Westat

Improving Lives Through Research®

Assessing Utility of Synthetic Data for Two Studies:

Applications to the Survey of Doctoral Recipients and Census Transportation Planning Products

Robyn Ferg*, Minsun Riddles, Tom Krenzke

WESTAT @ FCSM 2024

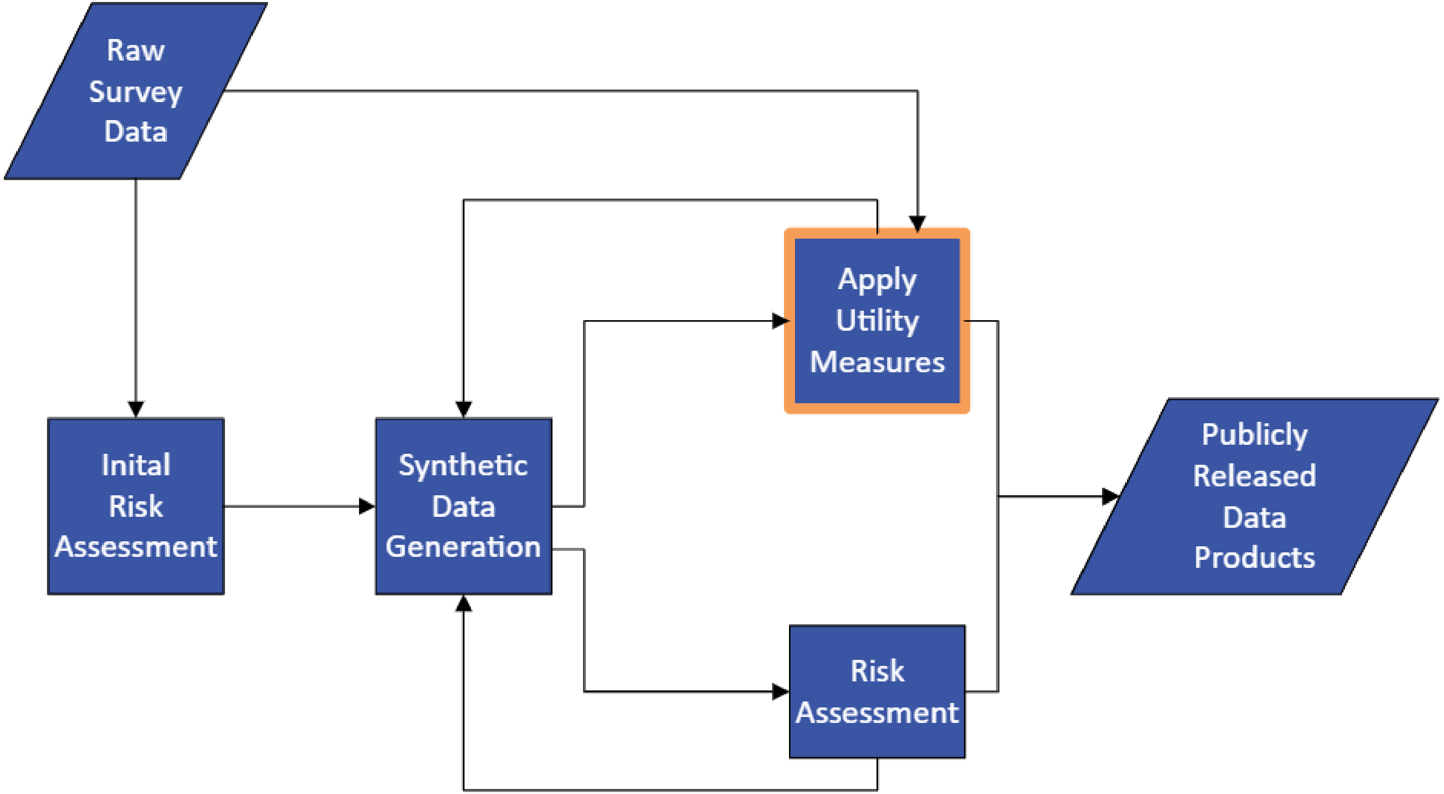
The views presented are those of the author(s) and do not represent the views of any Government Agency/Department or Westat

