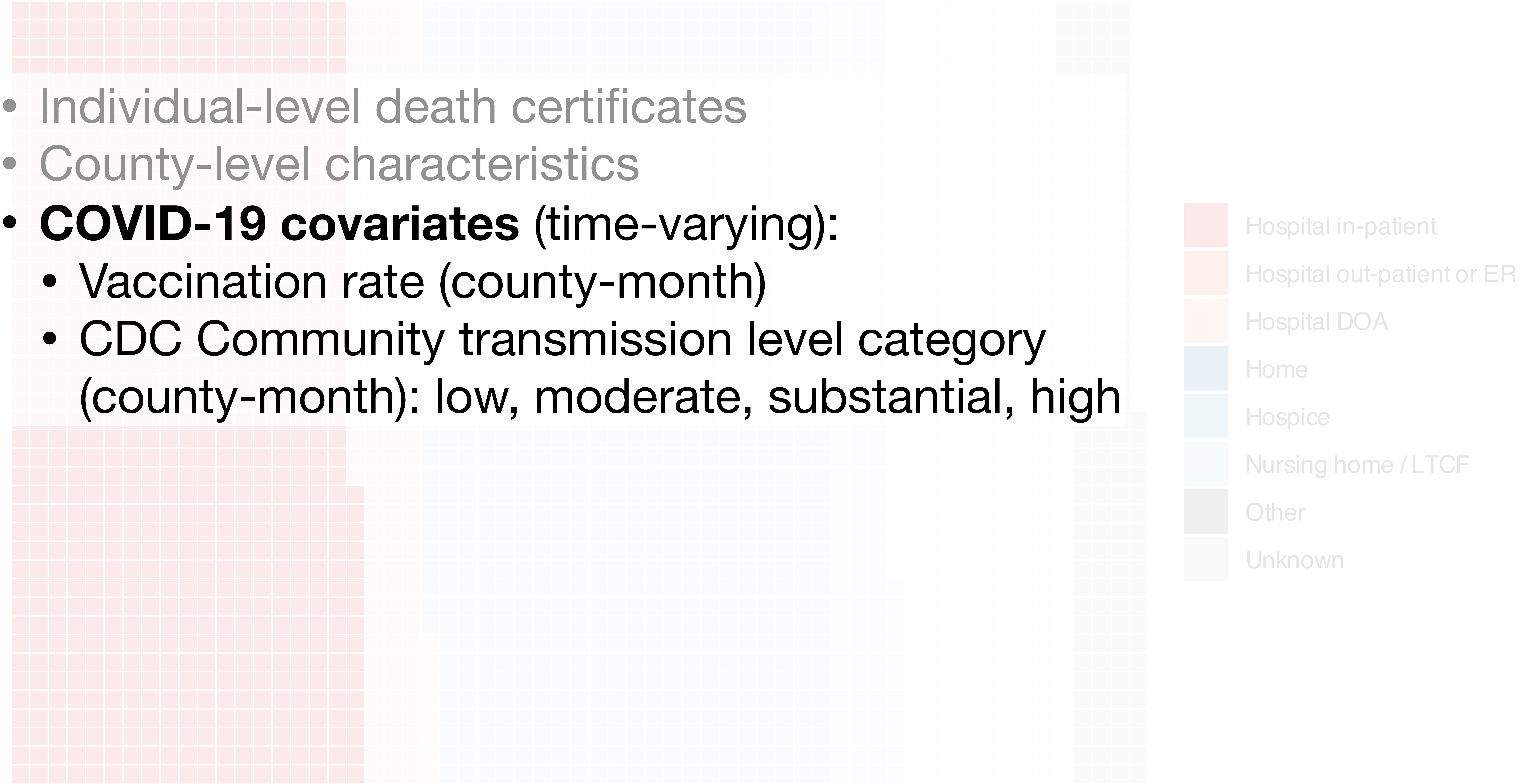
- **Individual-level death certificates** (static):
  - Sociodemographics (age, sex, marital status, race, ethnicity, state of birth, education)
  - Contributing causes of death (pneumonia, Type 2 Diabetes, COPD)
  - Reporting information (day of week, county)

Data sources

- Individual-level death certificates
- **County-level characteristics** (static, pre-2020):
    - Urban/rural code
    - Median household income
    - Smoking rate
    - Obesity rate
    - Racial/composition
    - Residential segregation
    - Homeownership
    - Area educational attainment

