SYNTHETIC DATA GENERATION
FOR RARE DISEASE RESEARCH

OCT. 27, 2021 | 11AM EDT

Presented by

Dr. Khaled El Emam,
CEO, Replica Analytics

Jason Colquitt, CEO
Across Healthcare

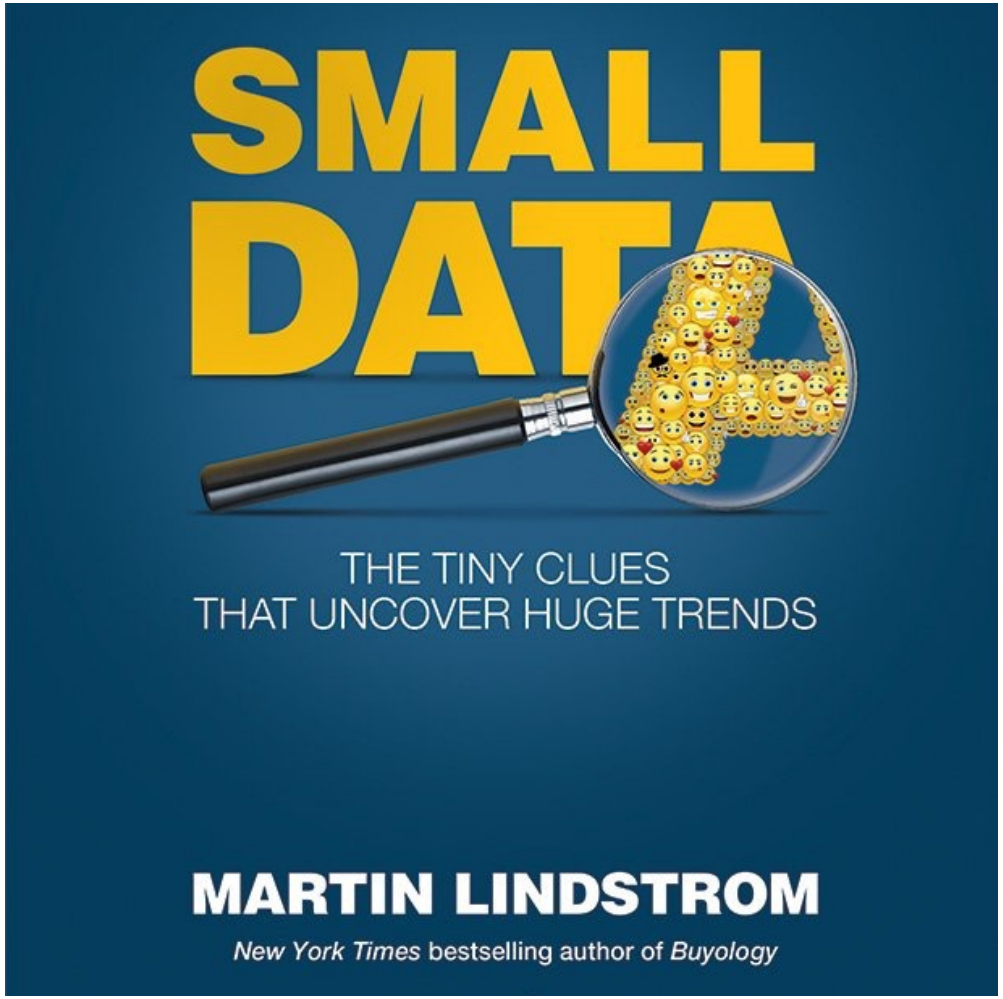
Agenda

Time	Speaker	Topic
11:00 – 11:05	Ash Kamath	Logistics & Introduction
11:05 – 11:20	Jason Colquitt	The challenges with getting access to rare disease data
11:20 – 11:40	Khaled El Emam	An overview of how synthetic data generation can solve some of the data access challenges with rare disease datasets
11:40 – 12:00	Ash Kamath	Q&A

Some Logistics

- The webinar will be recorded and we will post the recording on our web site afterwards
- You can use the chat function and the Q&A function to type your questions during or at the end of the presentations
- We expect to have a series of presentations and reports on the topic of small datasets in the next few months, so we will also keep you updated

Synthetic Data Generation for Rare Disease Research



While the rest of the world is chasing Big Data, in the study of rare diseases, Small Data is our passion.