DRB DAO approval # CBDRB-FY25-ACSO003-0004

Synthetic Data

- Synthetic data generation is gaining traction as a method for avoiding disclosure in publicly released data products
- Conclusions reached using the synthetic data must be similar to those reached using the original data
- A careful balance must be struck between utility and risk of disclosure
- This presentation focuses on the utility side of the risk-utility tradeoff

Synthetic Data Generation Process



Utility Measures

- We compiled a list of several utility measures, which we sorted into five broad groups:
  - QC Checks
  - Weighted Frequency Checks
  - Measures of Association
  - Dataset-Wide Checks
  - Equity-Focused Measures
- These measures were applied to data synthesized for the Survey of Doctoral Recipients (SDR), sponsored by NCSES, and the Census Transportation Planning Products (CTPP), sponsored by Census Bureau and American Association of State Highway and Transportation Officials

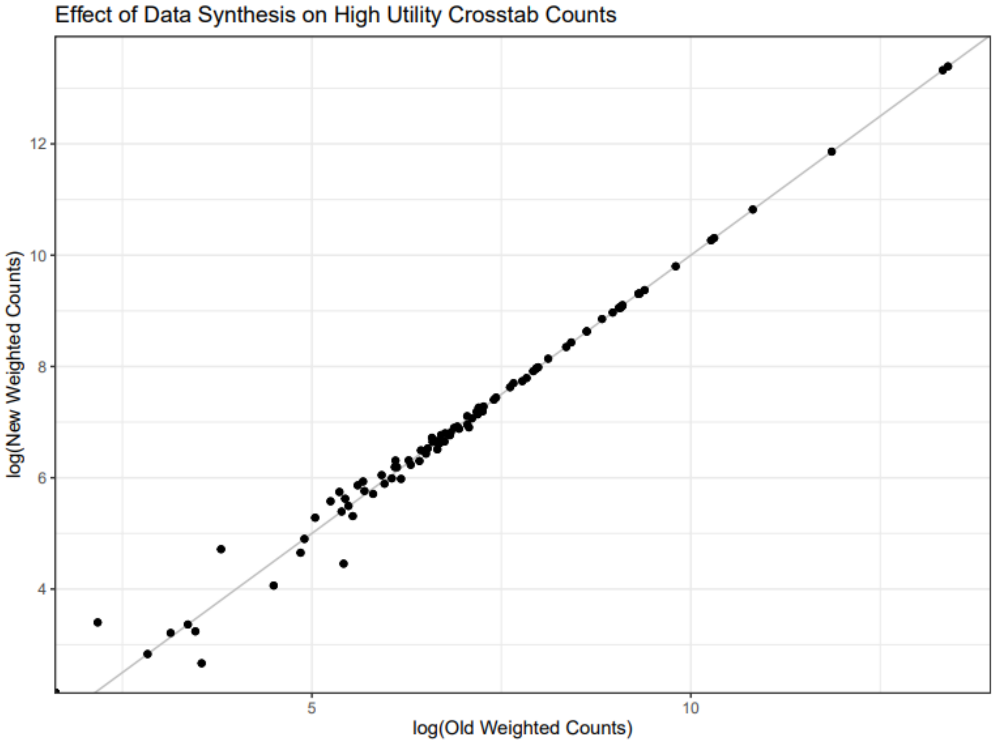
QC Checks

- Percent of records changed for each variable
- Percent of variables changed for each record
- Logic checks: differences in missing value patterns/combinations

Weighted Frequency Checks

- Distances between variables
  - Summary statistics of individual differences for continuous variables
  - Hellinger distance for categorical variables
- High-utility crosstabs
- Confidence interval overlap
- Indicators of whether synthesized estimate is within the original confidence interval

