Combining Data from Multiple Sources: Performance of different classes of estimators from Monte Carlo simulations

October 23, 2024 :: FCSM, Washington DC

Stas Kolenikov, Soubhik Barari, David Dutwin (NORC)
Katherine Irimata, James Dahlhamer (NCHS)

NORC at the University of Chicago

Views expressed in this presentation are those of the authors, and do not represent CDC, NCHS or NORC.

NCHS Team

Paul Scanlon

James Dahlhamer

Katherine Irimata

NORC Literature Review Team

David Dutwin

Ipek Bilgen

Stas Kolenikov

Michael Yang

Chien-Min Huang

Margrethe Montgomery

NORC R code team

Stas Kolenikov

Soubhik Barari

Núria Adell Raventos

Emerson Berry

Chien-Min Huang

Jiazhi Yang

Matt Gunther

Sabrina Sedovic

The Rapid Surveys System (RSS) is conducted by the Centers for Disease Control and Prevention's (CDC's) National Center for Health Statistics (NCHS), under contract 47QRAA20D001M with NORC with funding from CDC/NCHS and other cosponsors. The findings and conclusions in this presentation are those of the authors and do not necessarily represent the official position of CDC/NCHS or other funding partners.

Why combine data?

NORC at the University of Chicago

(How) Can NCHS use online probability panels to supplement its core surveys?

Methodology questions

- Panel recruitment
- Response rates
- Attrition
- Mode effects

Statistics questions

- Weighting
- Methods to combine estimates
- Methods to produce estimates on combined data

Agenda

01 Why combine data?

02 Simulation goals

03 Estimators and scenarios

04 Expectations

05 Results

06 Further work

Simulation goals

NCHS program of research into online panels

Identify the most promising methods, pit them against each other.

Methods

- Best methods to combine online panels and "gold standard" data?
  - NORC prepared a literature review for NCHS

Realistic Data

- Public use NHIS data 1997—2018
  - Largely compatible items over time
  - Consistent sampling design (stratified cluster samples)
  - N = 671,696
- Restricted use geography
- Range of outcomes: mental health, BMI/obesity, optometrist visit in past 12 months.

How do we define success?

Simulation Metrics (in lexicographic order of importance):

- Bias
- Standard errors and intervals coverage
- Variance and MSE
- Consistency of performance across scenarios
- Catastrophic biases or catastrophic coverage problems
  - Certain simulation scenarios and/or population subgroups where an estimator *drastically* underperforms.

How do we define success?

Usability Dimensions:

- Total estimation time
- Frequency of runtime compute problems
  - Lack of convergence in iterative procedures (e.g. calibration or REML of mixed models)
  - Time outs (set at 10 minutes per run)
  - Inexplicable crashes
- Positive calibration weights
  - We used linear calibration as the fastest calibration method; negative weights are a distinct possibility
- Manageability of external dependencies
  - `h2o` multi-node *software* cluster (needed even if *hardware* is not a cluster!)

Simulation implementation

Statistical tasks

- Create the finite population
- Draw samples
- Run estimators
- Summarize results

package

Simulation logistics

- Ensure all the required estimators are run on a given sample
  – Not a given when some code is added or debugged later
- Identify frozen runs
- Restart frozen / crashed runs
- Parallel threads

Code development process

- Unit tests: code behaves the way we expect it, changes don't break the past behavior
- Documentation: what is function's inputs and outputs

Estimators and scenarios

Nuts and bolts of the simulation

NORC at the University of Chicago

Four competing classes of estimators

Calibration:

- Demographic variables (age, sex, race/ethnicity, education)
- Demographic + health (from the major survey)
  - Three different standard error approaches
- Lasso prediction

Propensity score adjustment

- Stepwise variable selection
- Kernel weighting
- Attempts to aggregate across outcomes to produce omnibus weights

Small area estimation + calibration (TrueNorth)

- SAE modeling of outcome means within demographic domains with panel effects
- Lasso and stepwise model selection
- Prediction with panel effects removed
- Calibration to demographics + predictions

Double robust:

- Machine learning prediction for both selection and outcome equations

Drawing Monte Carlo samples from the finite population

Major survey sampling

- Stratified cluster sample
  - Respects the original PSUs
  - Gentle unequal probabilities
  - n ~ 4,400-4,500; 160+ PSUs
- Full unit response