A great example of passion and small data is Hao-Fountain Syndrome and their Project Artemis. This project launched in 2018 when 34 patients worldwide were known to have this USP7 gene mutation. The goal of the project was to find 66 more patients and get to 100 patients. This last month, three years later, Project Artemis achieved its goal, and 100 patients were found.

Matrix

Across Healthcare

Synthetic Data Generation for Rare Disease Research

Definitions:

- In the US, a rare disease is defined as a health condition that affects fewer than 200,000 individuals.
- 7,000 conditions meet this definition.
- A small number of individuals are affected by each rare disease, the estimated total number of individuals living with any rare disease is between 25 million and 30 million.

Challenges:

The study of rare diseases poses unique challenges. Creative study designs must be sought (examples: adaptive trials and self-controlled study designs). Researchers must be cautious of the analytic challenges (examples: available data representative of the entire population and sufficient statistical power).

Matrix

Across Healthcare

Synthetic Data Generation for Rare Disease Research

Reference: **Data DIY – Your Involvement in Driving Understanding, Discovery, and Treatments for Rare Disease**

A Global Genes Series supported by the Chan Zuckerberg Initiative

https://globalgenes.org/data-diy/

https://chanzuckerberg.com/

Matrix

Across Healthcare

Synthetic Data Generation for Rare Disease Data

Khaled El Emam
kelemam@replica-analytics.com

27th October 2021

Agenda

Introduction to Synthesis 1 General description of what synthetic data is and basic concepts and techniques

Data Amplification 2 Basic principles of data amplification

Example 3 Case study on a colon cancer clinical research study

Replica Analytics

An overview of synthetic data generation

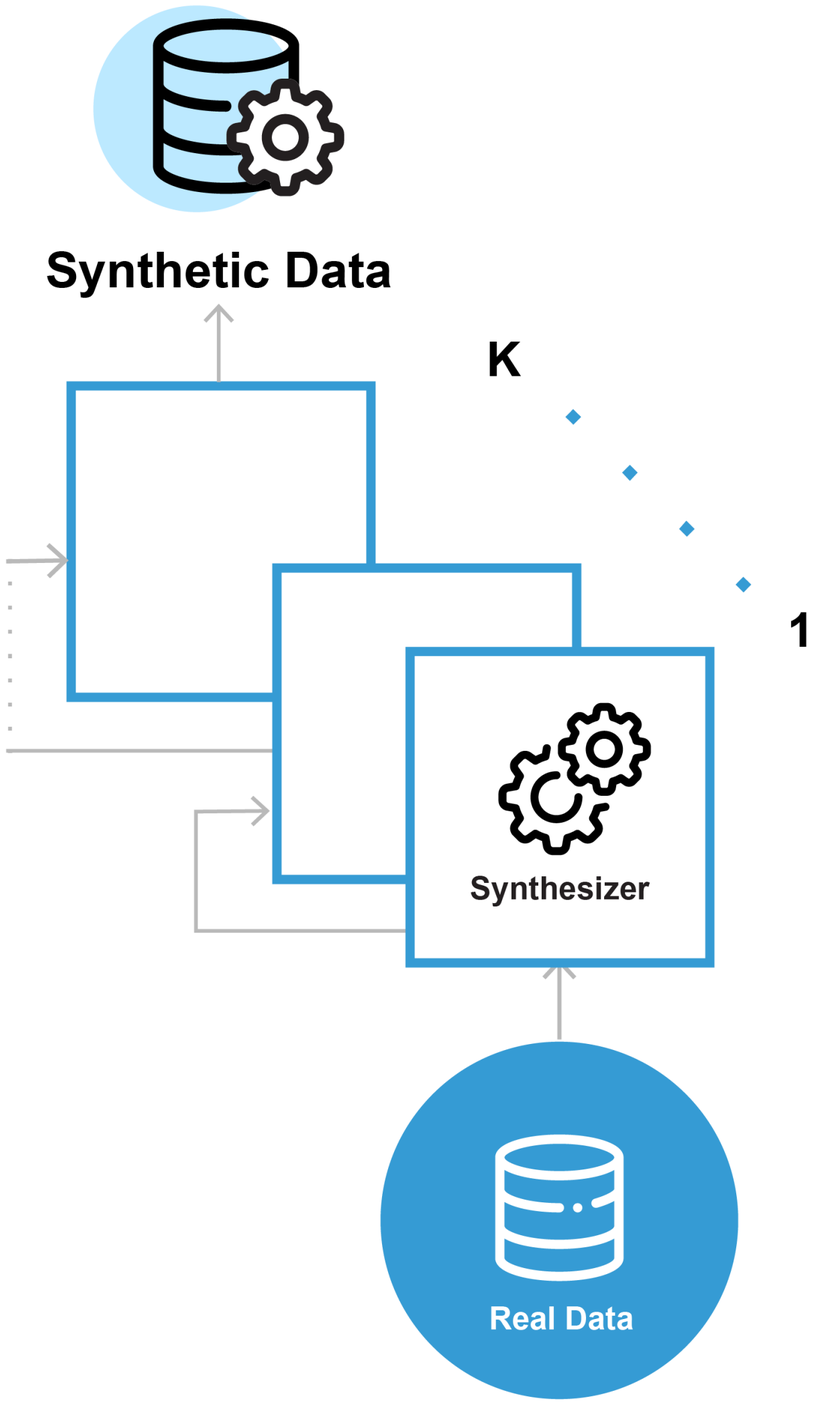
COU1A	AGECAT	AGELE70	WHITE	MALE	BMI
United States	2	1	1	1	33.75155
United States	2	1	1	0	39.24707
United States	1	1	1	0	26.5625
United States	4	1	1	1	40.58273
United States	5	0	0	1	24.42046
United States	5	0	1	0	19.07124
United States	3	1	1	1	26.04938
United States	4	1	1	1	25.46939