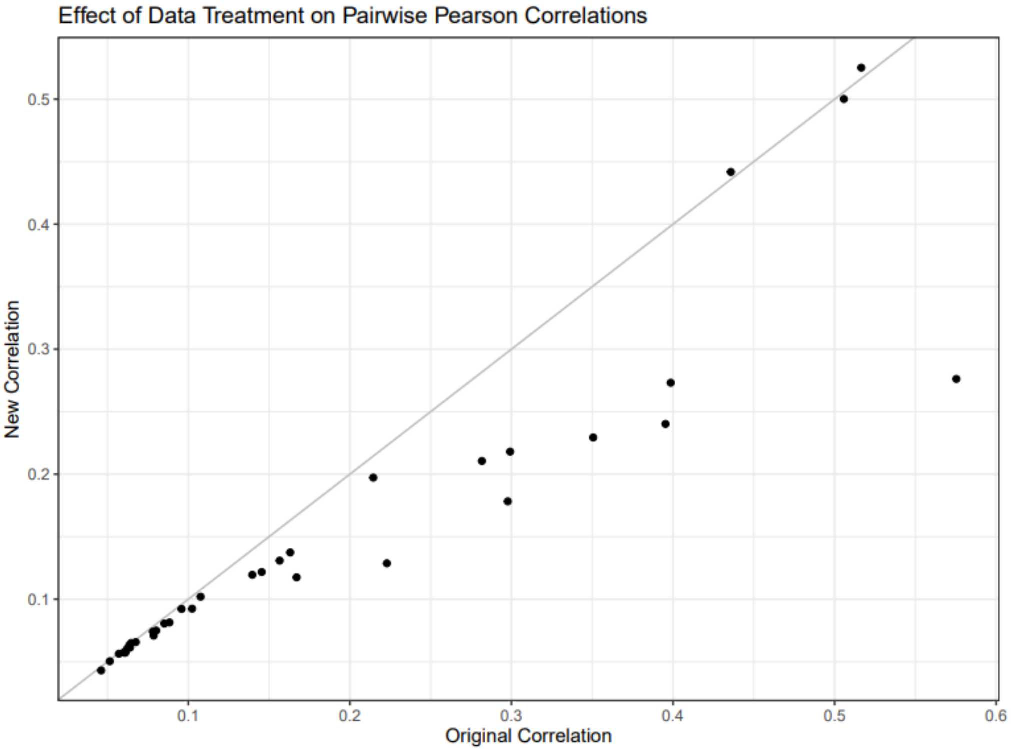
Example: High-Utility Crosstabs



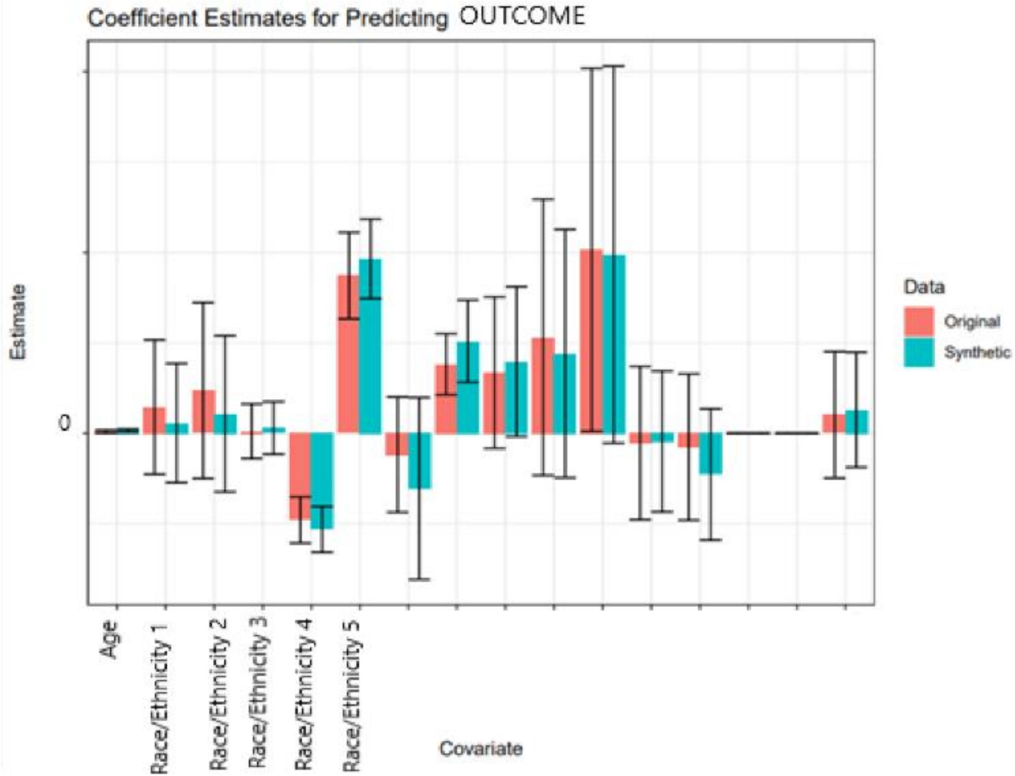
Measures of Association

- Pairwise associations
