50TH ANNIVERSARY SEER

TURNING CANCER DATA INTO DISCOVERY

Bias-corrected Cancer Incidence Rate to account for Differential Privacy Error in the Population Data

Mandi Yu, National Cancer Institute

Jiming Jiang, University of California, Davis

October 24-26, 2023

College Park, MD

NIH NATIONAL CANCER INSTITUTE

TURNING CANCER DATA INTO DISCOVERY

Disclaimer:

The opinions expressed in this presentation are the author's own and do not reflect the view of the National Cancer Institute, National Institutes of Health, the Department of Health and Human Services, or the United States government.

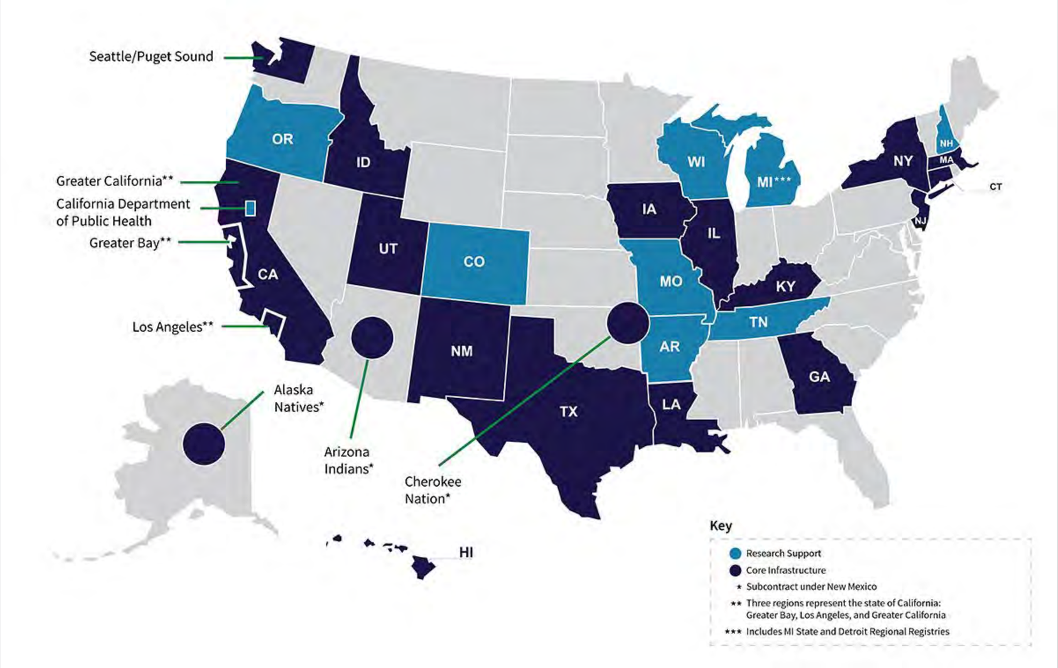
NIH NATIONAL CANCER INSTITUTE

Outline

- Background
  - NCI's SEER program and Cancer Rates
  - Differential Private (DP) Census Population Estimates
  - Bias-corrected Rate Estimator
- Study Objective – performance of correcting for bias due to DP error in Population data
- Discussion