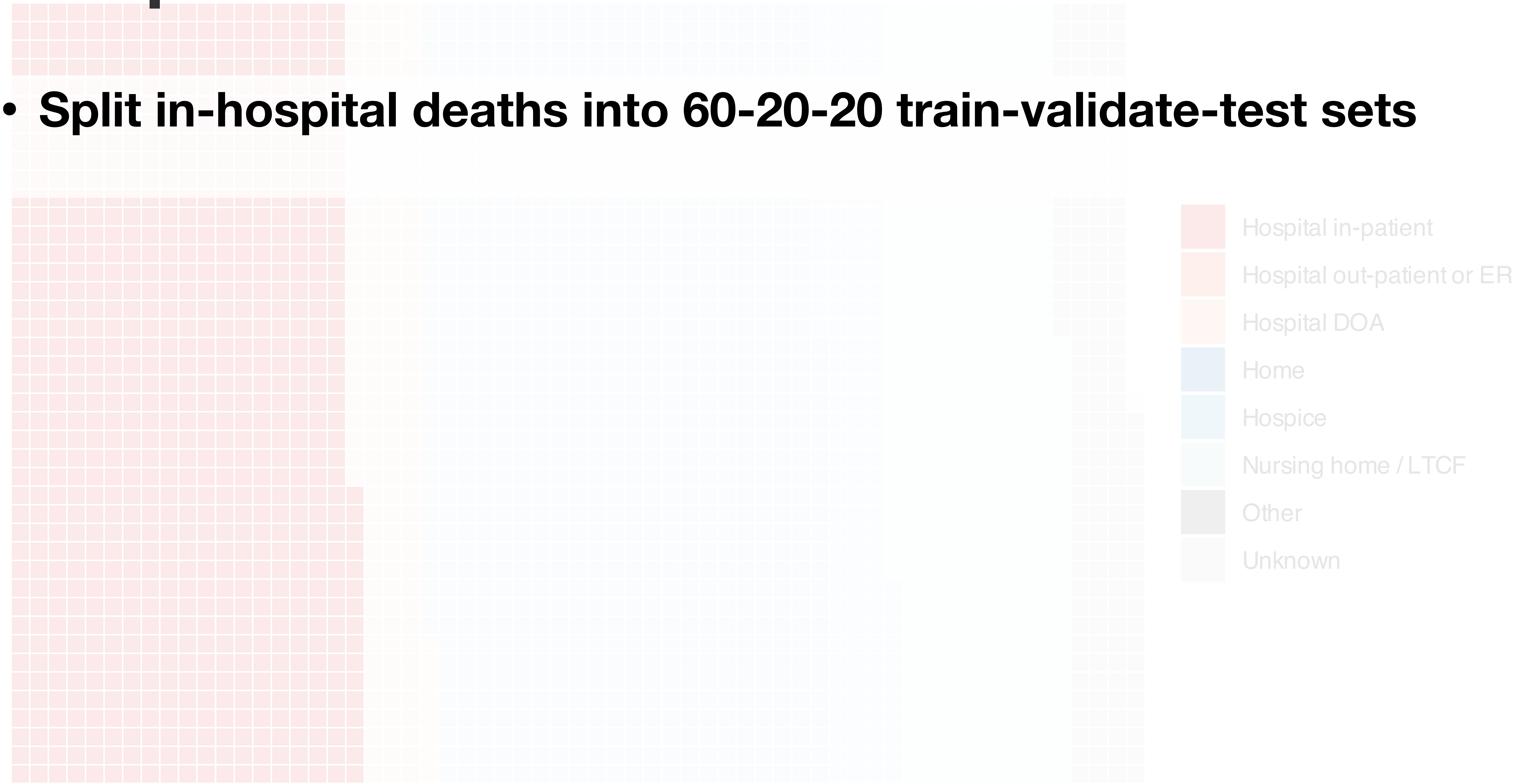
Data sources

- Individual-level death certificates
- County-level characteristics
- **COVID-19 covariates** (time-varying):
  - Vaccination rate (county-month)
  - CDC Community transmission level category (county-month): low, moderate, substantial, high



Step 1: Train a classifier

- **Split in-hospital deaths into 60-20-20 train-validate-test sets**



Step 2: Predict out of hospital deaths

- Split in-hospital deaths into 60-20-20 train-validate-test sets
- Fit different models, training and tuning on the train-validate
- Select the best model based on performance on the test set
- Use SHAP values to understand what the model is doing
- **Refit main model with all in-hospital death data**