- Significance of regression coefficients
- U-Statistic (Woo et al., 2009, Snoke et al., 2018)

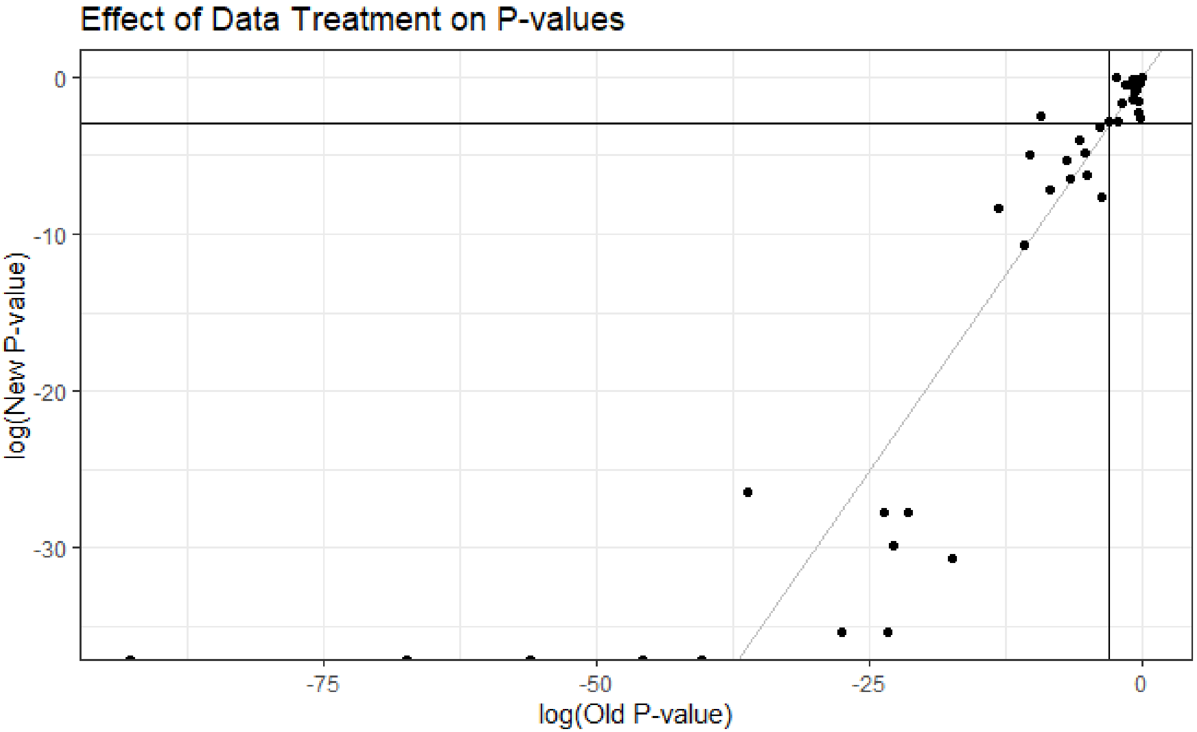
Example: Pairwise Associations for an Initial Diagnostic Run