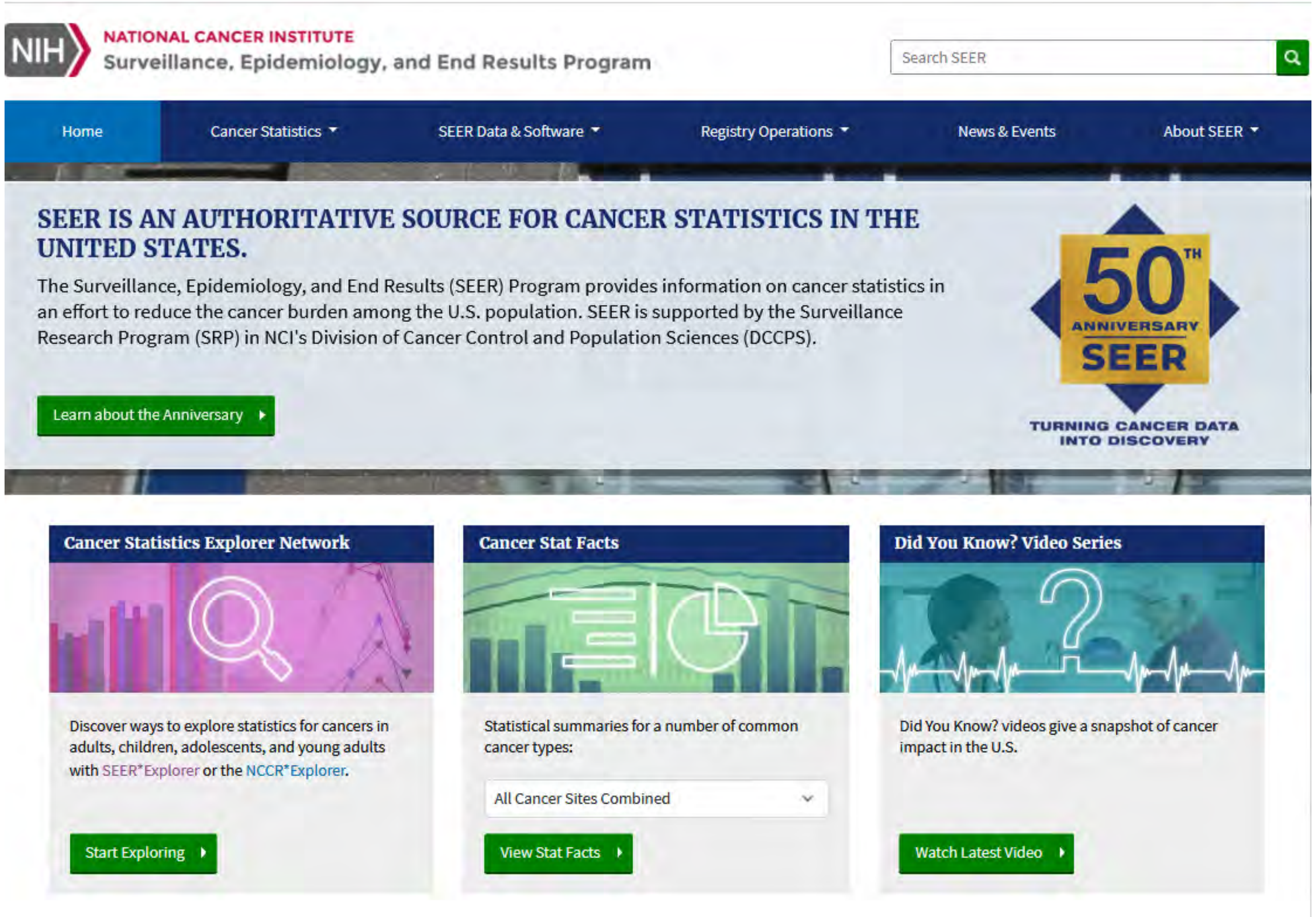
Cancer Surveillance and SEER Areas

- Surveillance, Epidemiology, and End Results (SEER) program is established and supported by the NCI since 1973.
- SEER collects and publishes cancer incidence and survival data from population-based registries across the nation.
- Information include patient demographics, primary tumor site, tumor morphology and stage at diagnosis, first course of treatment, and follow-up for vital status (survival)