Pick a model, any model

Example of a low risk individual

Example of a low risk individual

Example of a low risk individual

Example of a low risk individual

Example of a low risk individual

Example of a low risk individual

Example of a high risk individual

Low vaccination rate increases risk

Top 10 predictors (by SHAP value)

Higher transmission increases risk

Socioeconomic vulnerability increases risk

No comorbidities decreases risk

Final prediction for this individual

Example of a low risk individual

Example of a high risk individual

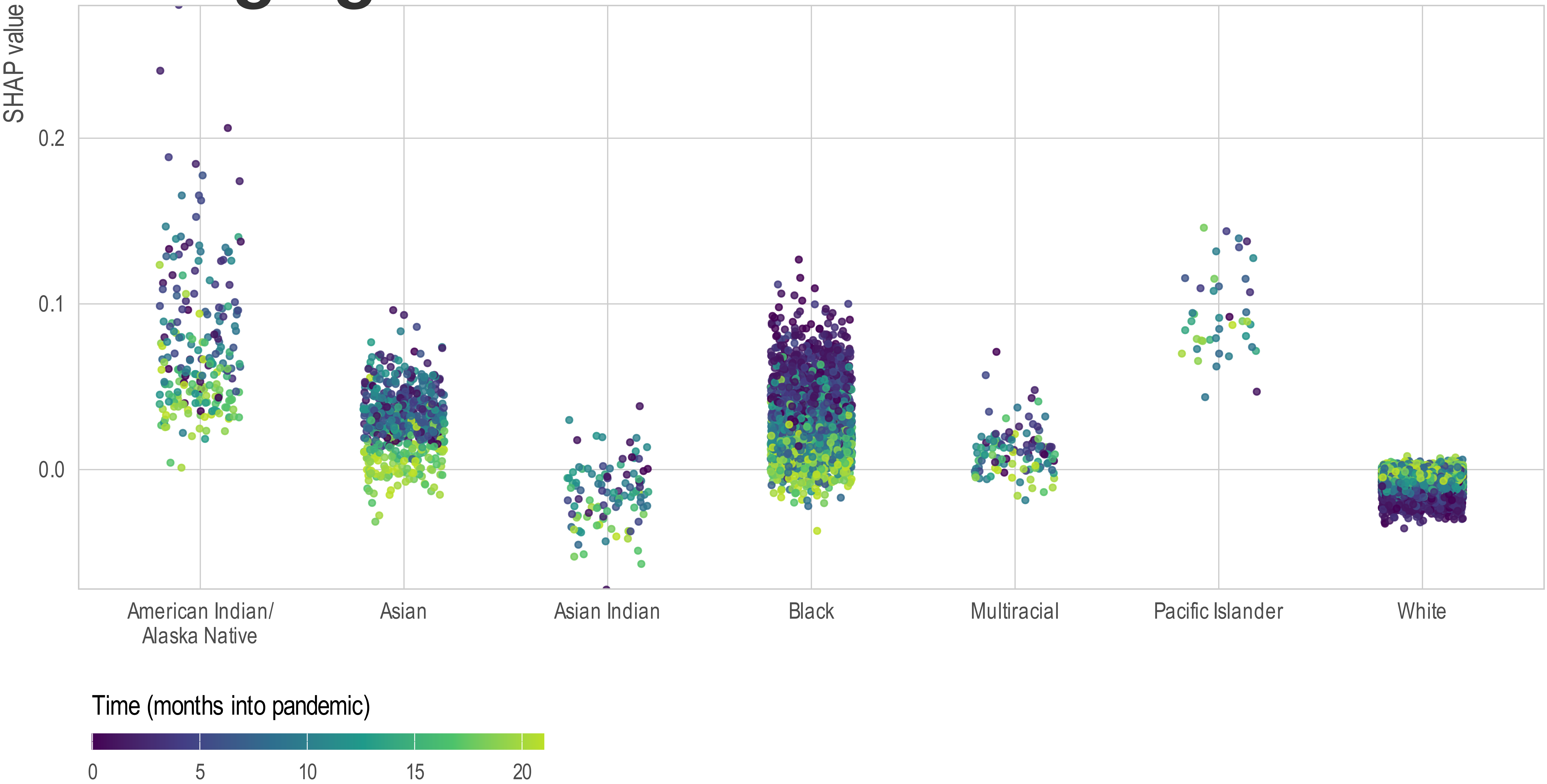
Example of a high risk individual

SHAP values of top 15 features

SHAP values of top 15 features

SHAP values of top 15 features

Changing contribution of race over time



Step 3: Estimate over-/under-reporting

$$ARR = \frac{D_{in} + \widehat{D_{out}}}{D_{in} + D_{out}}$$

Step 3: Estimate over-/under-reporting

$$\boxed{ARR} = \frac{D_{in} + \widehat{D_{out}}}{D_{in} + D_{out}}$$

- Adjusted Reporting Ratio (ARR)