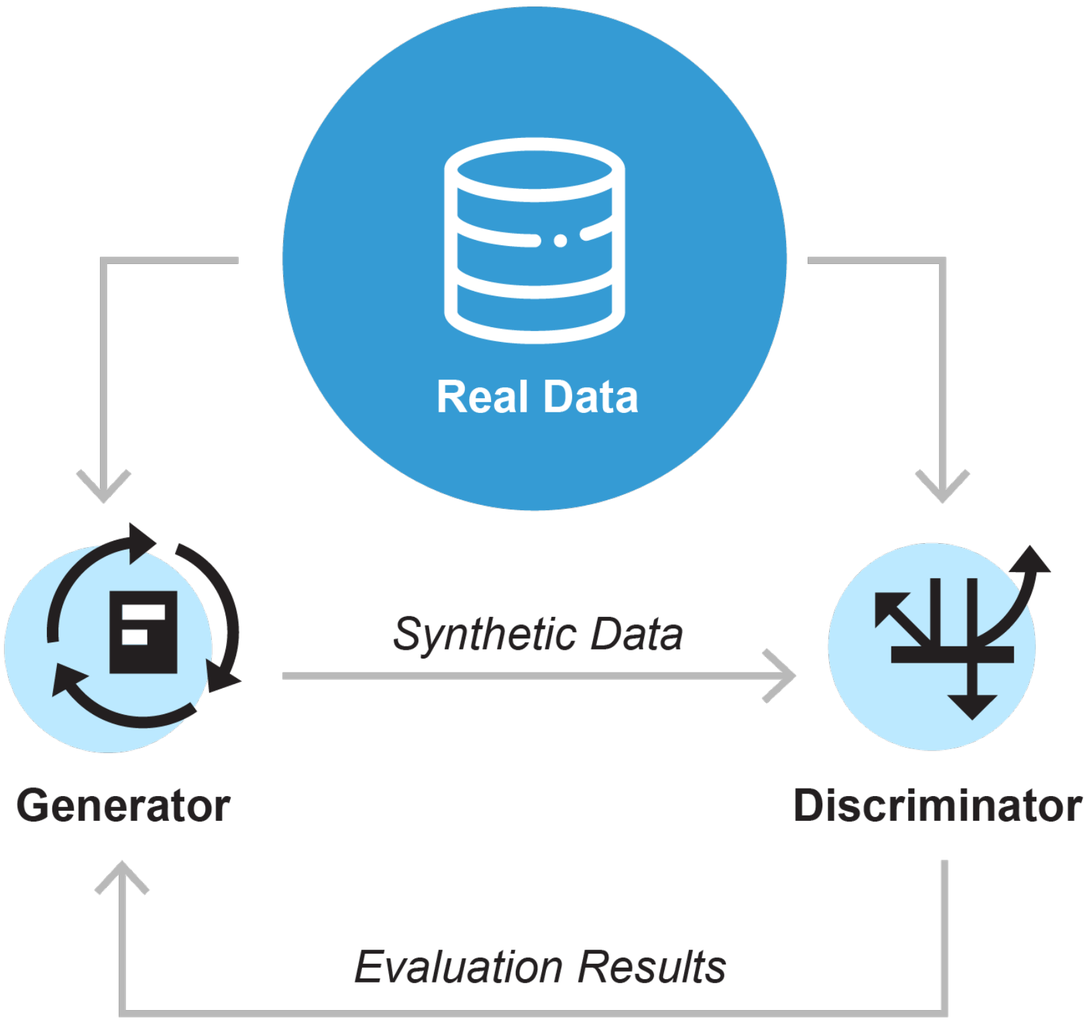
Additional Clarifications

- The source datasets can be relatively small. We have developed generative modeling techniques that will work for small datasets.
- The source datasets can also be very large – then it becomes a function of the 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.

Replica Analytics

