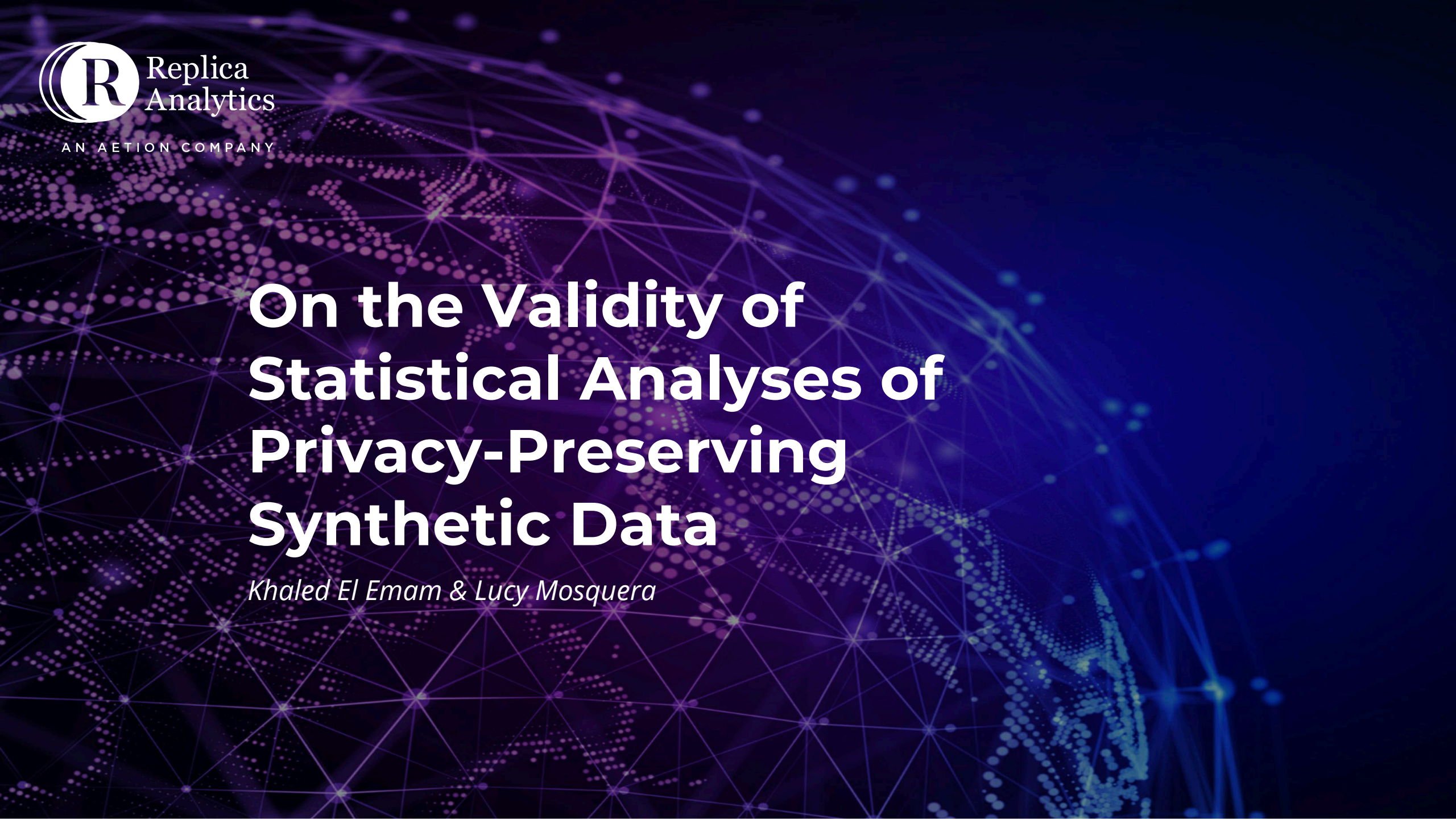


On the Validity of Statistical Analyses of Privacy-Preserving Synthetic Data

Khaled El Emam & Lucy Mosquera

Disclosures

Both KEE and LM are employed by Replica Analytics, an Aetion company. This study was performed with the University of Ottawa, where KEE is a professor, and the academic participation was funded by the Canada Research Chairs, NSERC, and CIHR. The study was also partially funded by the Gates Foundation.

Replica Analytics

AN AETION COMPANY

Agenda

- Introduction to synthetic data and its use cases
- Statistical inference and synthetic data
- Synthetic data for reproducing findings
- Synthetic data for population inference
- Privacy of synthetic data
- Conclusions

Replica Analytics

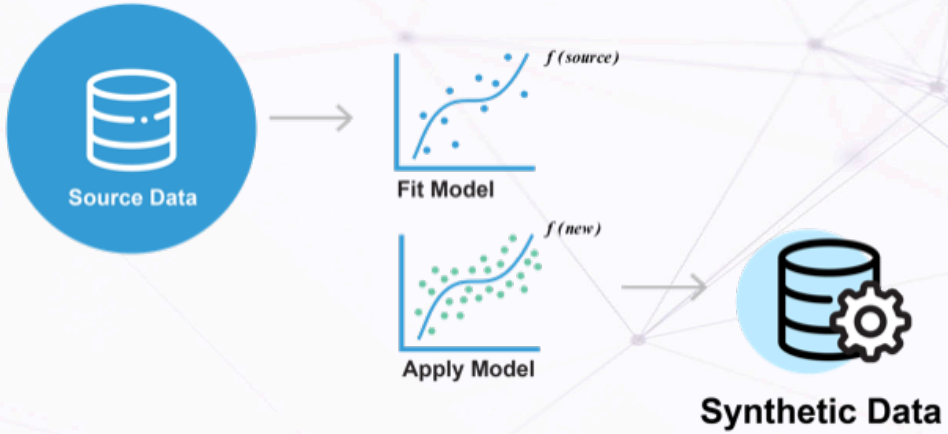
AN AETION COMPANY

What is Synthetic Data ?

Replica Analytics

AN AETION COMPANY

What is synthetic data ?



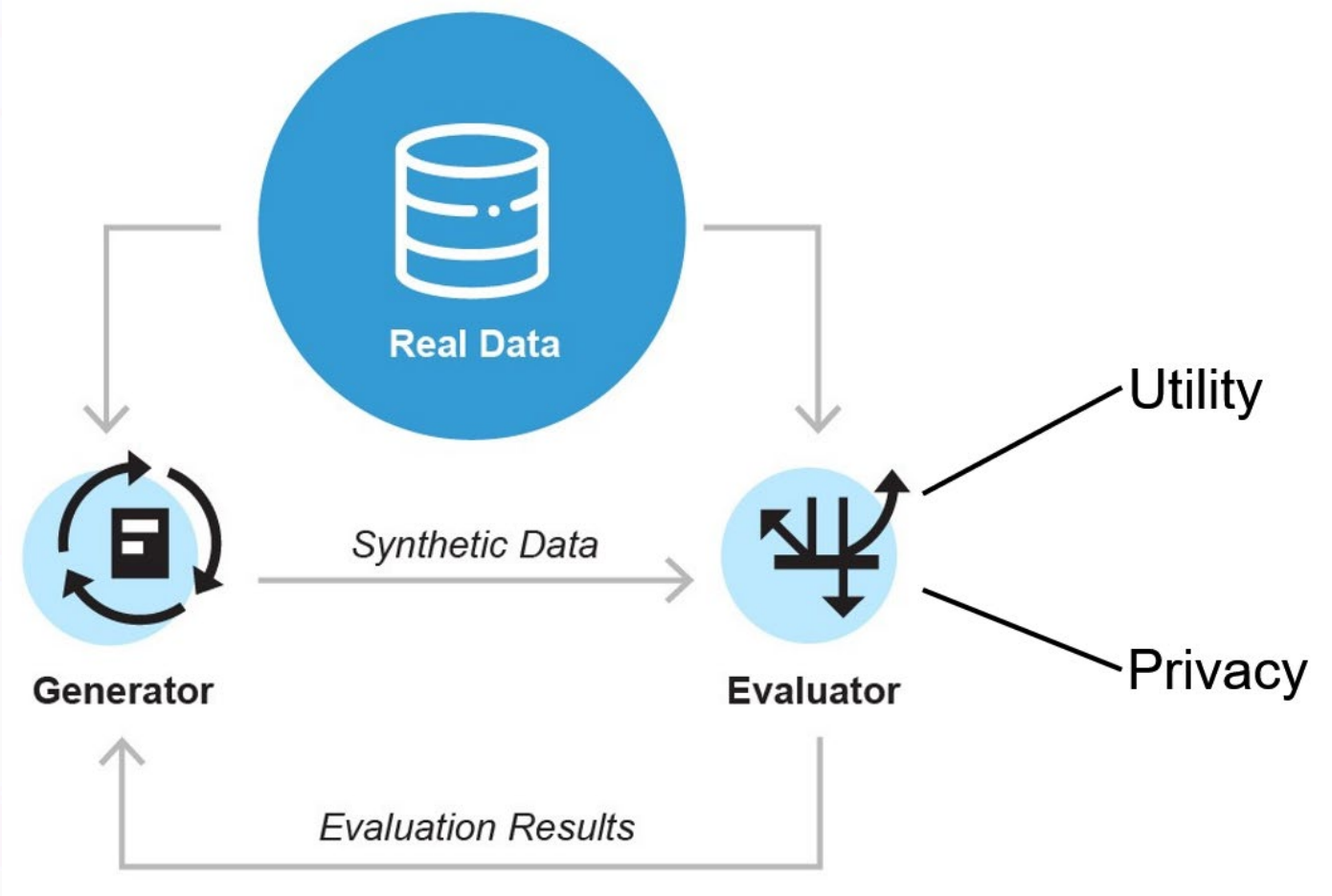
Additional Clarifications

- The source datasets can be as small as 150 patients or so. We have developed generative modeling techniques that will work reasonably well for small datasets. But this also depends on the number of variables in the dataset
- The source datasets can be very large – then it becomes a function of compute capacity that is available.
- It is not necessary to know how the synthetic data will be analyzed to build the generative models. The generative models capture many of the patterns in the source data.