Online panel sampling scenarios

- n = 1,000 in each scenario

Benchmark: SRS

Low correctable nonresponse: gently varying function of age

High correctable nonresponse: highly varying function of age, marital status, race and education

Non-correctable nonresponse: function of a secret variable (not used in calibration)

Moderate noncoverage: omit Midwest census region

- Retain 80% population + mild correctable nonresponse

High noncoverage: omit U.S.-born white individuals

- Retain 40% population + mild correctable nonresponse

Expectations

(Any Monte Carlo simulation is only worth doing if you have some baseline expectations)

Expectations

1. All methods work fine in the benchmark SRS scenario
    a. Somewhat more complex methods may have efficiency lower than that of the simplest method (demographic calibration)
2. The more difficult the scenario, the greater the bias of the demographic calibration
3. The more complex estimators will have lower biases than the demographic calibration
4. No expectation of the relative ordering of the complex estimators in terms of…
    a. Bias…
    b. Variance…
    c. Confidence interval coverage

That’s why we simulate!

Simulation results

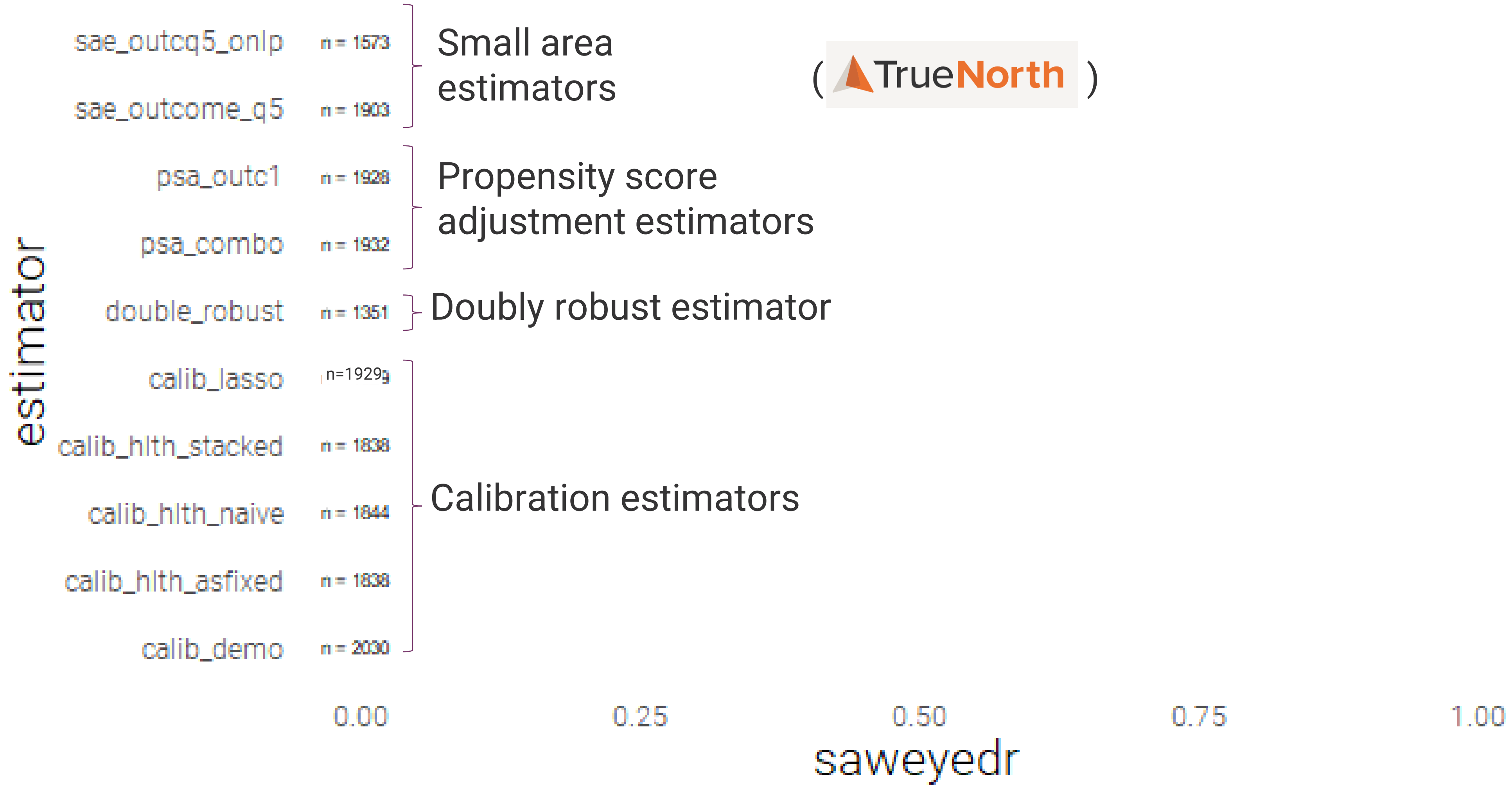
A high level overview of the results

- Several largely ineffective methods: biased across *many* scenarios and outcomes
- Some contextually useful estimators
  - Unbiased for the benchmark scenario
  - Low bias in complex scenarios
  - Decent confidence interval coverage
- ~2,000 boxplots of all estimates for all subgroup breaks, outcomes, scenarios
- We present striking, but representative results illustrating differences between estimators
  - Note: inflated type I error!

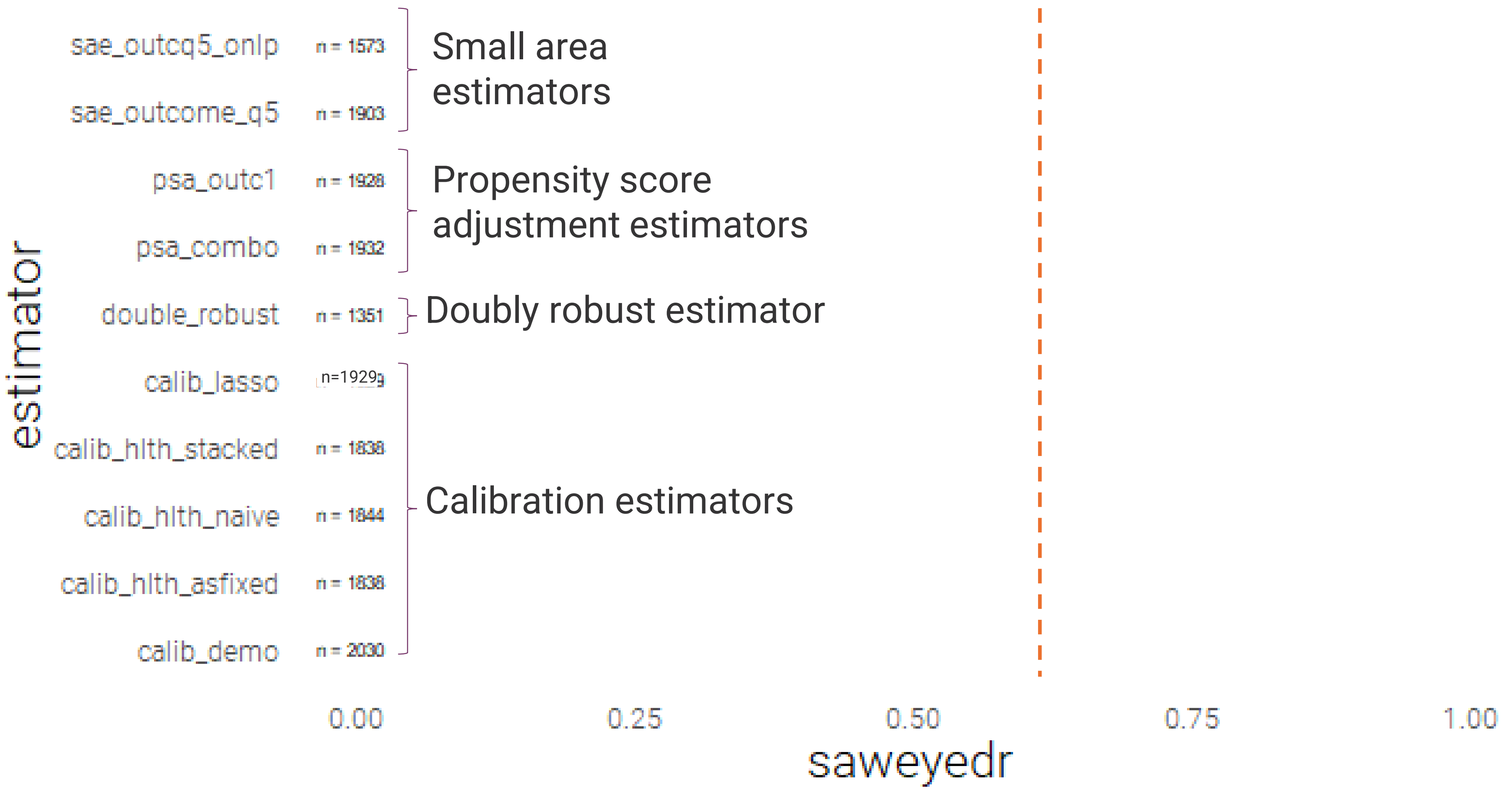
Full Sample (age 18+):