Step 3: Estimate over-/under-reporting

$$\boxed{ARR} = \frac{\boxed{D_{in}} + \widehat{D_{out}}}{\boxed{D_{in}} + D_{out}}$$

- Adjusted Reporting Ratio (ARR)
- Number of observed in-hospital deaths

Step 3: Estimate over-/under-reporting

$$\boxed{ARR} = \frac{\boxed{D_{in}} + \widehat{D_{out}}}{\boxed{D_{in}} + \boxed{D_{out}}}$$

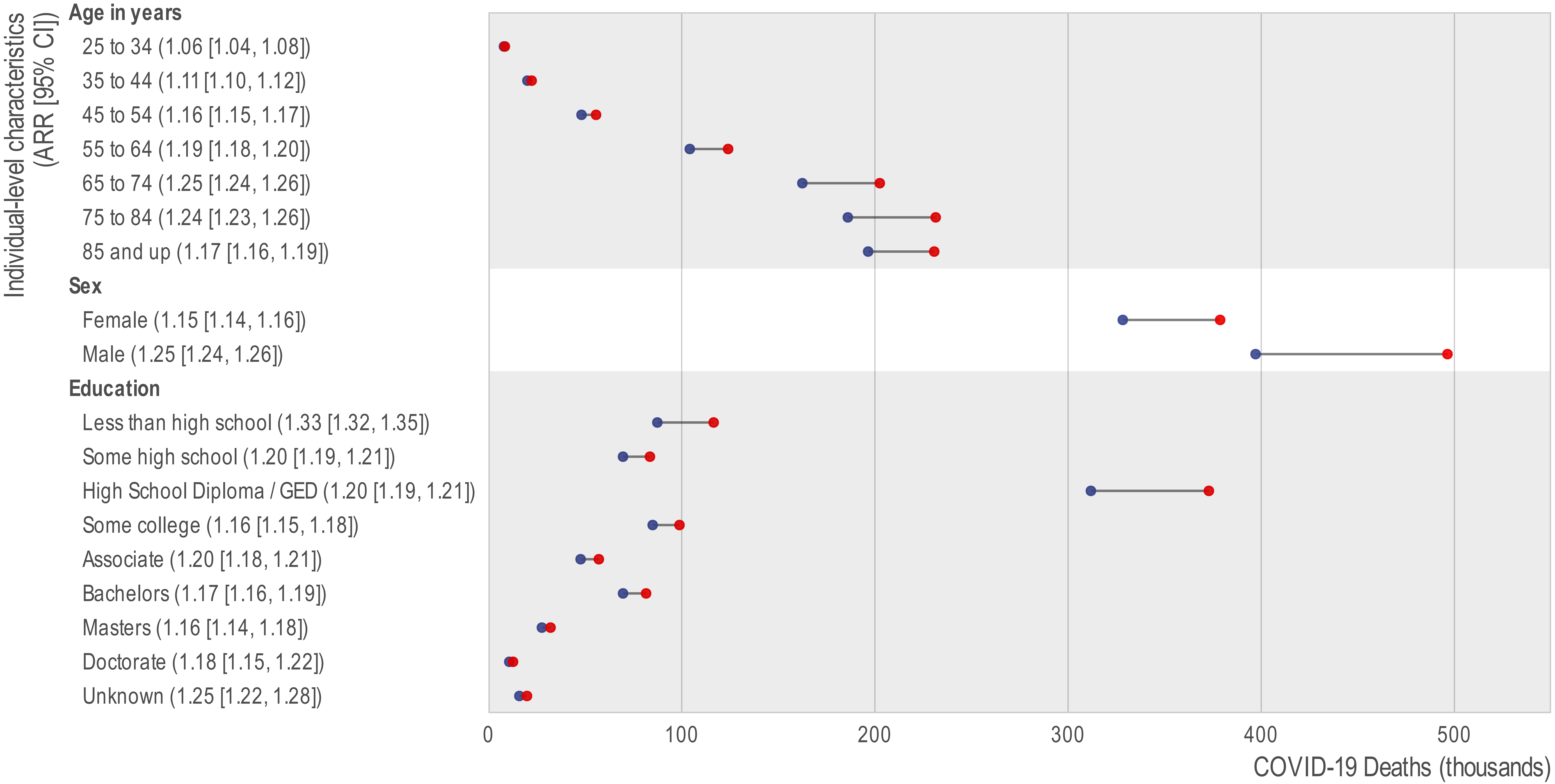
- Adjusted Reporting Ratio (ARR)
- Number of observed in-hospital deaths
- Number of observed out-of-hospital deaths