COU1A	AGECAT	AGELE70	WHITE	MALE	BMI
United States	2	1	1	1	33.75155
United States	2	1	1	0	39.24707
United States	1	1	1	0	26.5625
United States	4	1	1	1	40.58273
United States	5	0	0	1	24.42046
United States	5	0	1	0	19.07124
United States	3	1	1	1	26.04938
United States	4	1	1	1	25.46939

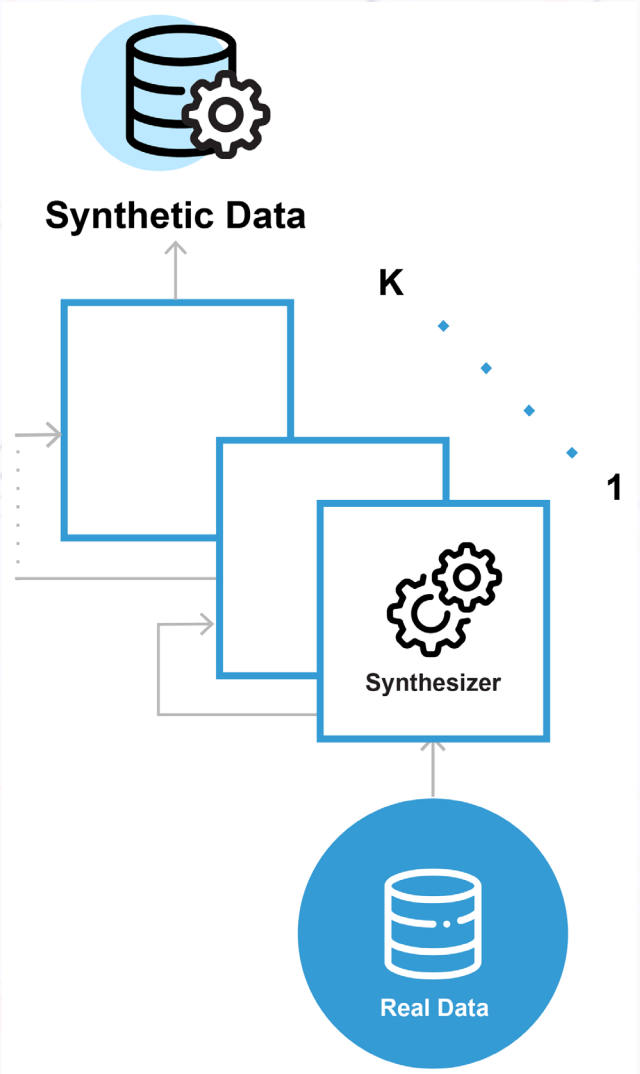
Replica Analytics
AN AETION COMPANY

How generative models work



Replica Analytics
AN AETION COMPANY

Sequential synthesis generative models



Khaled El Emam, Lucy Mosquera, Chaoyi Zheng, Optimizing the synthesis of clinical trial data using sequential trees, *Journal of the American Medical Informatics Association*, Volume 28, Issue 1, January 2021, Pages 3–13, https://doi.org/10.1093/jamia/ocaa249

Replica Analytics
AN AETION COMPANY

Main use cases for synthetic data

	Access	Efficacy	Privacy
	Cost, interoperability, processing power	*Robustness, diversity, fidelity*	*Transparency, data protection, regulations*
Insight improvement	Pharmacoepidemiology	Control group synthesis	Anonymization via synthesis
Topline opportunity	Simulated data asset creation	Powered hypothesis testing	Streamline data projects
Risk reduction	Cross-border synthesis	Enable AI	Risk based de-identification

Replica Analytics
AN AETION COMPANY

Operating models for secondary analysis using synthetic data

1. Data custodians share the synthetic data and conclusions are drawn from the analysis of synthetic datasets
2. Data custodians make synthetic data available for exploratory analysis and if there are interesting results, data users make a request for the full dataset (which may be a long and complicated process, but at least there is confidence that there are interesting results before initiating that process)
3. Perform the analysis on the synthetic data and then submit the analysis code (R, SAS, Python, …) to be executed on the real dataset behind a firewall – the external analysts never work with the real data

Replica Analytics
AN AETION COMPANY

Can synthetic data be a proxy for real data ?

Replica Analytics

AN AETION COMPANY

Assessing synthetic data

Generic utility

Show how similar synthetic data is to the real data it was generated from without referencing a specific analysis

Workload aware utility

Illustrate how well synthetic data can be used as a drop-in replacement or proxy for real data for a specific analysis

JMIR MEDICAL INFORMATICS El Emam et al

Original Paper

Utility Metrics for Evaluating Synthetic Health Data Generation Methods: Validation Study

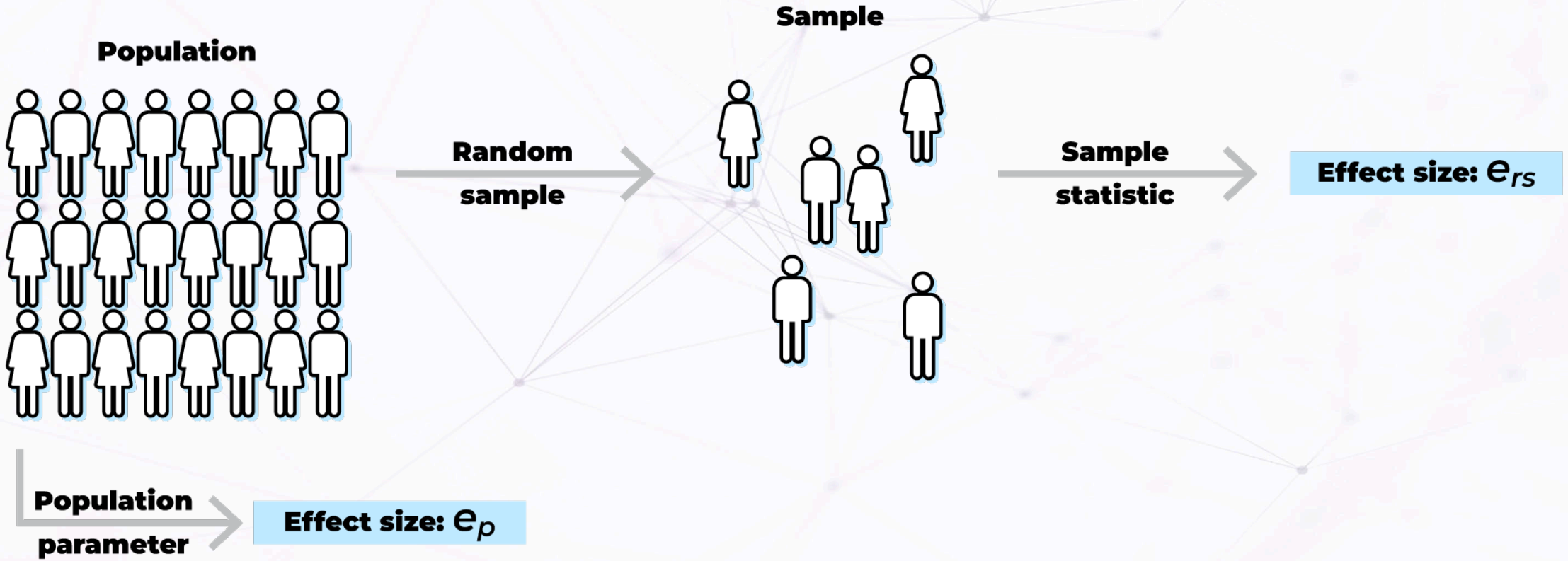
Khaled El Emam[1,2,3], BEng, PhD; Lucy Mosquera[2,3], BA, MSc; Xi Fang[3], BA, MSc; Alaa El-Hussuna[4], MSc, MD

[1]School of Epidemiology and Public Health, University of Ottawa, Ottawa, ON, Canada
[2]Children's Hospital of Eastern Ontario Research Institute, Ottawa, ON, Canada
[3]Replica Analytics Ltd, Ottawa, ON, Canada
[4]Open Source Research Collaboration, Aarlberg, Denmark

Today we will focus on using synthetic data as a proxy for real data in statistical analysis which is a kind of workload aware assessment.