Full Sample (age 18+):

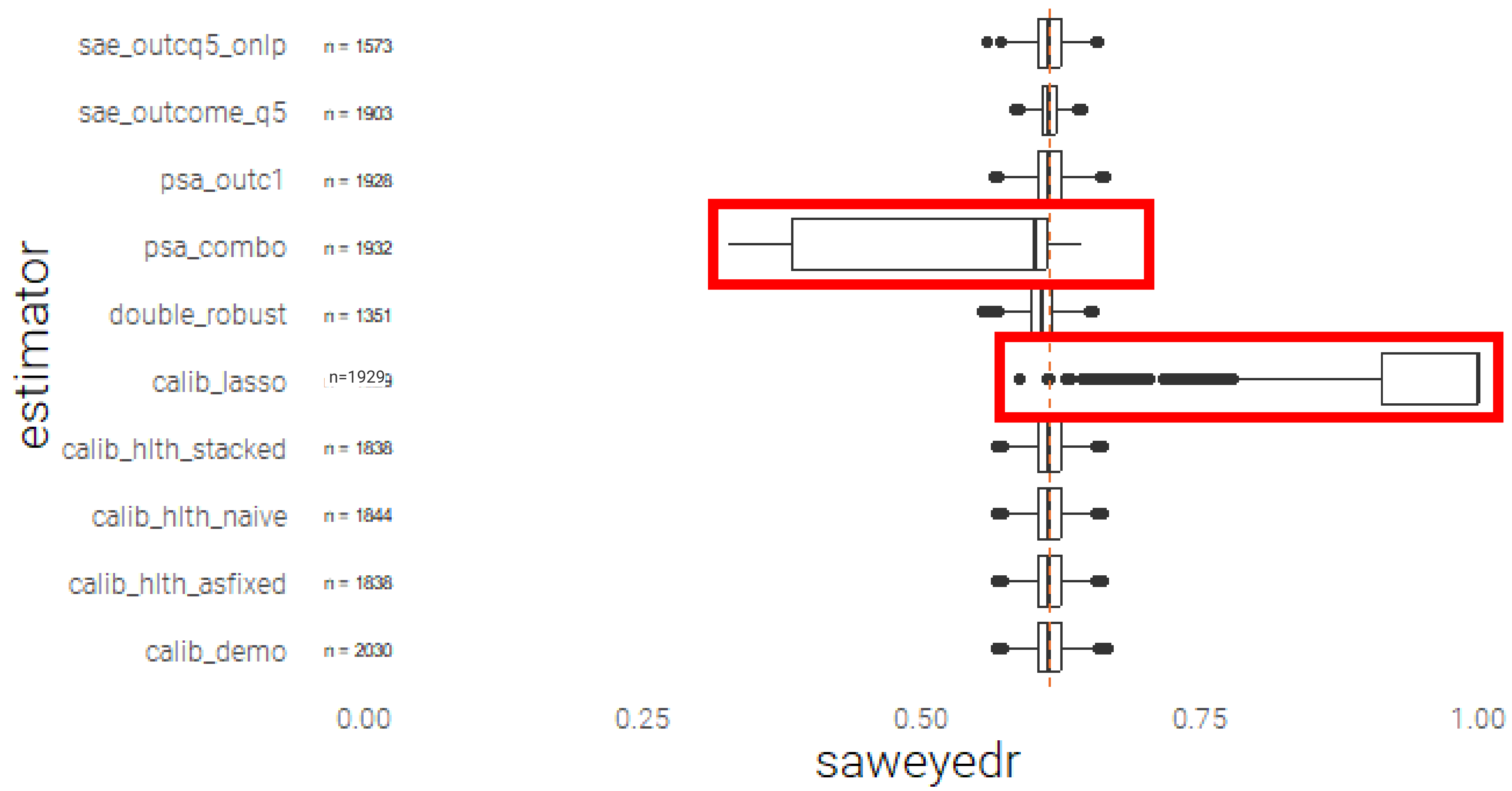
(TrueNorth)