Step 3: Estimate over-/under-reporting

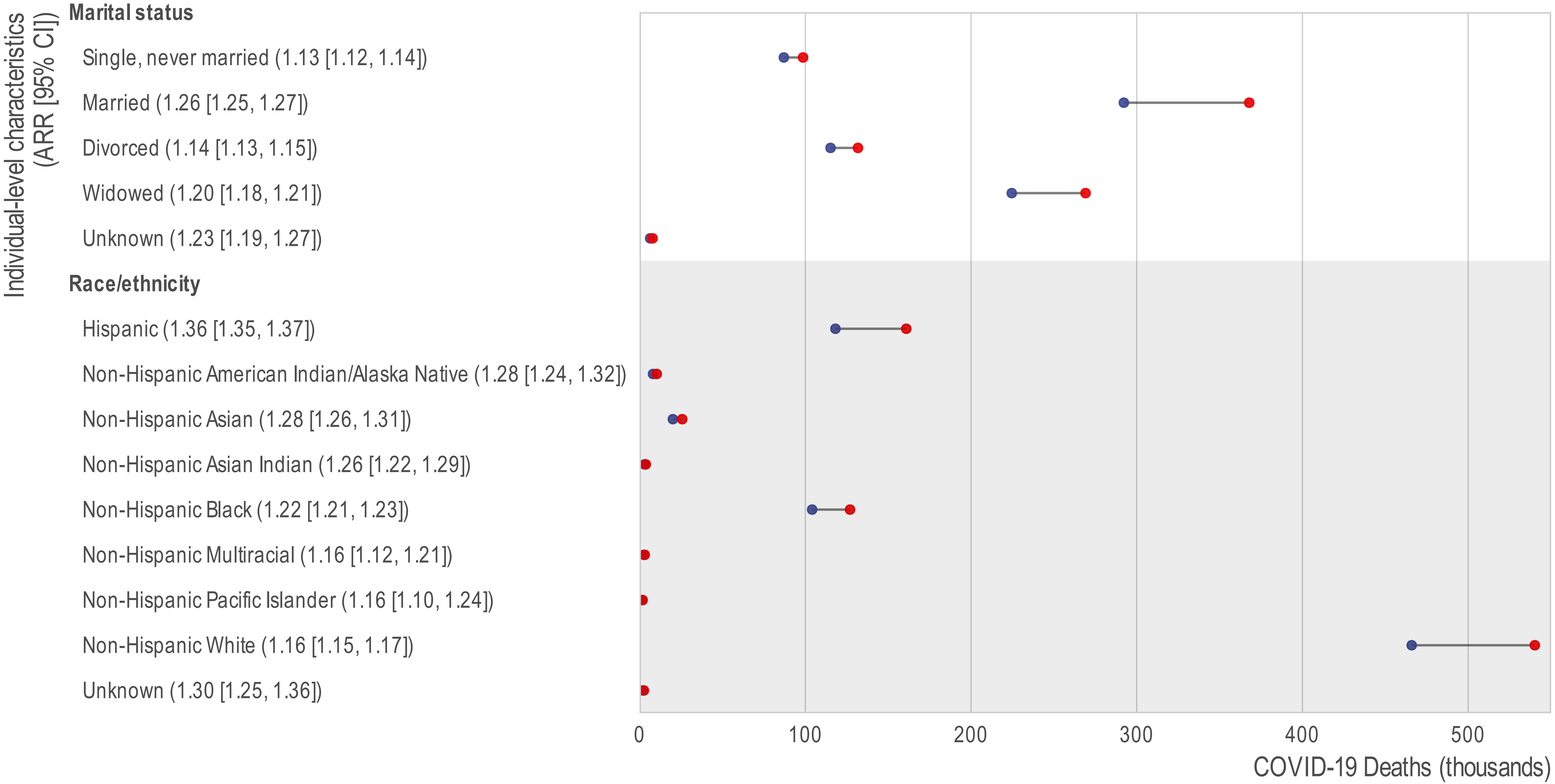
$$\boxed{ARR} = \frac{\boxed{D_{in}} + \boxed{\widehat{D_{out}}}}{\boxed{D_{in}} + \boxed{D_{out}}}$$