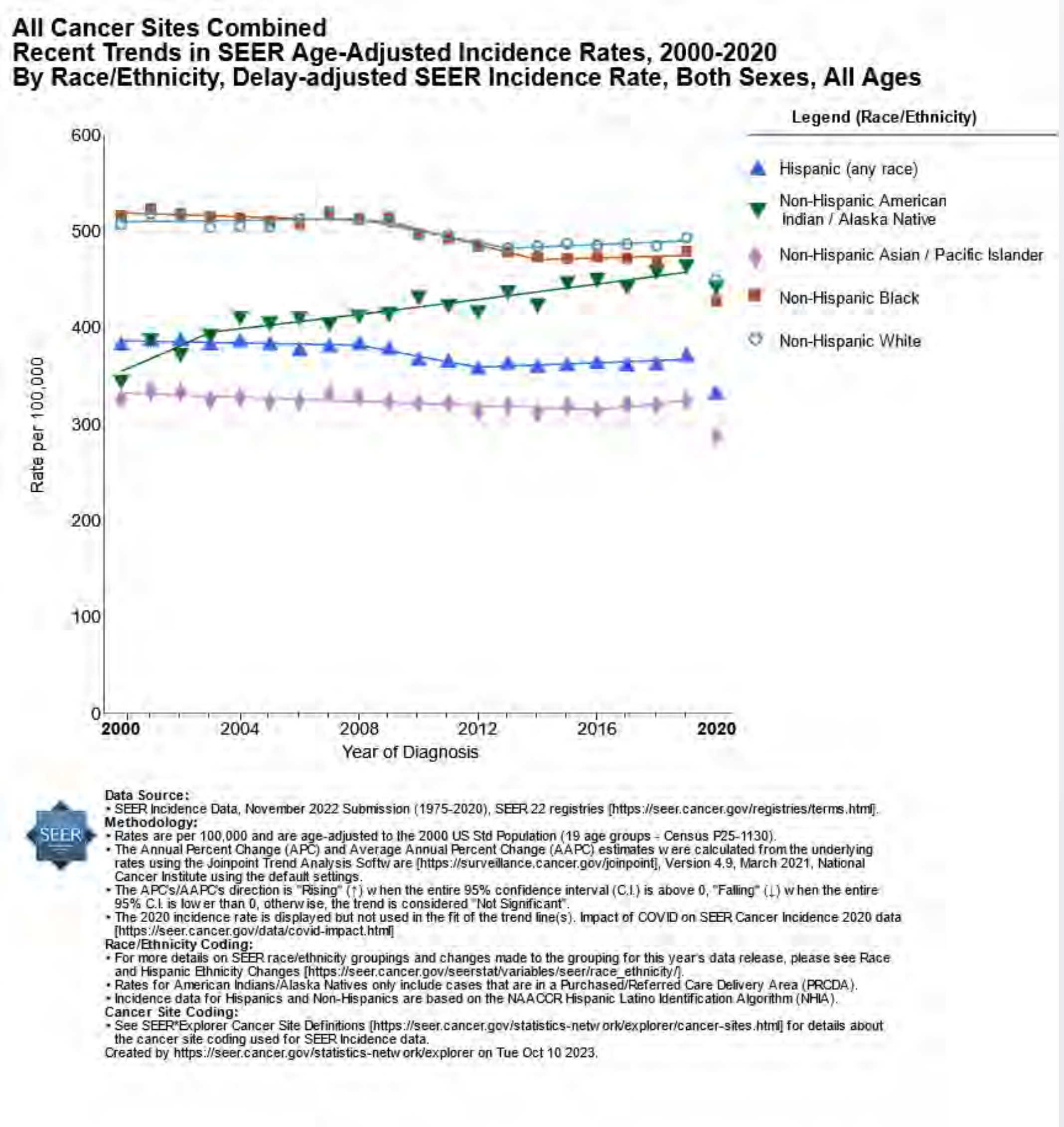


SEER registries cover ~48% of the U.S. population.

50TH ANNIVERSARY SEER
TURNING CANCER DATA INTO DISCOVERY


NIH
NATIONAL CANCER INSTITUTE
Surveillance, Epidemiology, and End Results Program
Search SEER
Home
Cancer Statistics
SEER Data & Software
Registry Operations
News & Events
About SEER
SEER IS AN AUTHORITATIVE SOURCE FOR CANCER STATISTICS IN THE UNITED STATES.
The Surveillance, Epidemiology, and End Results (SEER) Program provides information on cancer statistics in an effort to reduce the cancer burden among the U.S. population. SEER is supported by the Surveillance Research Program (SRP) in NCI's Division of Cancer Control and Population Sciences (DCCPS).
Learn about the Anniversary
50TH ANNIVERSARY SEER
TURNING CANCER DATA INTO DISCOVERY
Cancer Statistics Explorer Network
Discover ways to explore statistics for cancers in adults, children, adolescents, and young adults with SEER*Explorer or the NCCR*Explorer.
Start Exploring
Cancer Stat Facts
Statistical summaries for a number of common cancer types:
All Cancer Sites Combined
View Stat Facts
Did You Know? Video Series
Did You Know? videos give a snapshot of cancer impact in the U.S.
Watch Latest Video