Sequential synthesis is a good way to generate synthetic data from small source data

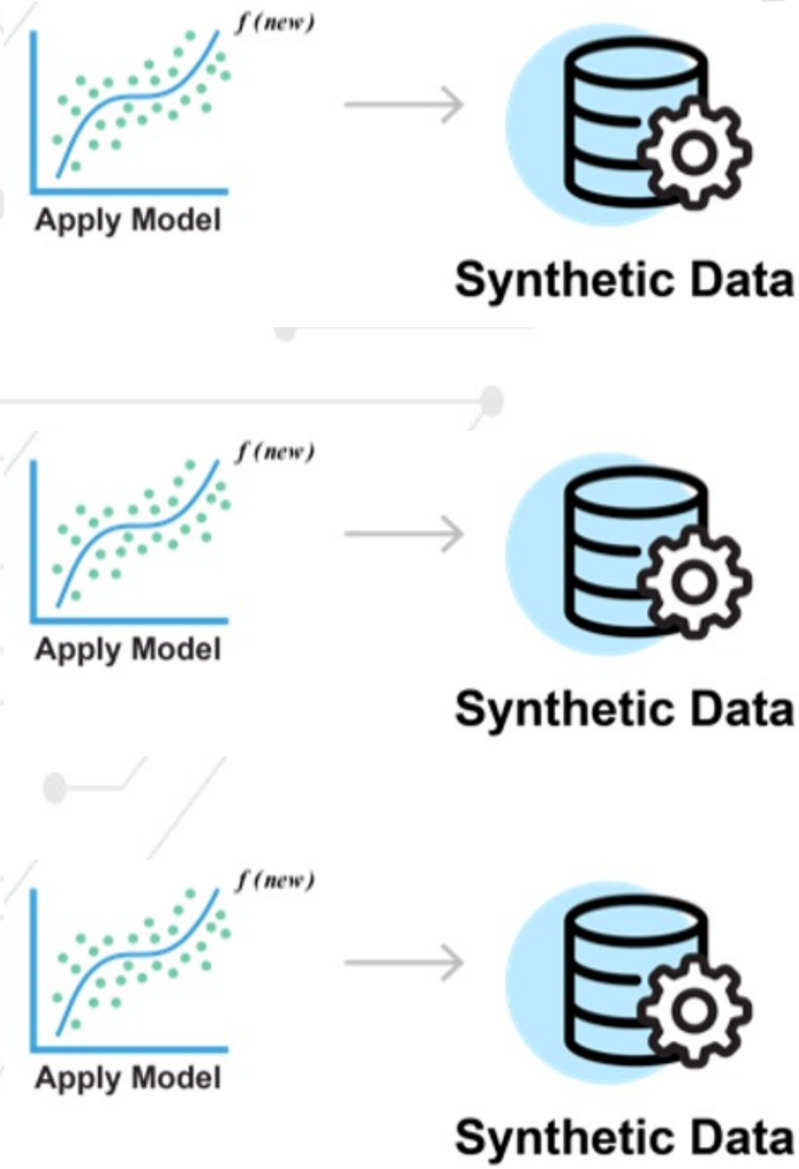


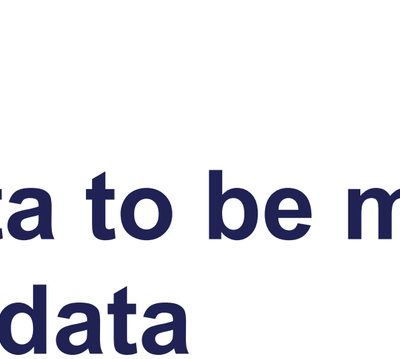


JAMIA
A SCHOLARLY JOURNAL OF INFORMATICS IN HEALTH AND BIOMEDICINE

Replica Analytics

A simulator exchange allows data to be made available without sharing actual data