- Adjusted Reporting Ratio (ARR)
- Number of observed in-hospital deaths
- Number of observed out-of-hospital deaths
- Number of **predicted** out-of-hospital deaths

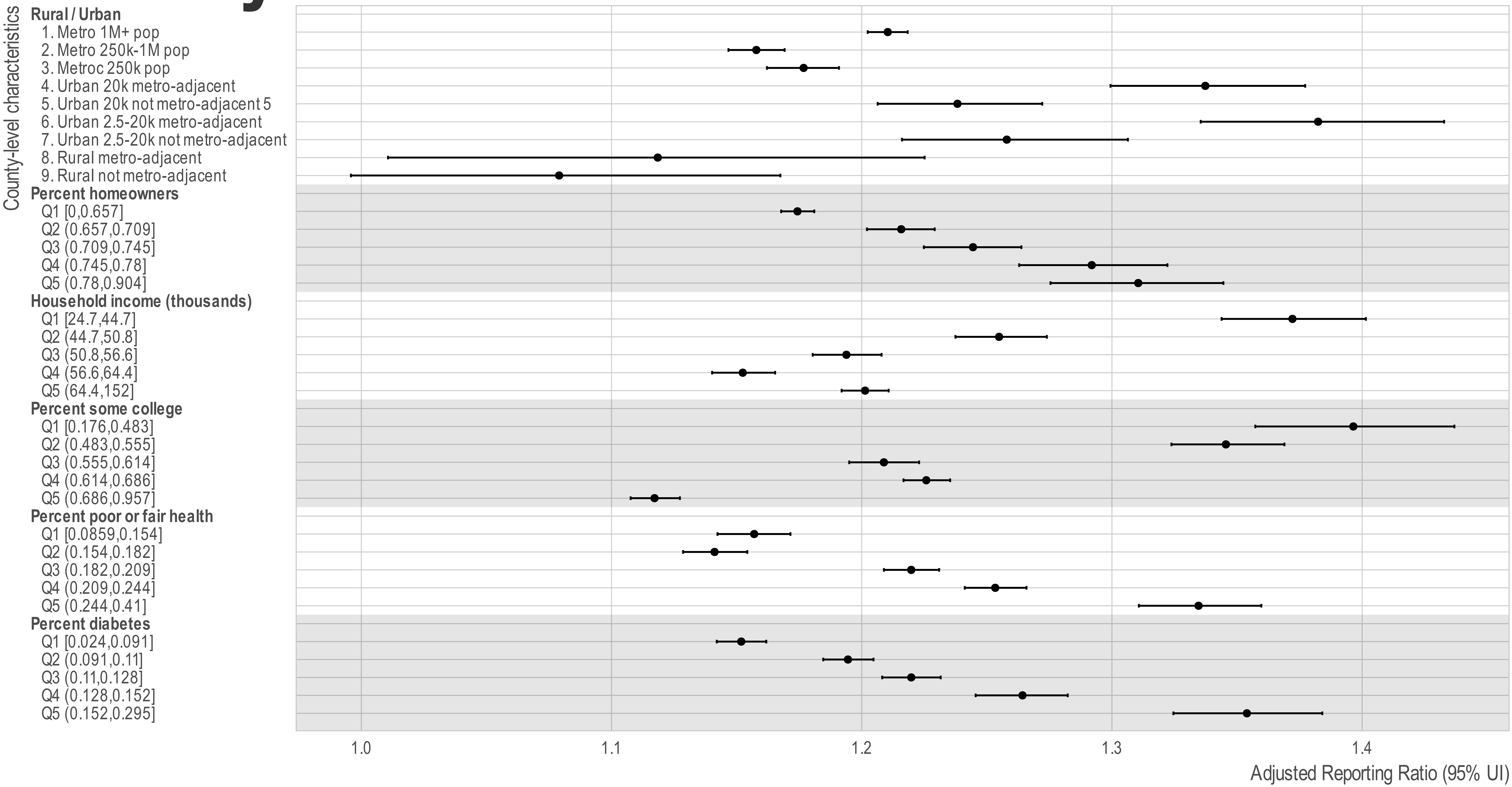