Example: Regression Coefficients



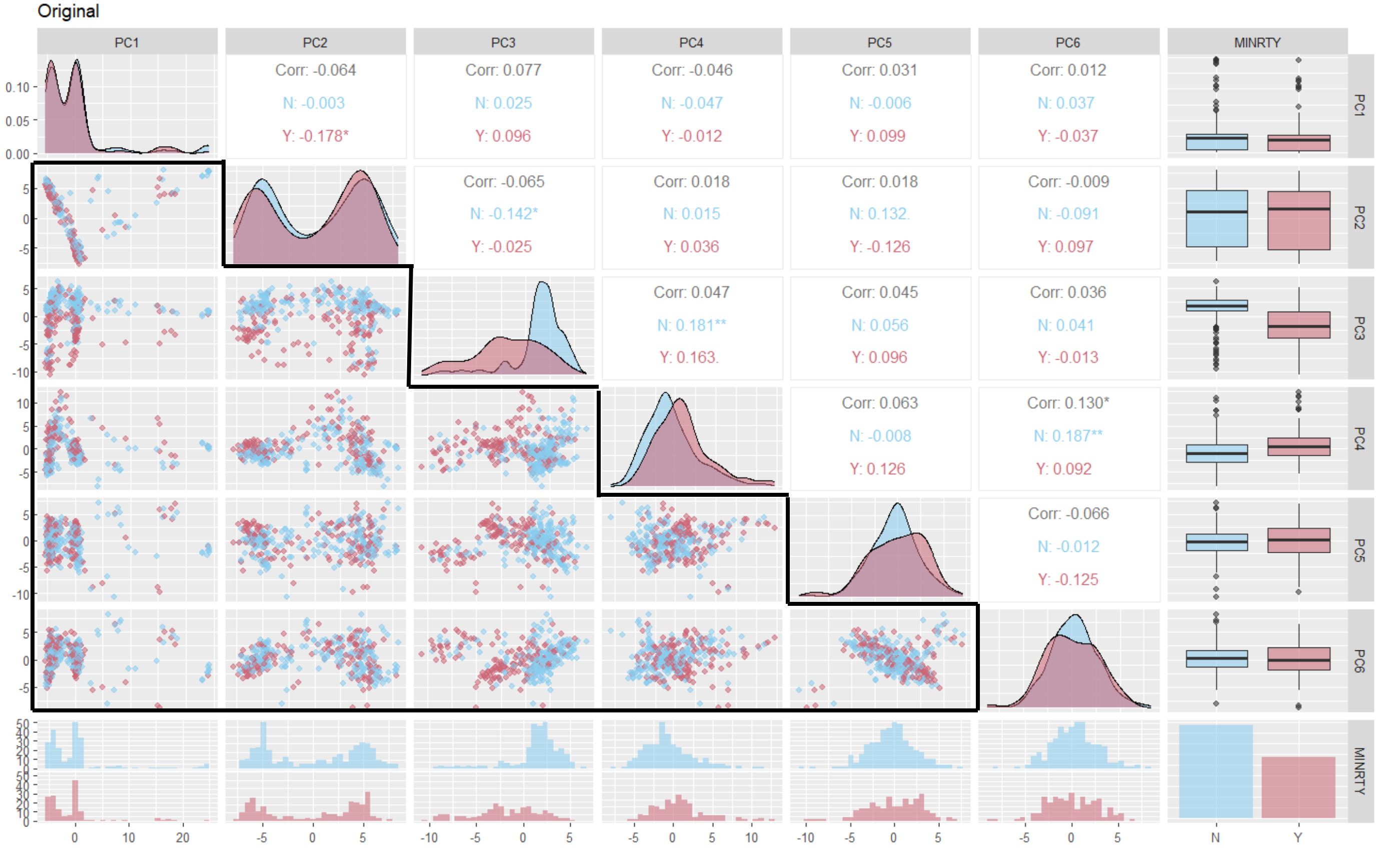
Example: Significance of Regression Coefficients