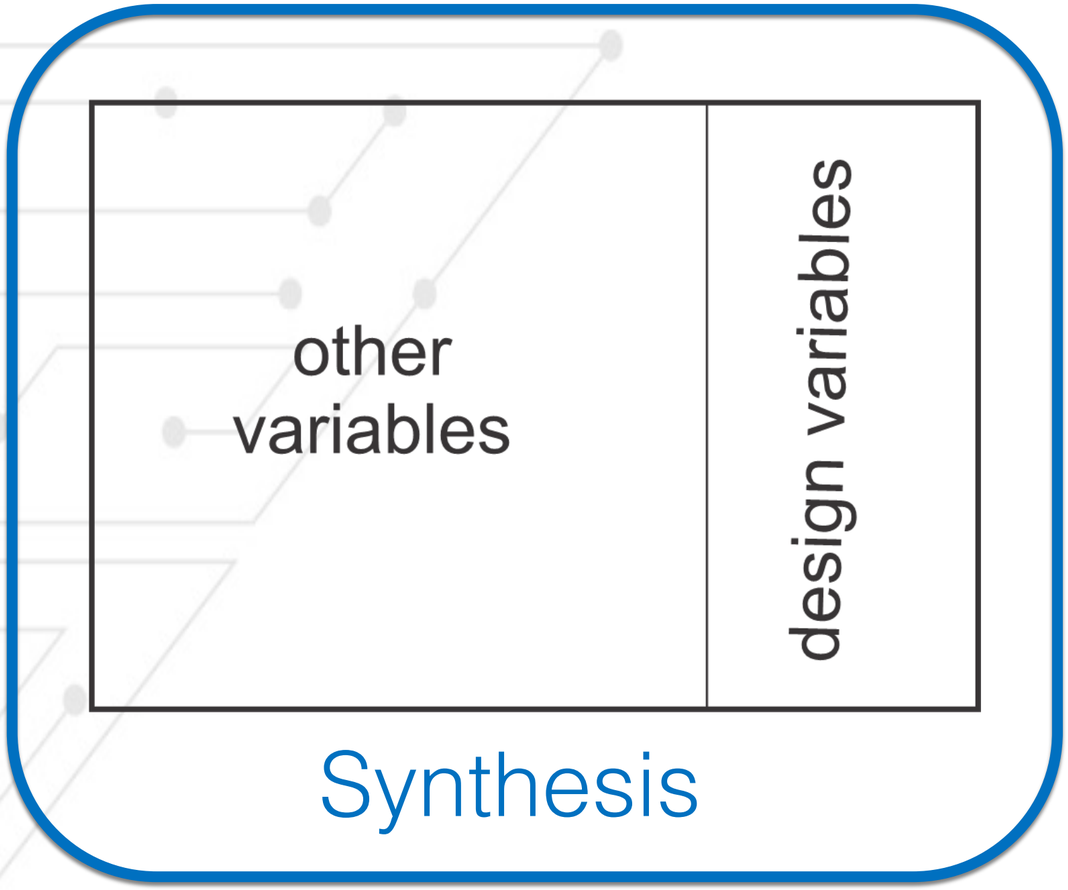
Additional Clarifications

- The simulators would not be given to the data consumers – they would only have access to them through an interface.
- This access would be monitored and throttled to reduce the risk of attacks on the models.
- Data consumers would also need to agree to terms of use around the access to the simulators.