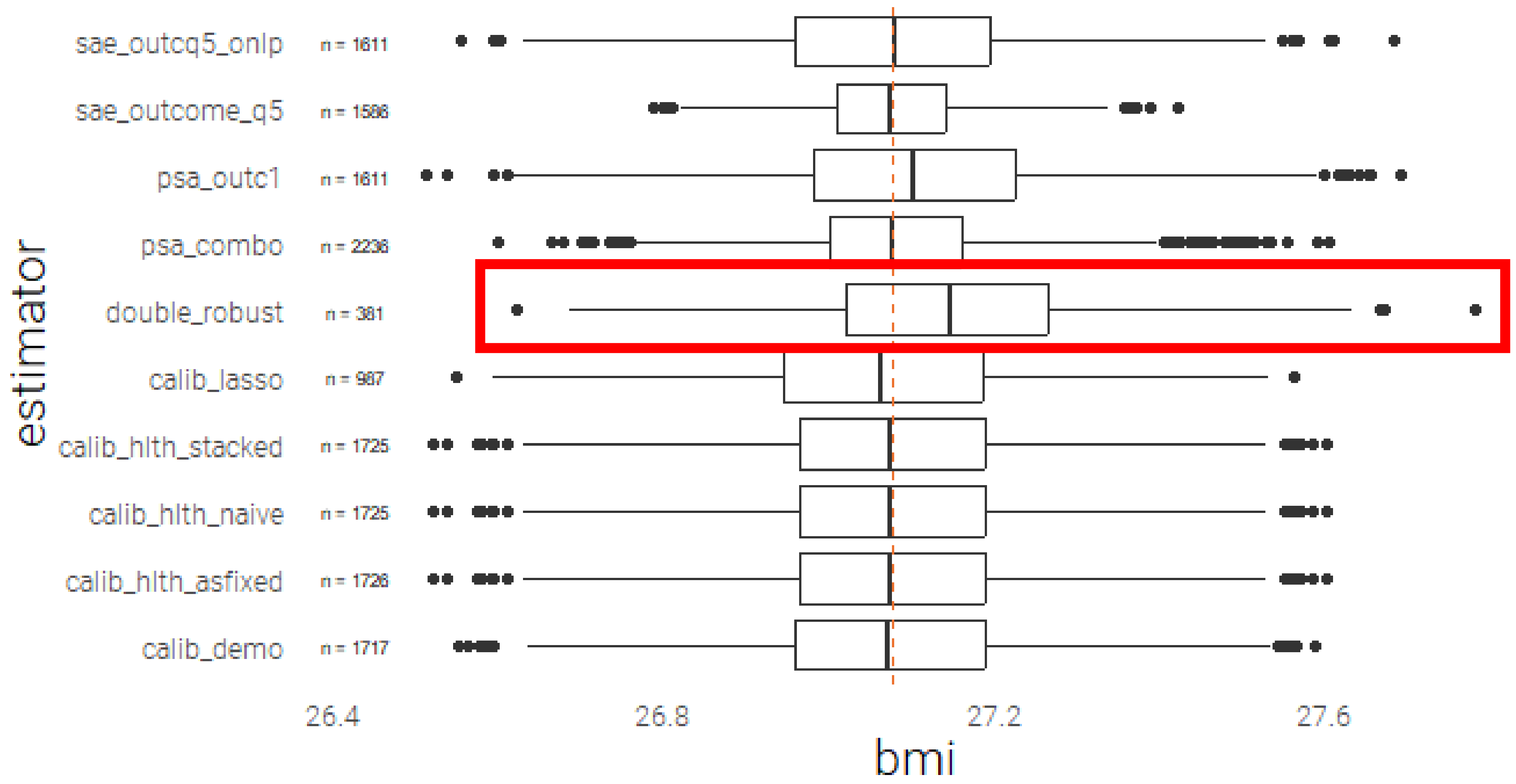
Full Sample (age 18+):