Replica Analytics
AN AETION COMPANY

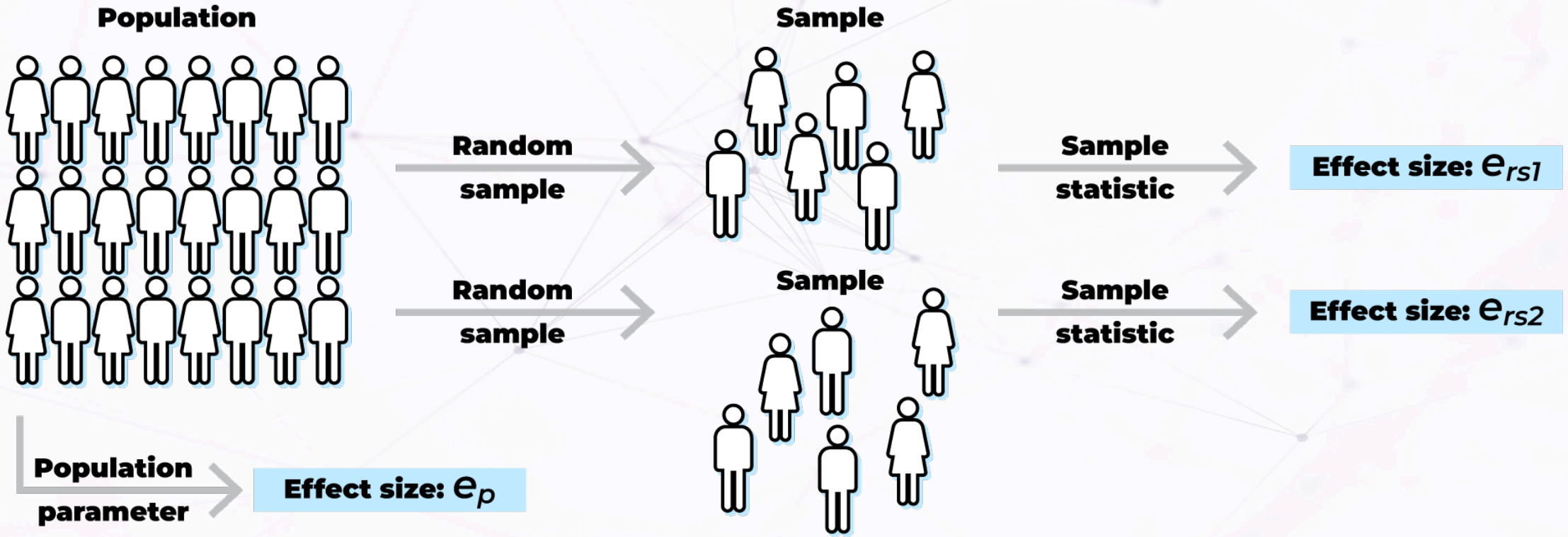
Statistical inference



The goal of statistical inference is to use observable sample statistics e_{rs} to infer unobservable population values e_p

Replica Analytics
AN AETION COMPANY

Statistical inference

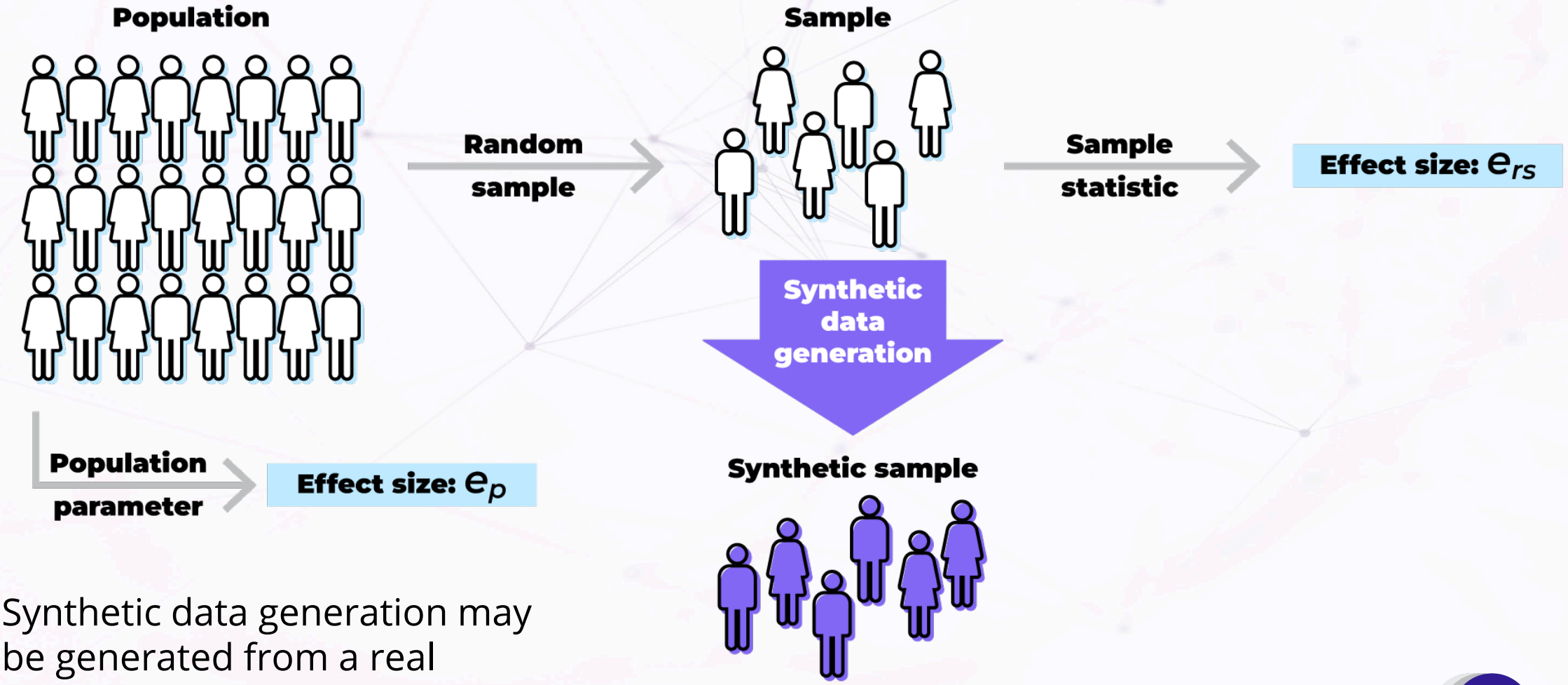


Random sampling introduces variability, each sample will produce a slightly different effect size, $e_{rs1} \neq e_{rs2}$

With appropriate study design, each sample statistic should be representative of the true population effect size e_p

Replica Analytics
AN AETION COMPANY

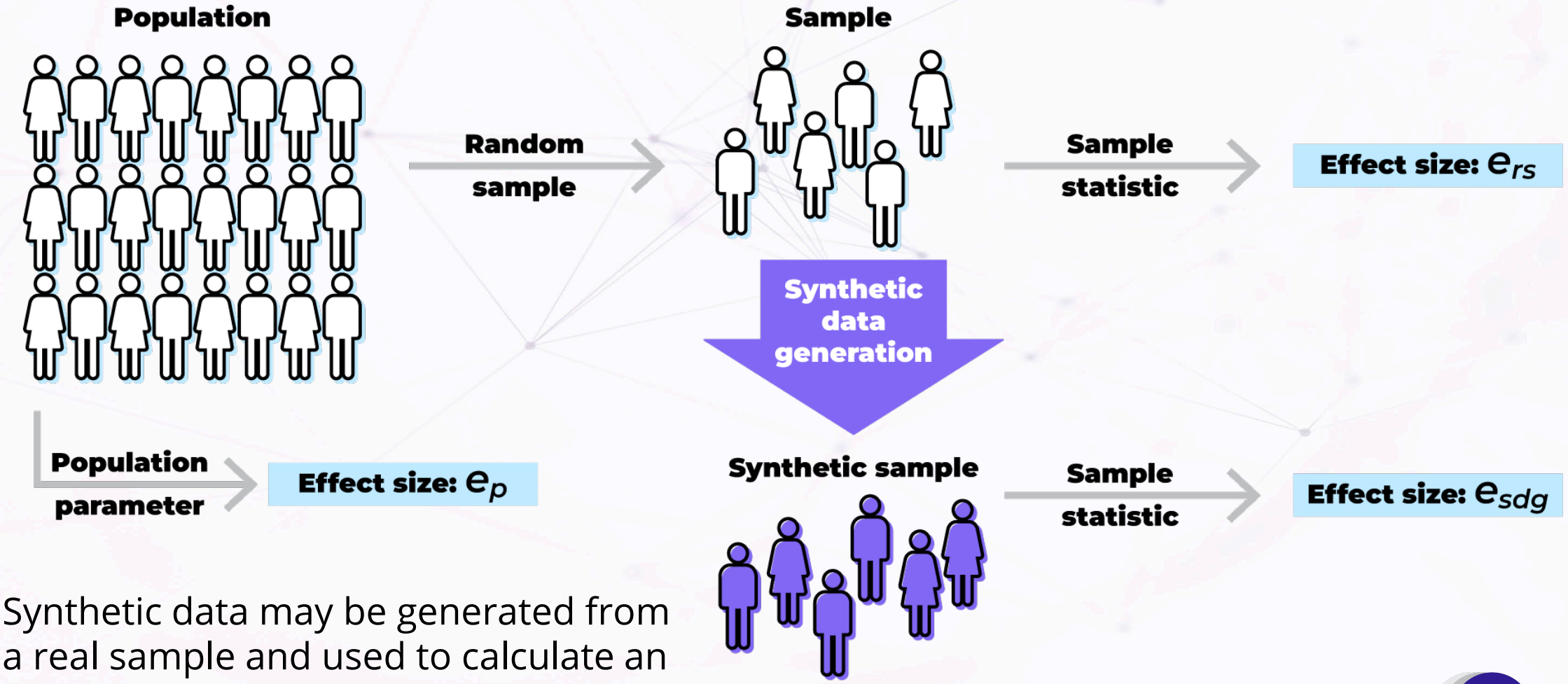
Statistical inference and synthetic data