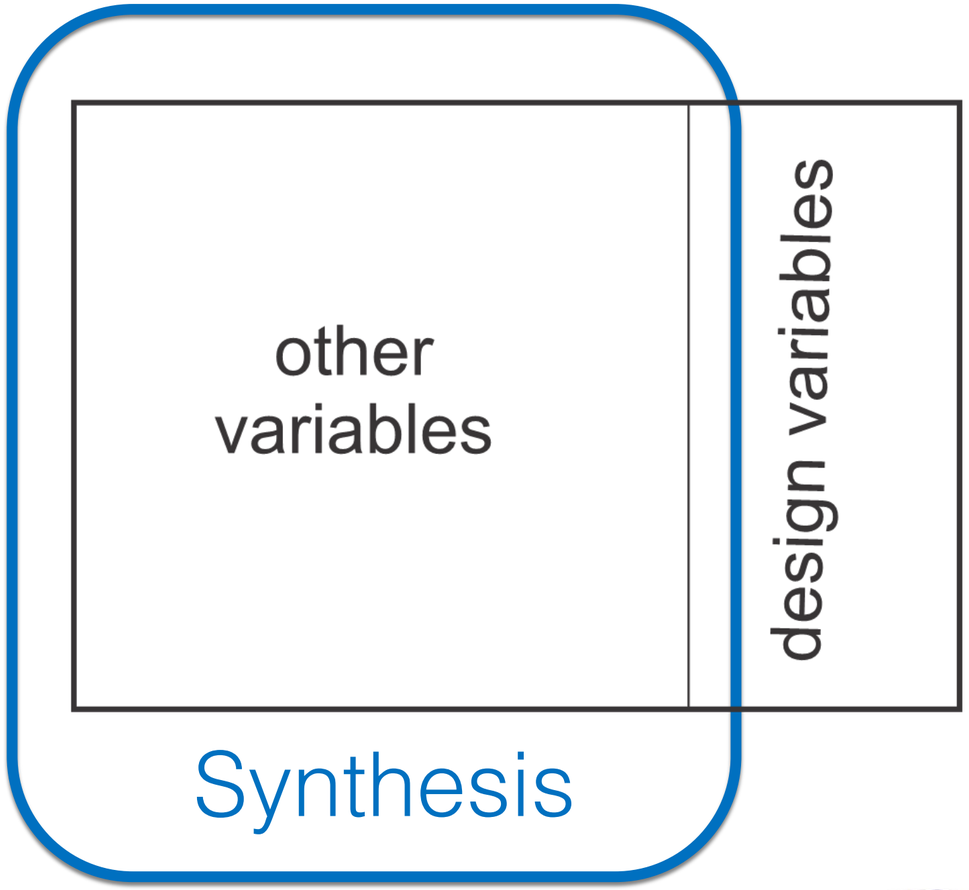
Replica Analytics

Two Synthesis Strategies

Full Synthesis
Synthesize all variables

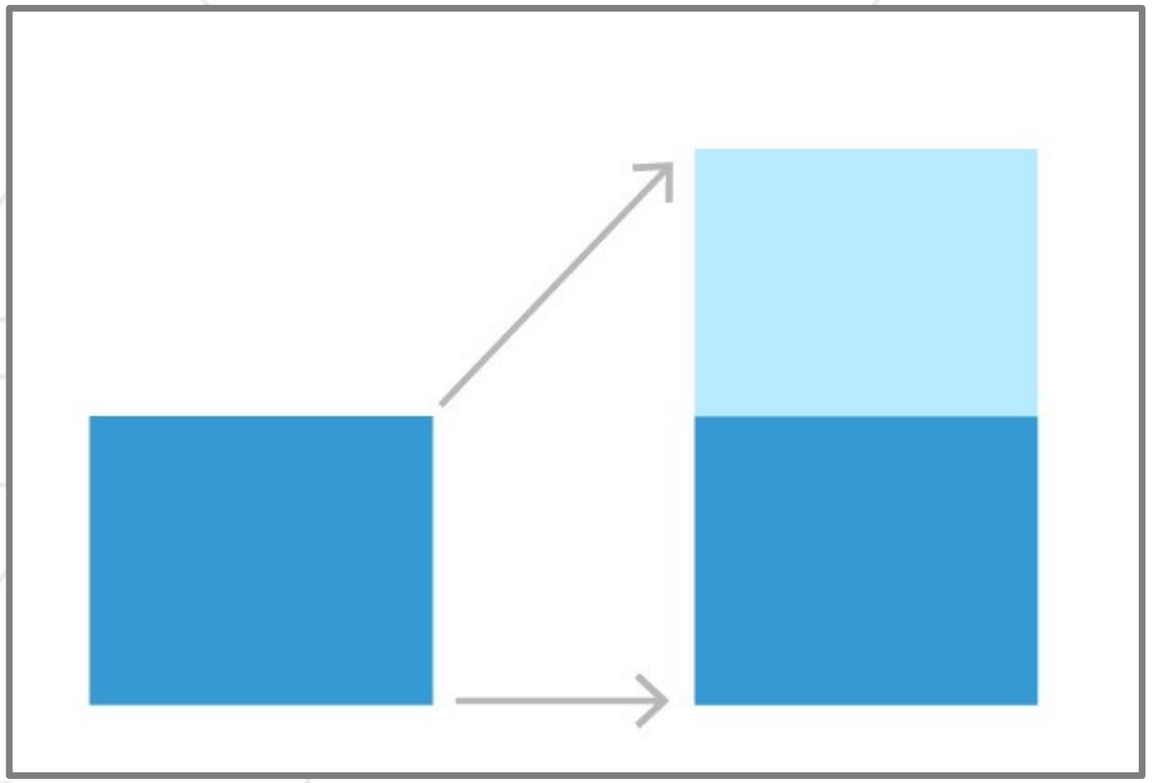


Partial Synthesis