Age-Adjusted Rate of Cancer Incidence



$$\text{Age Adjusted Rate} = \sum_J w_j \frac{c_j}{N_j}$$

c_j = number of tumors in age group j

w_j = age-adjusting weight in age group j

N_j = at risk population in age group j, and it is assumed to be error-free (despite of small enumeration error)

Bias-corrected Rate

$$\text{Age Adjusted Rate}_{bc} = \sum_J w_j \frac{c_j}{\hat{N}_j}\left(\mathbf{1} - \boldsymbol{CV}^{\mathbf{2}}\right)$$

$\hat{N}_j$ = *estimated population in age group j*

CV_j = *estimated **C**oefficient of **V**ariation of population in age group j*

> Stat Methods Med Res. 2021 Feb;30(2):535-548. doi: 10.1177/0962280220962516. Epub 2020 Oct 15.

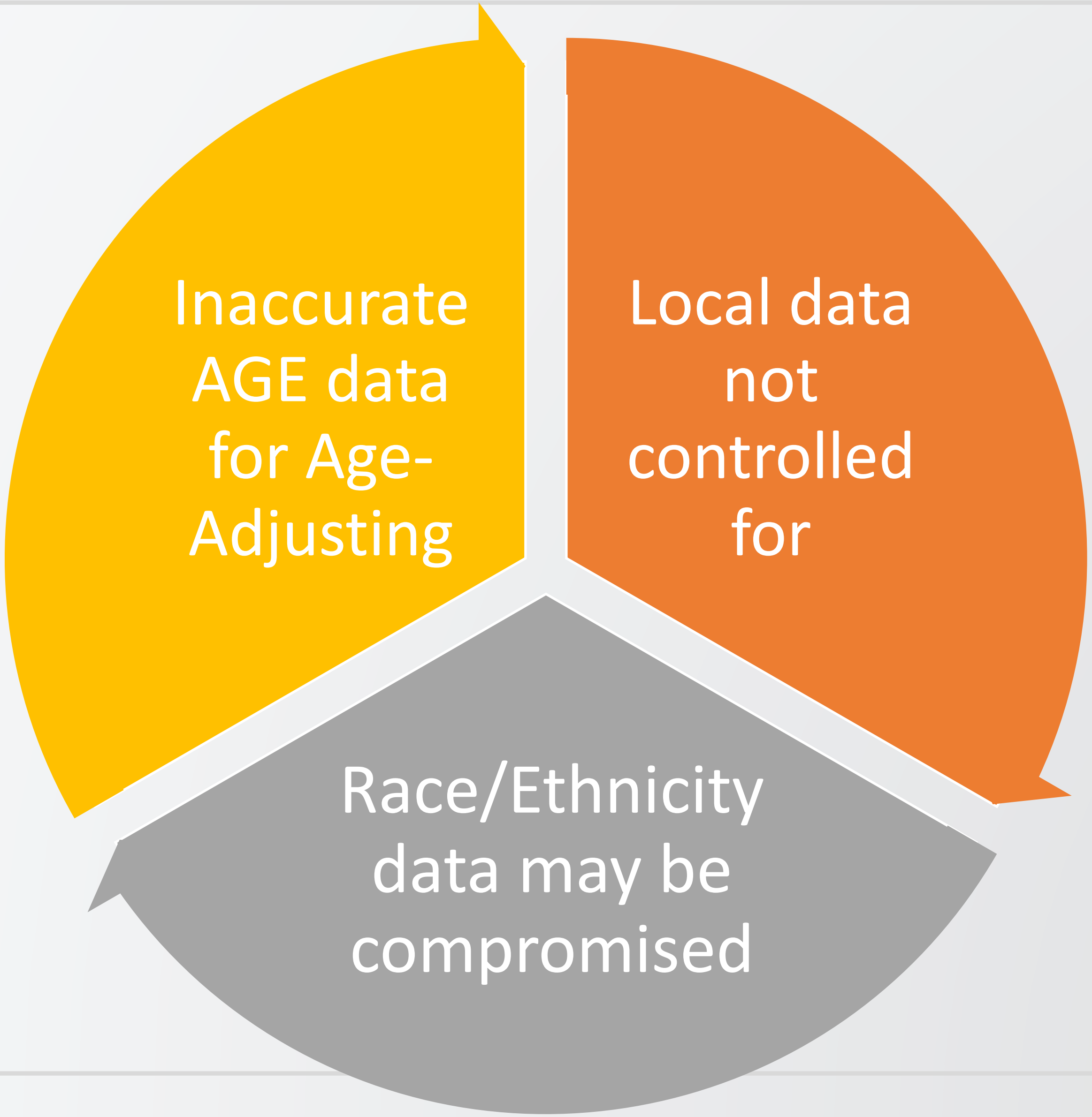
Inference about age-standardized rates with sampling errors in the denominators