Synthetic data generation may be generated from a real sample

Replica Analytics
AN AETION COMPANY

Statistical inference and synthetic data



Synthetic data may be generated from a real sample and used to calculate an effect size e_{sdg}

Replica Analytics
AN AETION COMPANY

Statistical inference and synthetic data

There are two perspectives on analysis using synthetic data:

1. **Reproducing** real results: aims to see $e_{sdg} \simeq e_{rs}$
2. Making **inferences** about the underlying population: $e_{sdg} \simeq e_p$

One of the enduring questions is whether **synthetic datasets are a good proxy for real data for analysis purposes**, while simultaneously addressing privacy concerns

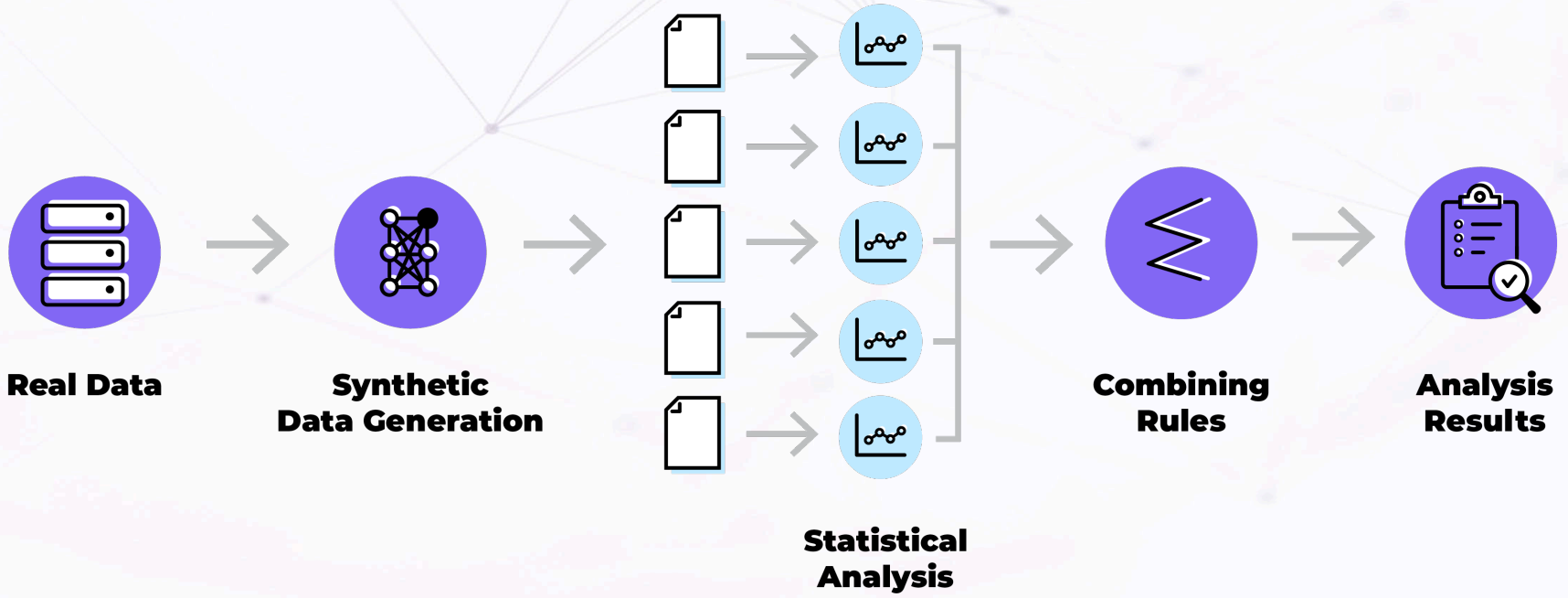
Our presentation today will cover both perspectives to address:

- Whether valid inferences can be made from synthetic data
- Understand the parameters behind such valid inferences

Replica Analytics
AN AETION COMPANY

Because synthesis introduced additional variation, this needs to be accounted for in models to get valid estimates

This means that it is necessary to take a multiple imputation approach to account for this additional variability



Replica Analytics
AN AETION COMPANY

Different types of analyses for synthetic data

- There are four different approaches to evaluating the analysis results from synthetic data

	No Multiple Imputation	With Multiple Imputation
Reproducibility	?	?
Inferences	?	?

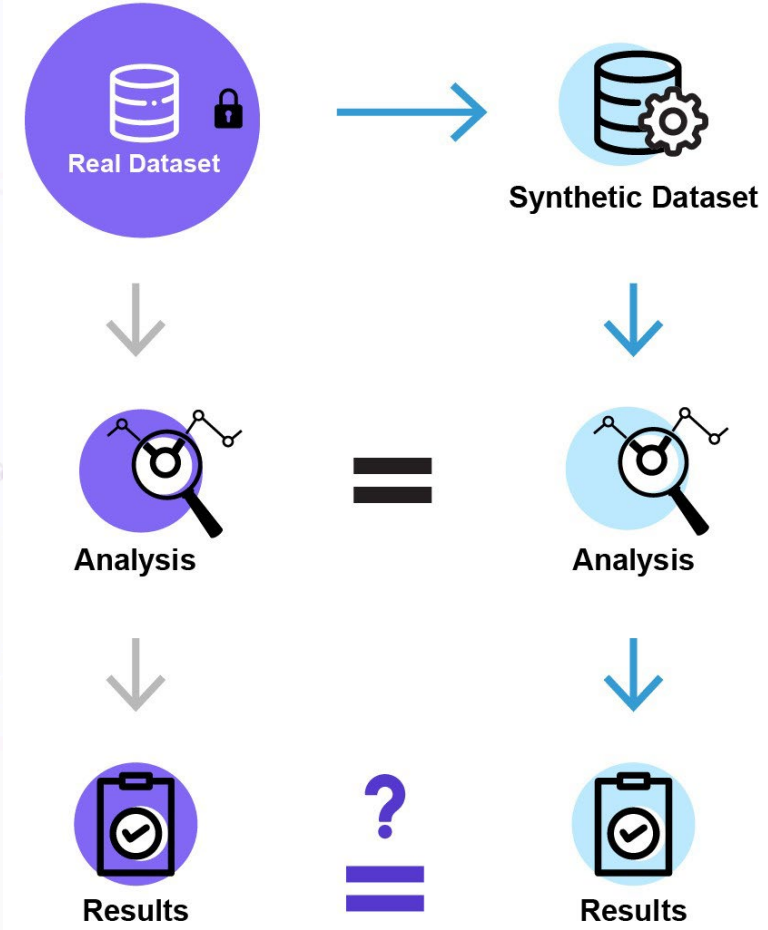
Replica Analytics
AN AETION COMPANY

Case study on reproducibility

Replica Analytics

AN AETION COMPANY

Can synthetic data reproduce real data analysis results?