Fix design variables

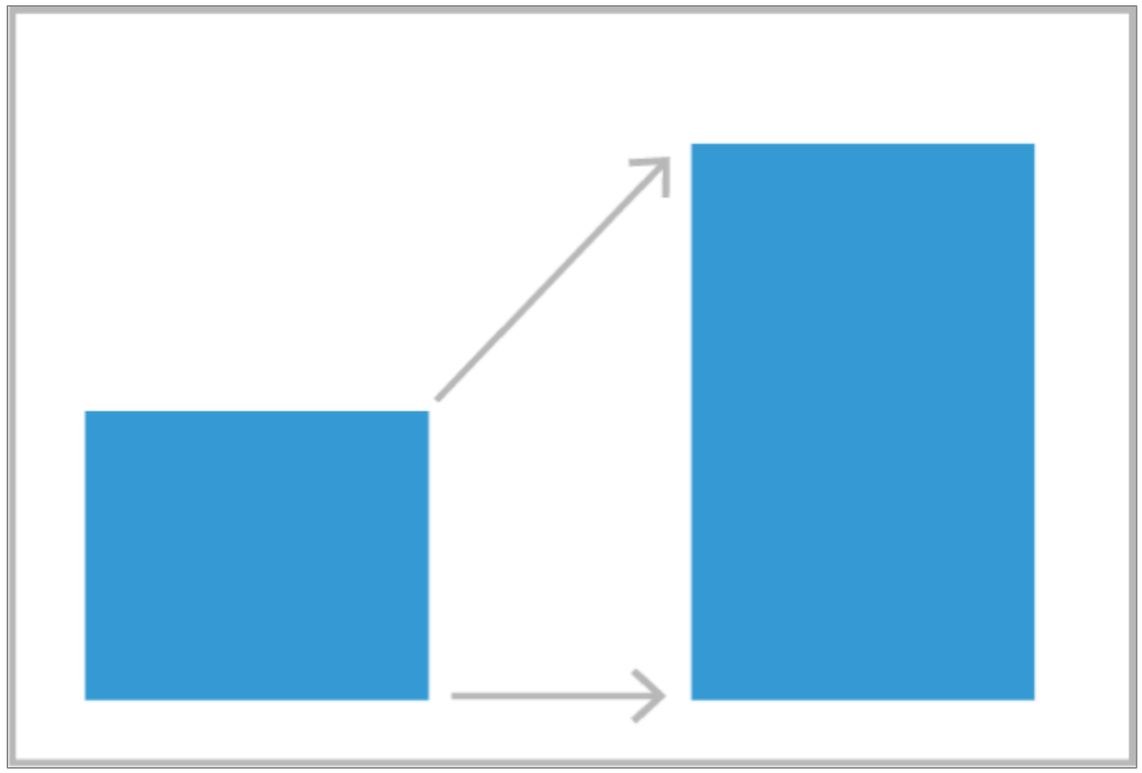


Replica Analytics

Data augmentation vs data amplification – two different approaches for getting more data



(a)



(b)