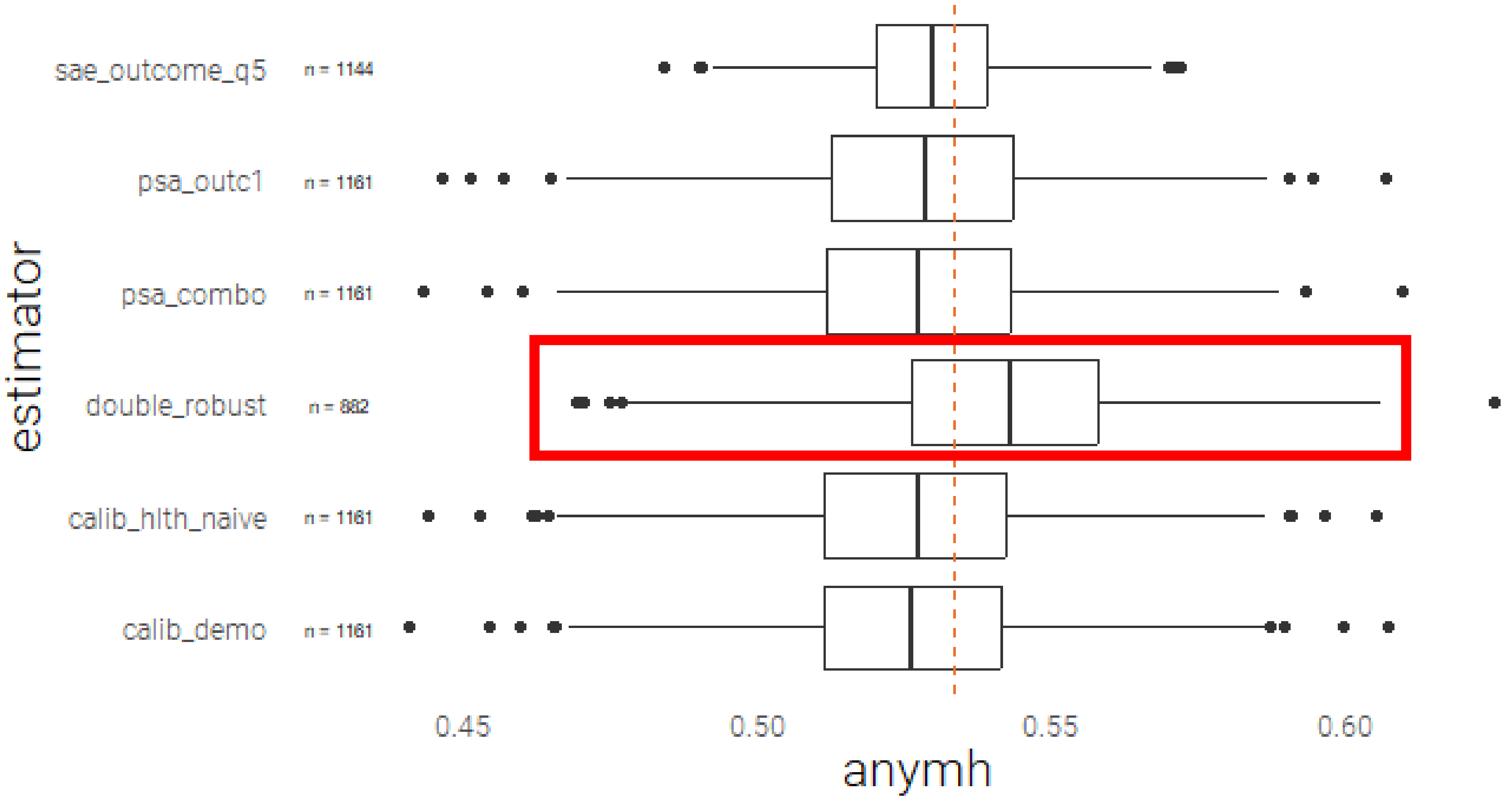
Full Sample (age 18+):