Replica Analytics
AN AETION COMPANY

Case study: reproducing analysis

Contents lists available at ScienceDirect

Contraception

journal homepage: www.elsevier.com/locate/contraception

Prescription opioid fills following surgical abortion ☆

Liza R. Gibbs[a], Julia A. Pisc[a], Kari P. Braaten[b,c], Brian T. Bateman[d,e], Elizabeth M. Garry[a,*]

[a] Science, Aetion, Inc. Boston, MA United States
[b] Department of Obstetrics and Gynecology, Brigham and Women's Hospital; Boston, MA United States
[c] Planned Parenthood League of Massachusetts; Boston, MA United States
[d] Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School; Boston, MA United States
[e] Obstetric Anesthesia, Department of Anesthesiology, Perioperative and Pain Medicine, Brigham and Women's Hospital and Harvard Medical School; Boston, MA United States

Analysis using commercial insurance data to assess rate and predictors of opioid prescription fills following surgical abortion

Utilized data from 28,252 individuals with recorded surgical abortions

Replica Analytics
AN AETION COMPANY

Methods

- Synthesized m =10 copies of the real dataset using sequential synthesis
- Produced univariate and multivariate logistic regression models using a single synthetic data (unadjusted) or all m=10 datasets (adjusted)

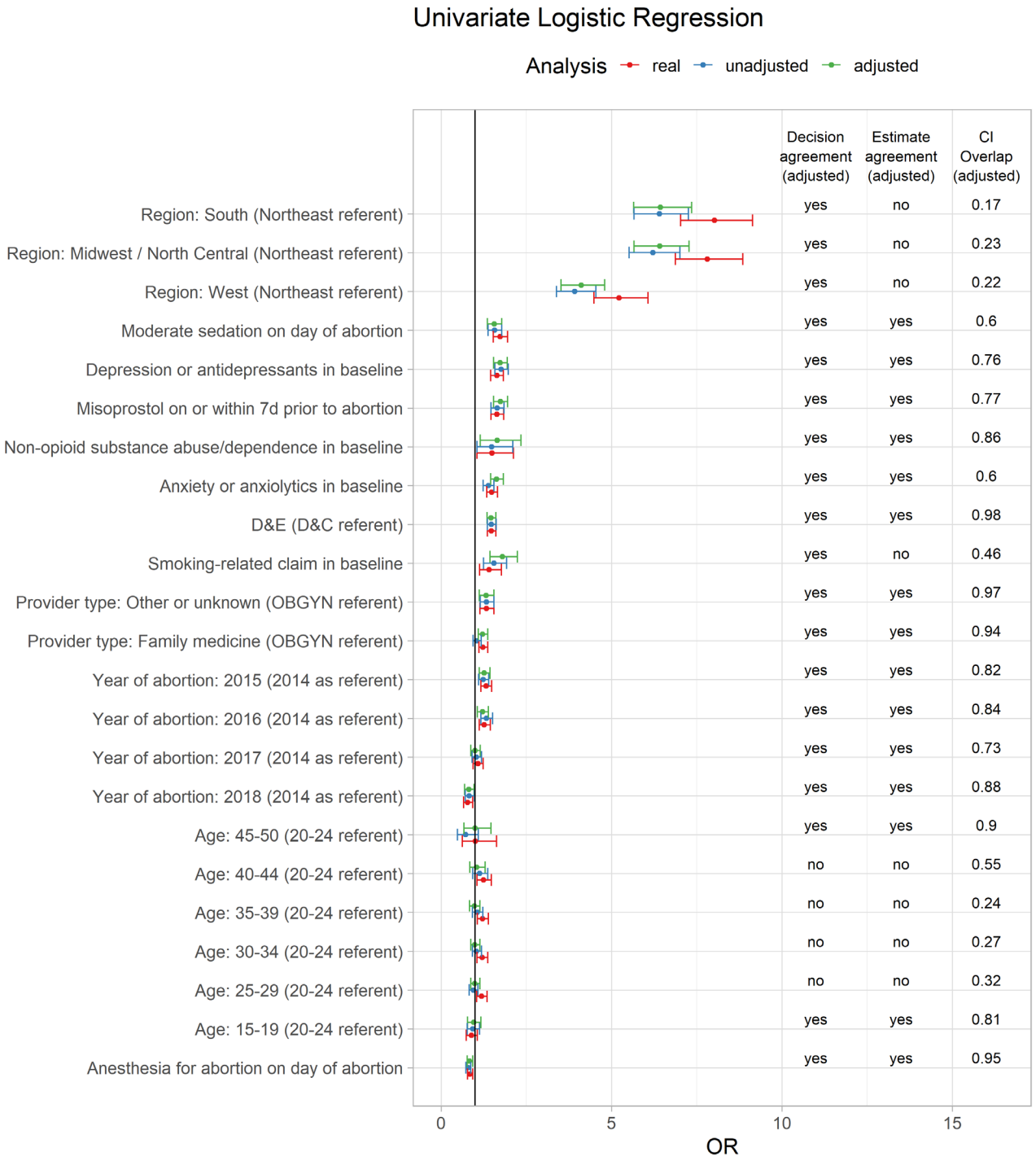
3 metrics to assess how well synthetic data reproduces the findings of the real data:

- **Decision agreement** (are the same conclusions drawn from the synthetic data?)
- **Estimate agreement** (does the estimate from the synthetic data fall within the real data CI?)
- **Confidence interval overlap** (extent of overlap of the CI between real and synthetic data)

Replica Analytics
AN AETION COMPANY

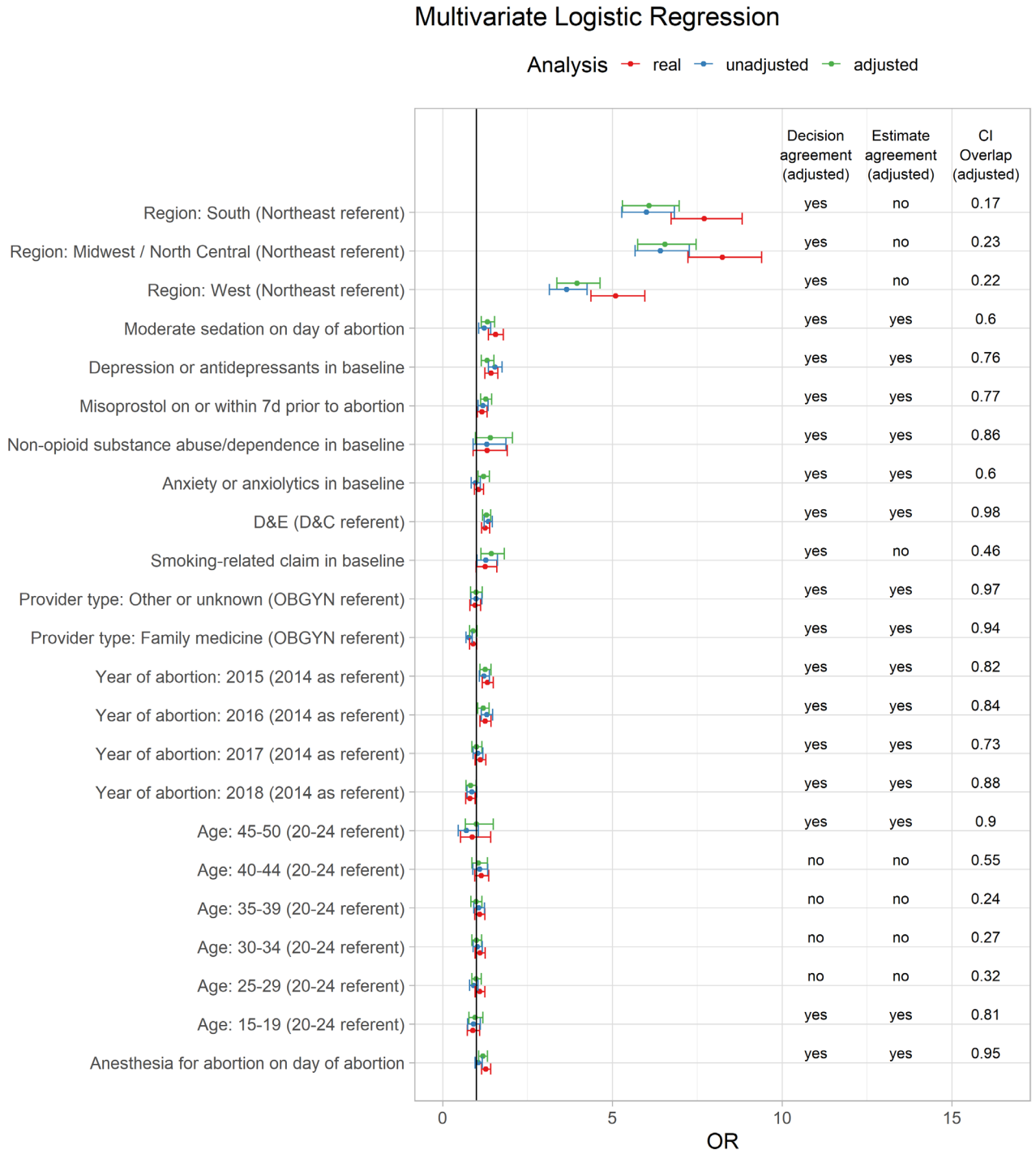
Univariate Results

	Unadjusted	Adjusted
Decision agreement	0.78	0.83
Estimate agreement	0.74	0.65
Average CI overlap	0.63	0.65



Multivariate results

	Unadjusted	Adjusted
Decision agreement	0.69	0.91
Estimate agreement	0.69	0.83
Average CI overlap	0.62	0.67



Conclusions about reproducibility

- The level of agreement and CI overlap to be expected, on average, should be quite high
- Utilizing a multiple imputation approach (rather than single imputation) generally gives better reproducibility results
- These results have also been validated using simulations