Jiming Jiang [1], Eric J Feuer [2], Yuanyuan Li [1], Thuan Nguyen [3], Mandi Yu [2]

Differential Private Population Data

- A Topdown algorithm to adds noise – or variations from the actual count – to the collected data
- Geographic pop control totals were assigned a privacy budget (state totals are exactly matched)
- However, pop totals by demographics, such as age and race, are not controlled for and may be subject to systematic deviations
- Size of deviation tends to be greater for small groups despite of iterations of improvements from Census Bureau on the Topdown algorithms

A Few Key Challenges for Calculating Cancer Rates



50TH ANNIVERSARY SEER
TURNING CANCER DATA INTO DISCOVERY

Study Objectives and Methods

- Objectives:
  - 1. Impact of DP on the validity of rates
  - 2. Performance of bias-correction to alleviate the DP impact
- Use California data as a test case
  - Stratification variables: county and 5-year age group
  - Outcome: age-adjusted rates for counties
- Most challenging part is how to simulate DP population estimates
  - Detailed algorithms are confidential and kept within Census Bureau
  - Size of DP errors derived from Census 2010 demonstration dataset
  - A TopDown approach similar in principle to Census's algorithm

Population Simulation-A TopDown Appraoch

- Add normal-distributed noise to pops totals
- With variance of noise proportional to the size of population totals (p=0.2)
- Resulted noise is similar to observed differences between the demonstration and real 2010 census data
- Optimized noise size using a two-step approach

Simulate Study

- Questions to be answered
  - Whether Bias-correction helps to adjust for DP error?
  - What is the population cutoffs for DP errors to be negligible for age-adjusted rates of incidence (AAR)?

- Study Methods:
  - Simulate Poisson incidence county by age and county (since cancers are random events)
  - Calculate AARs using DP-pop and Real-pop
  - Calculate % Relative Bias of AAR from simulation studies

Results: % Relative Bias in AAR (county-level)

Table 3: **Summary of %RB for Different ASR Estimators (TopDown Algorithm)**

Summary	Naive No DP	Naive with DP	Bias-correction with DP
Min.	-4.413	-2.238	-4.338
Q1	-0.110	-0.110	-0.111
Median	0.025	0.05	0.023
Mean	-0.098	-0.042	-0.102
Q3	0.263	0.283	0.282
Max.	1.603	2.071	1.310

Naïve No DP: Population does not have DP errors
Naïve with DP: Population has DP errors and AAR is calculated using the standard method (i.e. without bias correction)
Bias-correction with DP: Population has DP errors and AAR is bias-corrected