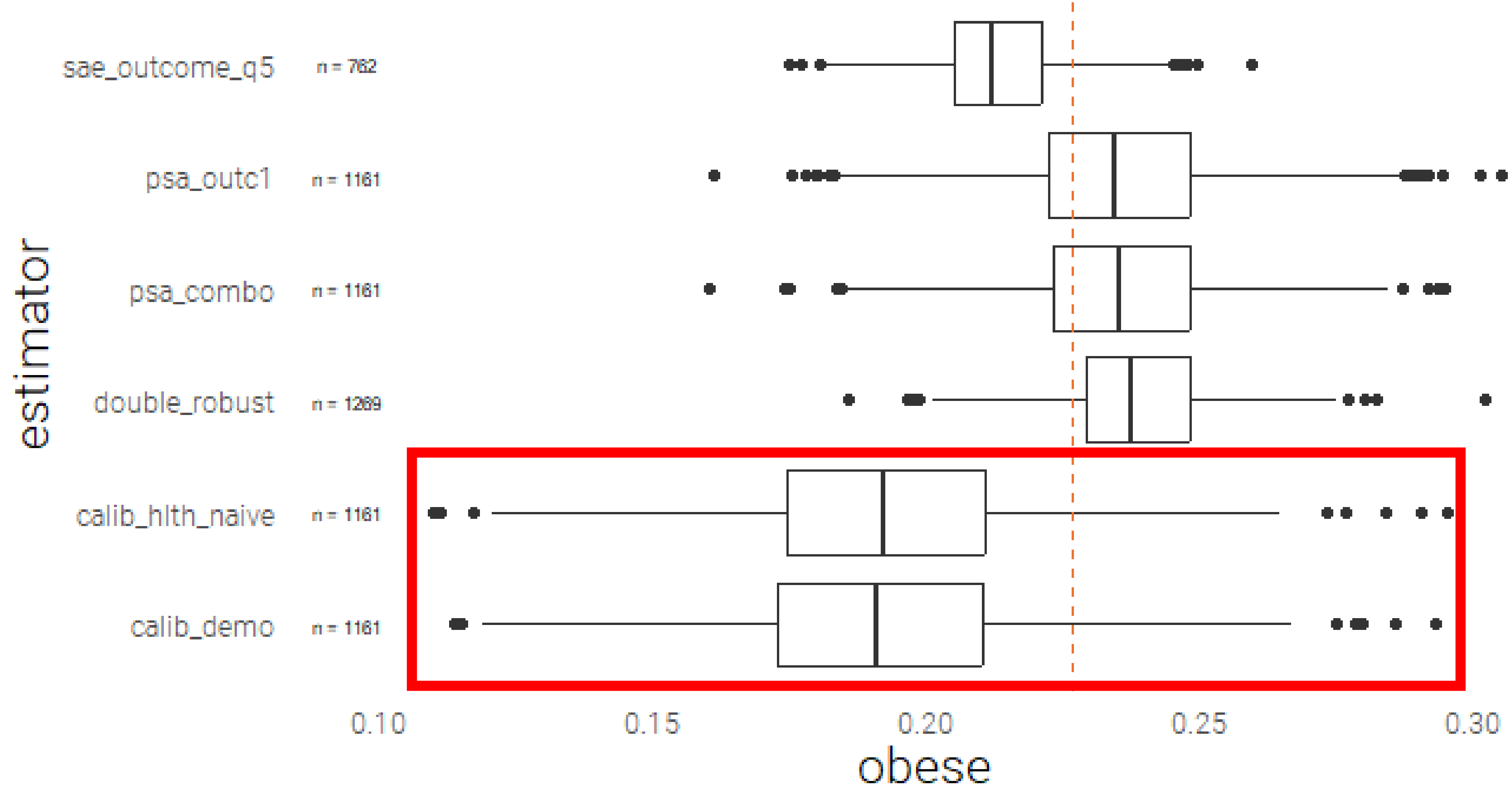
Full Sample (age 18+):