Replica Analytics
AN AETION COMPANY

Drawing inferences from synthetic data

Replica Analytics

AN AETION COMPANY

Evaluating the validity of population inferences

- A common way to evaluate the validity of population inferences (through simulations) is to consider:
  - Bias (we want it as close to zero as possible)
  - Coverage (we want it to be close to 95%)
  - Precision (we evaluate this using the empirical standard error, which we want to be small)
  - Power (we want this to be as close to 80% as possible)

Replica Analytics
AN AETION COMPANY

Simulation of population inference

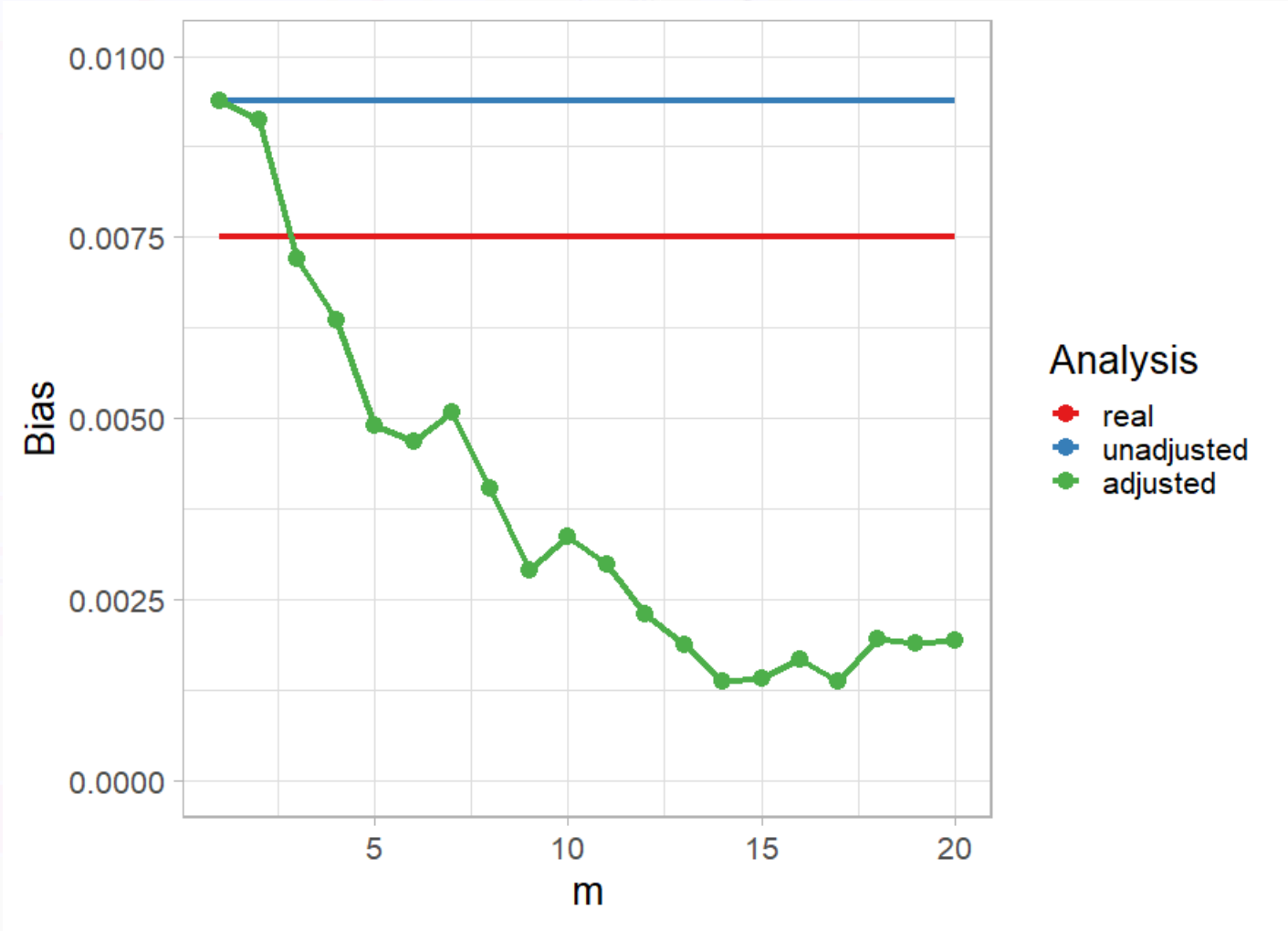
- We performed a simulation on four different datasets to evaluate how well synthetic data can be used to make population inferences

Dataset	Description	n
N0147	Colon cancer clinical trial; examine the relationship between bowel obstruction and overall survival	1,543
CCHS	Canadian Community Health Survey; impact of sex on cardiovascular health	63,522
Danish surgery	Danish colon cancer surgery registry; examine the relationship between age and medical complications from surgery	12,855
COVID-19 Data	Testing data for COVID-19 testing; impact of sex	4,150

Replica Analytics
AN AETION COMPANY

Bias results

(sequential synthesis w/ N0147 trial)

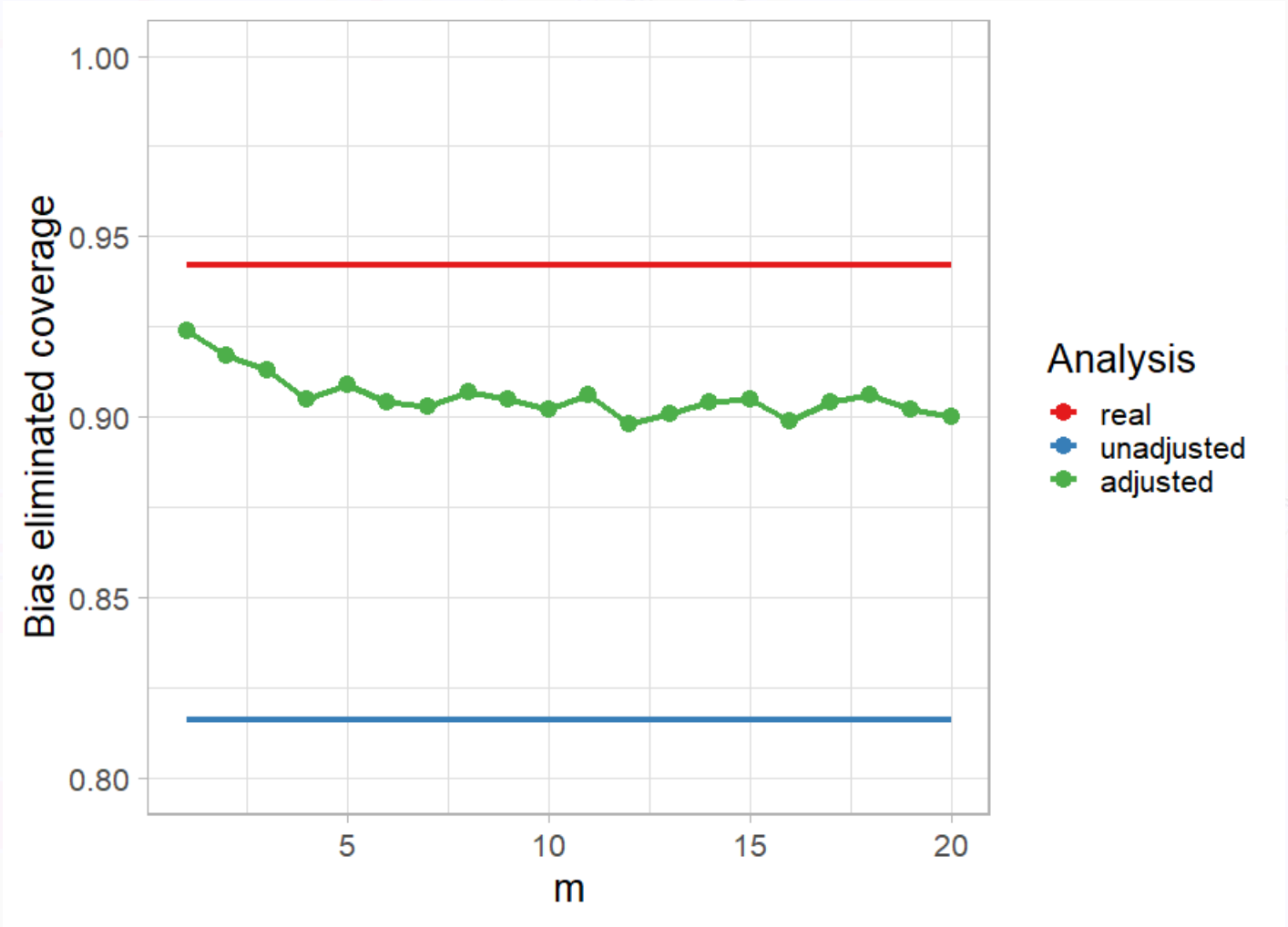


Replica Analytics

AN AETION COMPANY

Bias eliminated coverage

(sequential synthesis w/ N0147 trial)