Boxplots of % Relative Bias

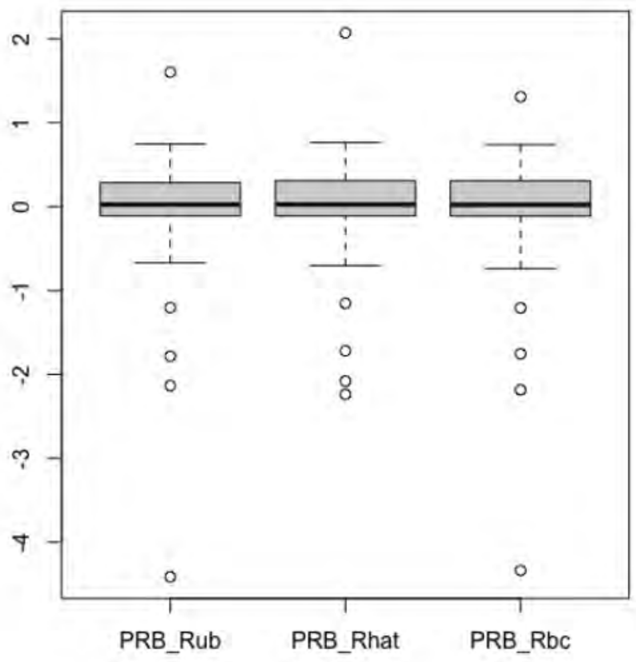
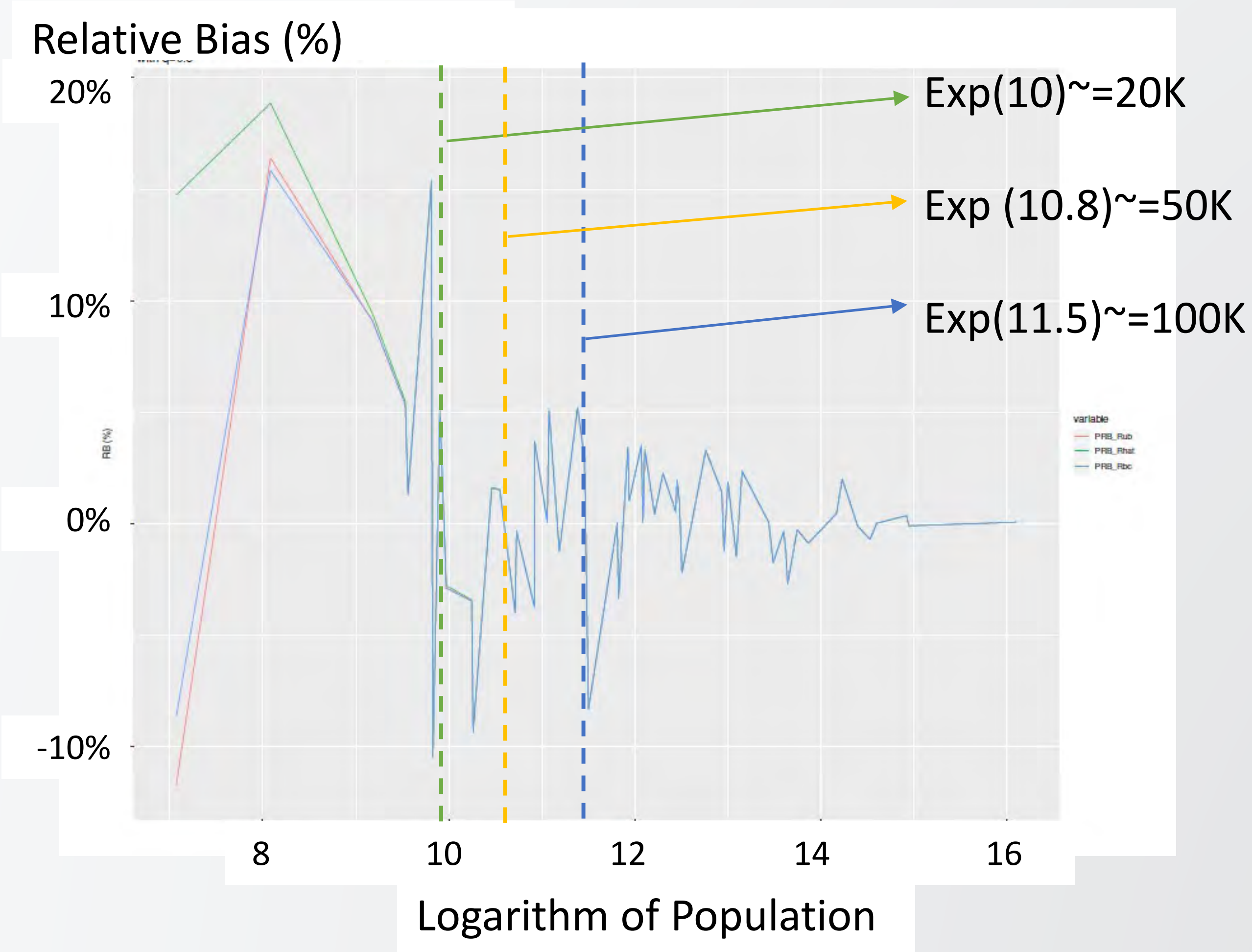