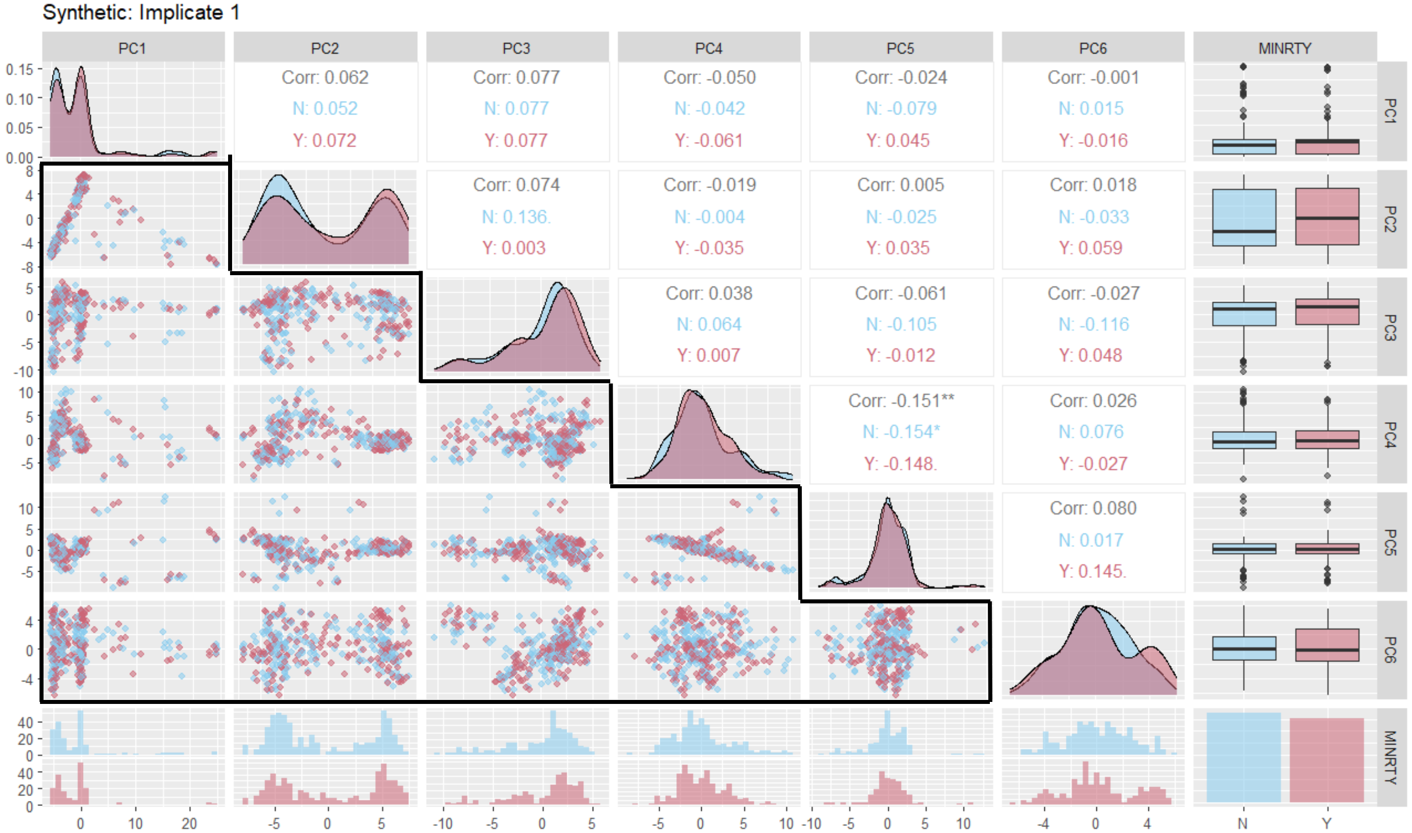
Dataset-Wide Checks

- Perform Principal Component Analysis (PCA) on the original and synthetic data sets to ensure patterns hold across the entire data set