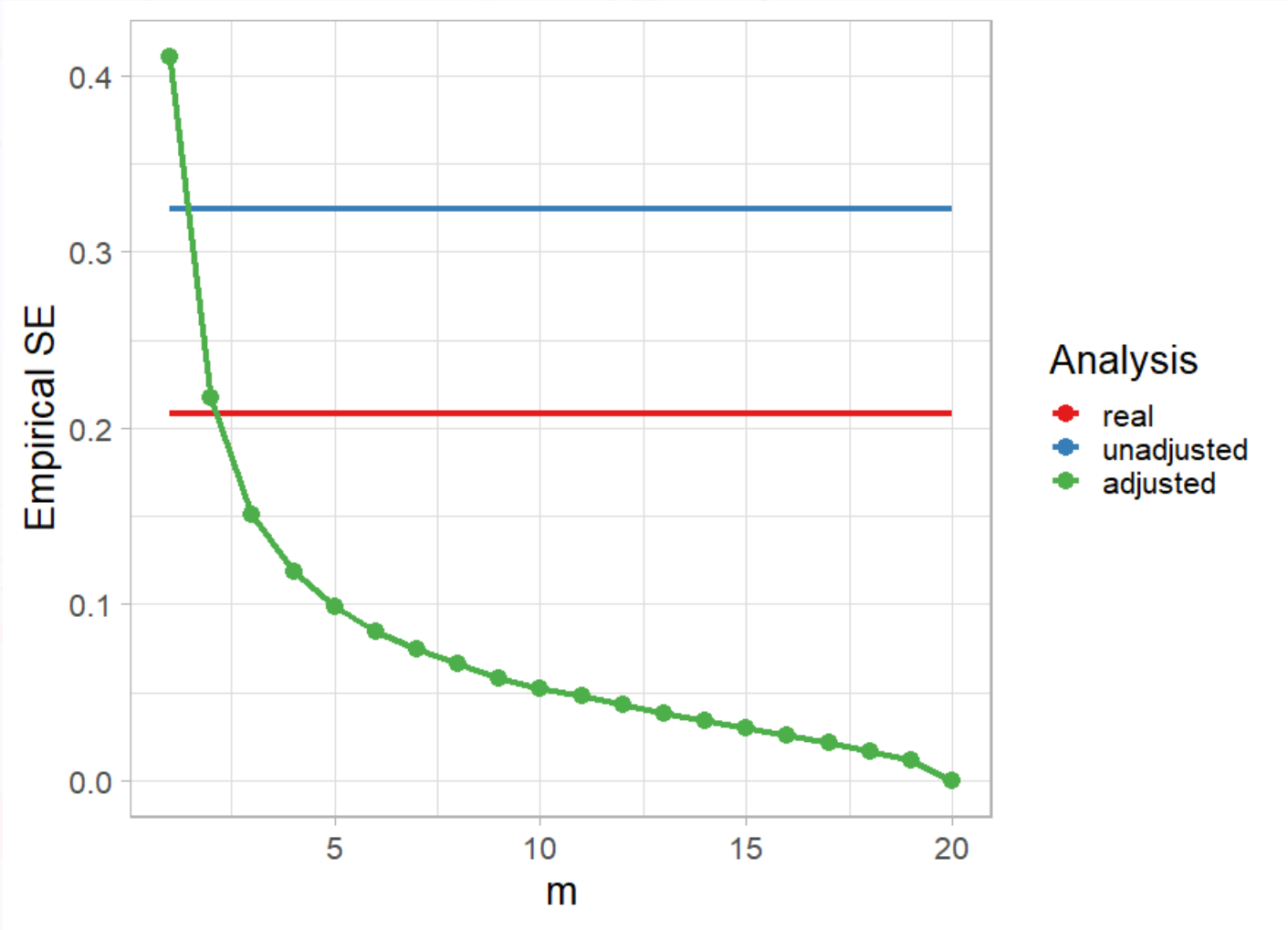


Replica Analytics

AN AETION COMPANY

Empirical SE

(sequential synthesis w/ N0147 trial)

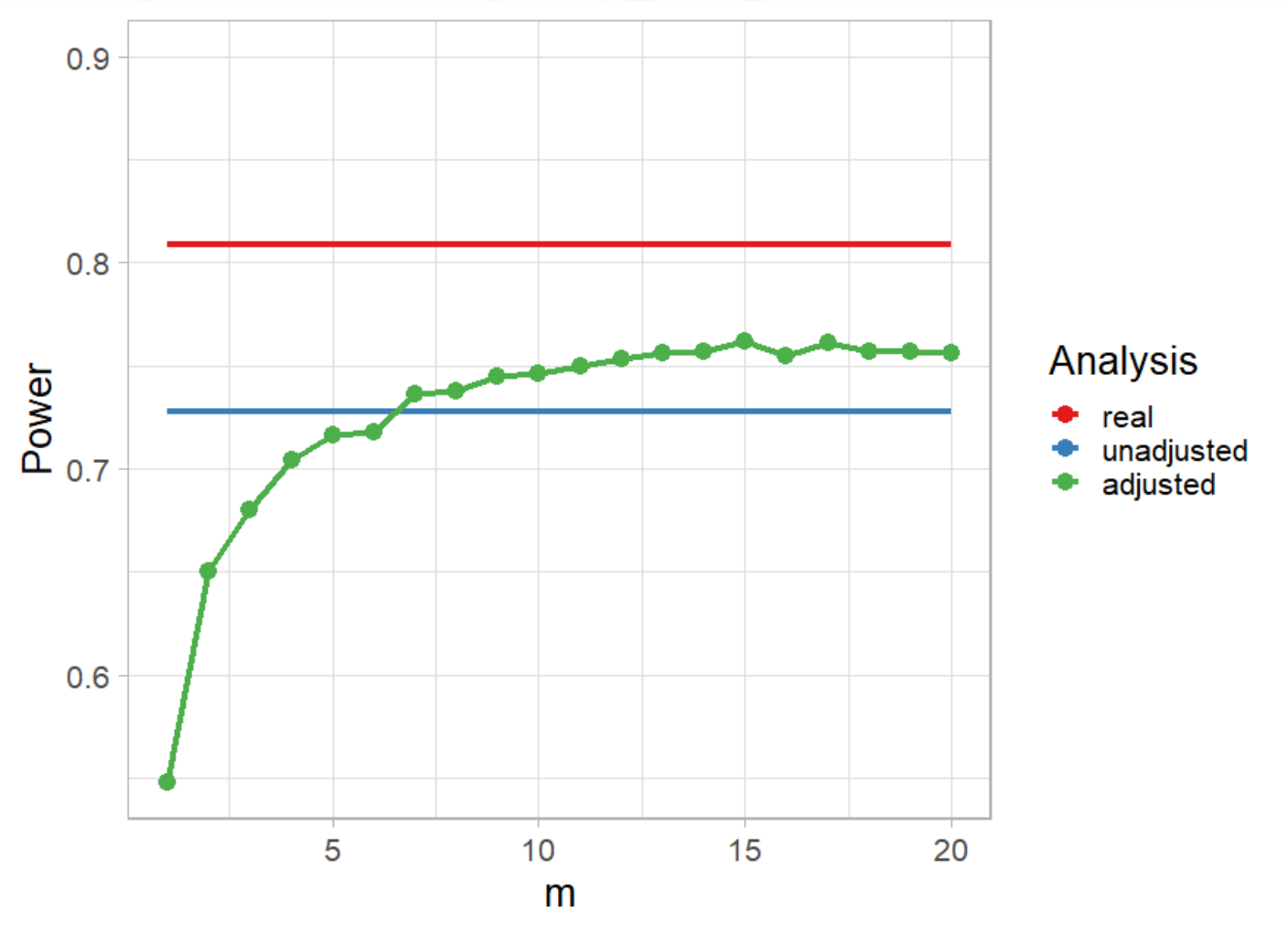


Replica Analytics

AN AETION COMPANY

Power

(sequential synthesis w/ N0147 trial)



Replica Analytics

AN AETION COMPANY

Conclusions about inferences

- Using a multiple imputation approach can result in valid inferences from synthetic datasets for sequential synthesis generative models
- An appropriate parameter is m=10
- Data amplification in this context only provides a marginal benefit
- Inferences without multiple imputation can often have low validity

Replica Analytics
AN AETION COMPANY

What have we learned ?

Replica Analytics

AN AETION COMPANY

What have we learned ?