Figure 2: Boxplots of %RB (TopDown DP algorithm): PRB_Rub: Naive No DP; PRB_Rhat: Naive with DP; PRB_Rbc: Bias-correction with DP

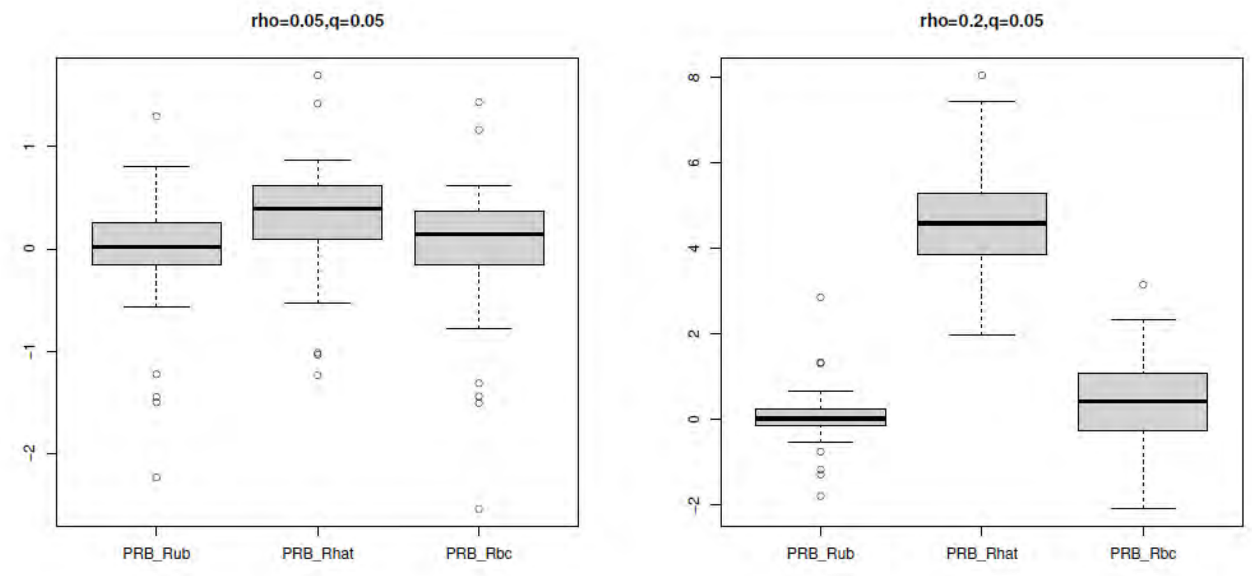
Note: Each dot corresponds to one county

% Relative Bias by County Population Size (log scale)



A further simulation of Sampling Error+ DP Error

- The case when DP error is added to survey samples
- E.g., DP modified American Community Survey estimates.
- ρ is the measure of sampling fraction and gauges sampling error
- q gauges DP error



(a) $\rho = 0.05$ (b) $\rho = 0.2$

Figure 3: Boxplots of %RB (TopDown DP algorithm): $q = 0.05$; (a) $\rho = 0.05$, (b) $\rho = 0.2$; PRB_Rub: Naive No DP; PRB_Rhat: Naive with DP; PRB_Rbc: Bias-correction with DP

Discussion

- The magnitude of DP bias is not comparable to that of sampling error
- The impact of bias-correct on AARs is small in relative to sampling error
- DP error's impacts become small/negligible only if the population is at least 100K or greater – limit the ability for detailed disparity analysis
- Current simulation is limited to Age variable and the next step is to consider race/ethnicity

Policy Implications

- Health burden studies for small subpopulations or at local geographic level are becoming impossible, however policies are mostly ‘local’ and ‘specific’
- Noise metrics released by Census Bureau are not detailed enough to help understand the extend of impact on cancer rates
- User community would benefit a guidance from Census Bureau regarding to how to use the noise metrics, e.g., how to relate noise measure to variance for common statistics

50TH ANNIVERSARY SEER

TURNING CANCER DATA INTO DISCOVERY

Contact: Mandi Yu (mandi.yu@nih.gov)

NIH NATIONAL CANCER INSTITUTE