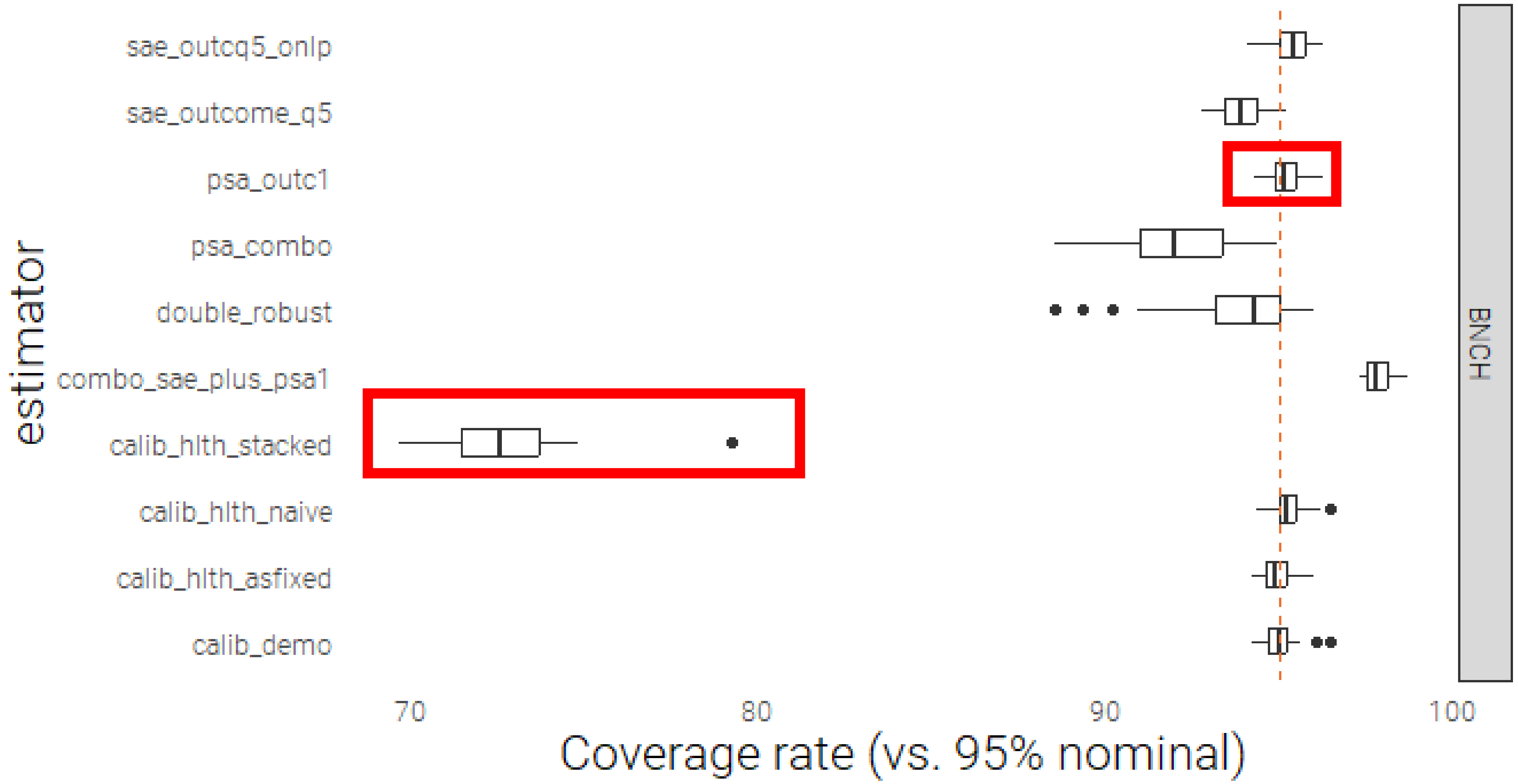
Non-Hispanic White Subgroup:



Full Sample (age 18+)



Confidence interval coverage across full sample/subgroups:



Lessons learned and future work

NCHS program of research into online panels

What did we learn?

Some clear losers:

- Double robust estimator failed inexplicably even in the simple situations
  - Violation of standard assumptions? (exchangeability, positivity; model specification)
- Lasso calibration failed with categorical variables
  - Lasso may be over-shrinking to nearly constant predictions, in which case calibration isn't doing anything
- Combining weights from individual PSA models did not work

No clear winners, but still in the game:

- Demographic domain small area + model calibration (TrueNorth)
  - Standard errors are biased down in the combined data – badly mismatched PSU sizes?
- Propensity score adjustments for individual outcomes
- Calibration (non-lasso)
  - Naïve standard errors; other methods produce standard errors that are too optimistic

Current work

Implementation in production

- Using the better performing methods in reporting with RSS real data

Re-trying methods

- Different libraries
- Improving model specification
- Other ways of producing omnibus propensity score-based weights

Keeping an eye on the blending literature

- Adding any new promising methods to the comparison

Questions?

BigSurv26: Reserve The Date!

March 2026 – Research Triangle, NC – https://bigsurv.org/

NORC at the University of Chicago

Thank you.

Stas Kolenikov
Principal Statistician
kolenikov-stas@norc.org

Research You Can Trust™

NORC at the University of Chicago