- Both approaches (reproducibility and inferences) to analyses using synthetic data are reasonable
- The results are not uniform across generative models – it is important to evaluate the validity of inferences for different types of generative models

	No Multiple Imputation	With Multiple Imputation
Reproducibility	✗	✓
Inferences	✗	✓

Replica Analytics
AN AETION COMPANY

Limitations

- We performed the simulations on four datasets which represent a limited set of possible effect sizes, causal relationships, and degrees of confounding
- These results may not apply exactly to machine learning problems where the benefits of data amplification may be more substantive, and synthetic datasets with m=1 may still provide high prognostic accuracy; also a primary criterion is generalizability which would be evaluated differently

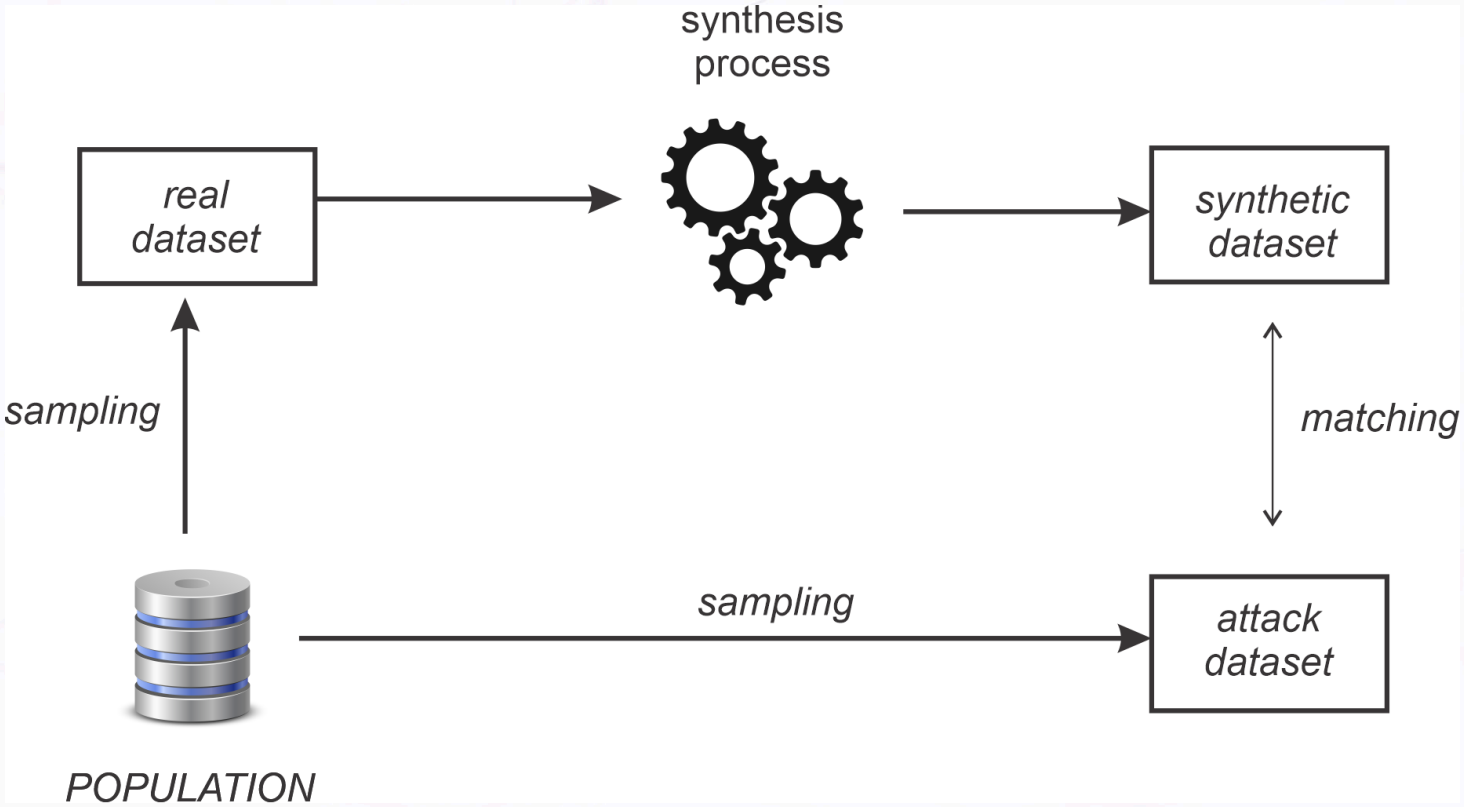
Replica Analytics
AN AETION COMPANY

Evaluating synthetic data privacy

Replica Analytics

AN AETION COMPANY

Privacy risk as membership disclosure

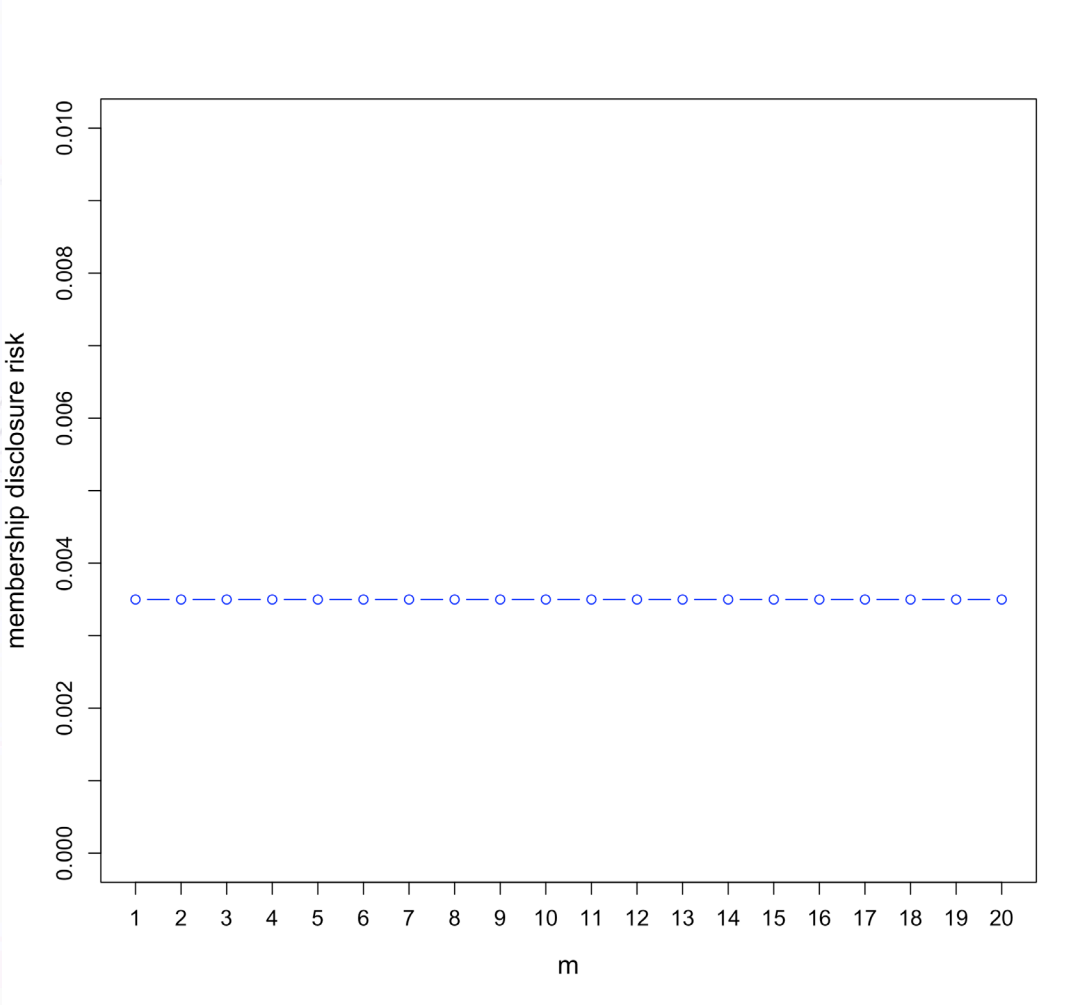


El Emam, K, Mosquera L, Fang, X. Validating A membership disclosure metric For synthetic health data. *JAMIA Open*. 2022; in press.

Replica Analytics
AN AETION COMPANY

Privacy

(sequential synthesis w/ N0147 trial)

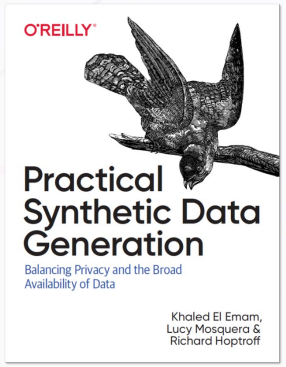


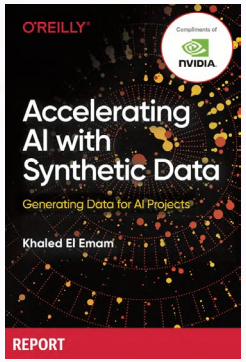
A maximum risk threshold is 0.2, and therefore any values below that are considered low risk.

Replica Analytics
AN AETION COMPANY

To Learn More

- Join our mailing list: https://bit.ly/3gRVAIi
- Follow us on Linkedin: https://bit.ly/2XS3KHF
- Listen to our comprehensive on-line tutorials on data synthesis: https://bit.ly/2TXI0Jy
- Review our detailed knowledgebase of technical articles on synthetic data generation
- Read our introductory report and book on the topic:





Replica Analytics
AN AETION COMPANY

Thank you!

kelemam@replica-analytics.com lmosquera@replica-analytics.com

www.replica-analytics.com

Replica Analytics

AN AETION COMPANY