Example: PCA on Original Data



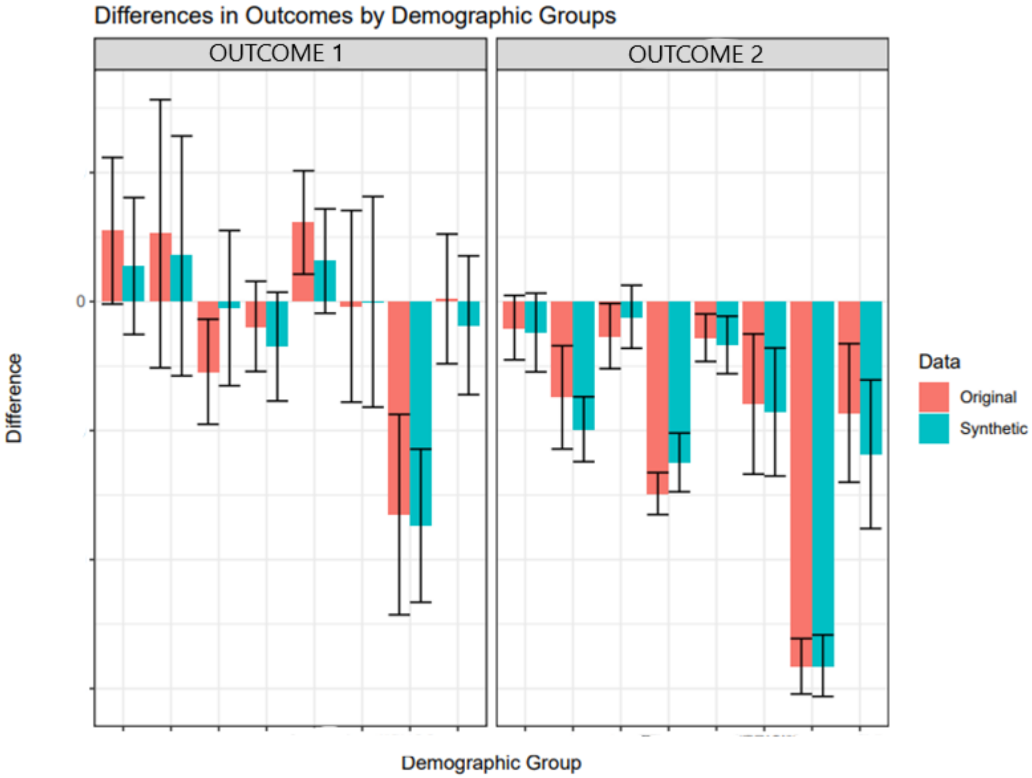
Example: PCA on Synthetic Data



Equity-Focused Measures

- Apply earlier utility measures, but subset to specific demographic groups
- Compare differences in outcomes between subgroups

Example: Differences in Outcomes by Demographic Groups



Conclusion

- The utility measures are important in determining whether a proposed publicly released synthesized data set is suitable for analysts' needs
  - Iterative process of synthetic data generation
- Results used in combination with risk assessment to ensure an appropriate balance between respondents' privacy and data usefulness
- Take into consideration minority and vulnerable demographic groups to ensure equitable analytic results

Thank you!

RobynFerg@westat.com

westat.com

Photos are for illustrative purposes only. All persons depicted, unless otherwise stated, are models.

References

- Joshua Snoke, Gillian M. Raab, Beata Nowok, Chris Dibben, Aleksandra Slavkovic, General and Specific Utility Measures for Synthetic Data, *Journal of the Royal Statistical Society Series A: Statistics in Society*, Volume 181, Issue 3, June 2018, Pages 663–688, https://doi.org/10.1111/rssa.12358

- Mi-Ja Woo, Jerome P. Reiter, Anna Oganian, Alan F. Karr, Global Measures of Data Utility for Microdata Masked for Disclosure Limitation. *Journal of Privacy and Confidentiality*, Volume 1, Issue 1, 2009, Pages 111-124, https://doi.org/10.29012/jpc.v1i1.568