Individual characteristics



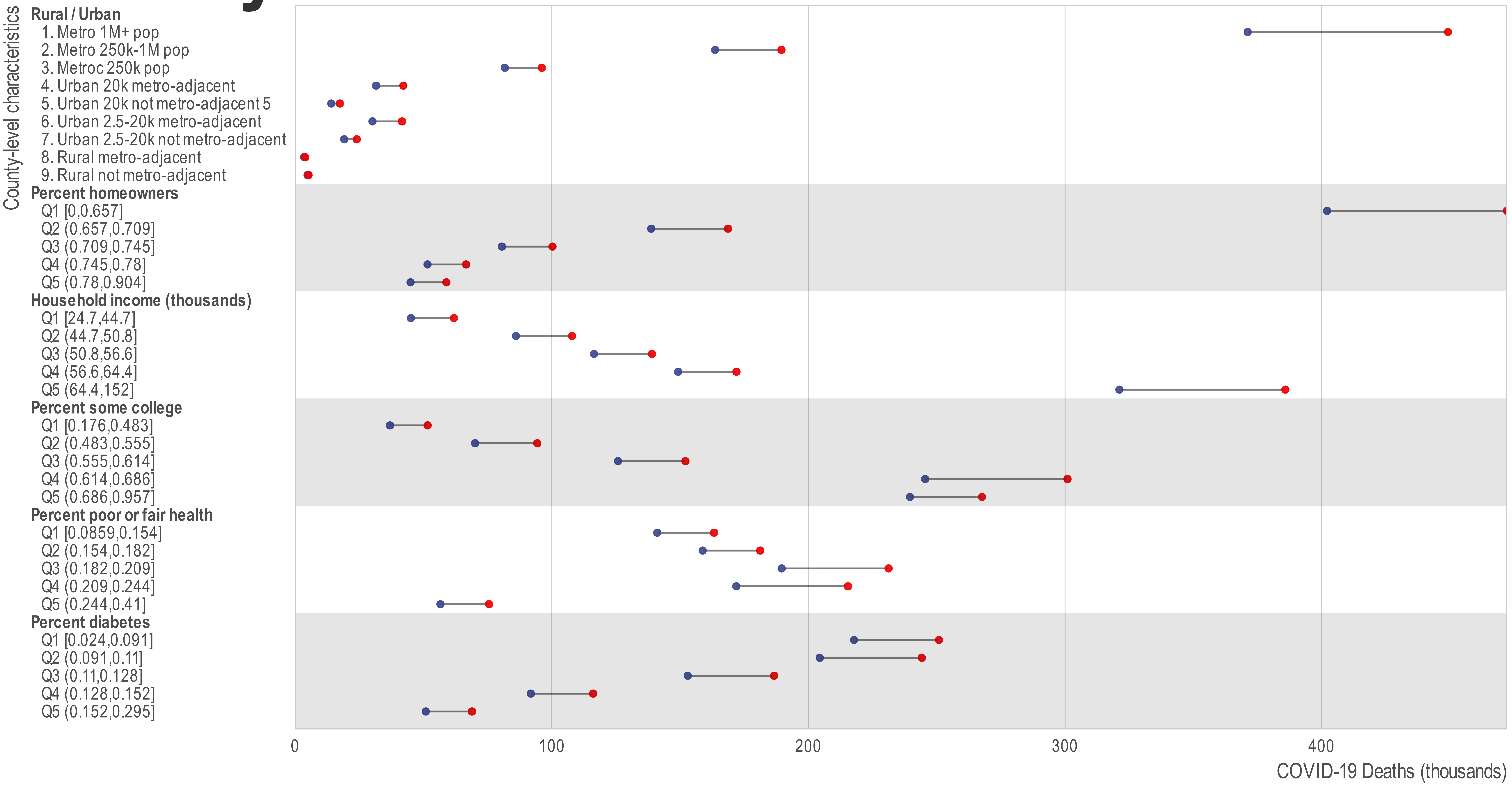
Individual characteristics



County characteristics



County characteristics



Observed vs predicted by county