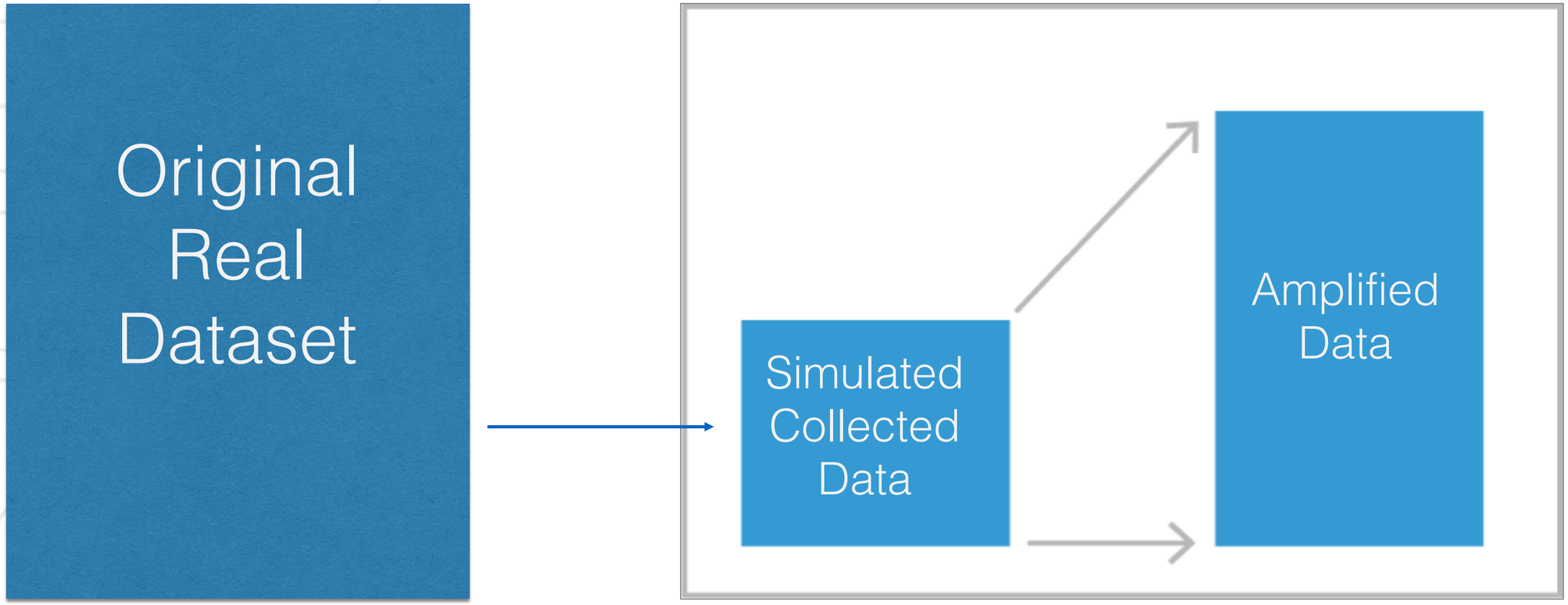
Replica Analytics

We are embarking on an effort to develop and validate synthetic data methodologies for amplifying small datasets

- While we have quite a bit of experience with synthetic data generation for clinical trial datasets (which are small datasets), there is no requirement to amplify these datasets
- The data amplification requirement becomes more relevant when there are challenges or high resource costs to collect data

Our methodology is a simulation to evaluate how well amplification works

- We have the real target dataset from which we select an example of the dataset that we actually have collected
- The collected dataset is then amplified, and the amplified synthetic data is used for analysis



Replica Analytics

The dataset we will use: N0147 Trial

Effect of cetuximab on survival among patients with resected stage III colon cancer

- A randomized trial 2004 – 2009
- 2,686 adult patients with stage III colon cancer
- Arms:
  - Control: adjuvant regimens of folic acid, fluorouracil, and oxaliplatin / fluorouracil, leucovorin, and irinotecan
  - Treatment: Cetuximab + control regimens

1. Alberts SR, Sargent DJ, Nair S, Mahoney MR, Mooney M, Thibodeau SN, Smyrk TC, Sinicrope FA, Chan E, Gill S, Kahlenberg MS. Effect of oxaliplatin, fluorouracil, and leucovorin with or without cetuximab on survival among patients with resected stage III colon cancer: a randomized trial. Jama. 2012 Apr 4;307(13):1383-93.(NCT00079274)

Replica Analytics

Secondary Analysis of N0147

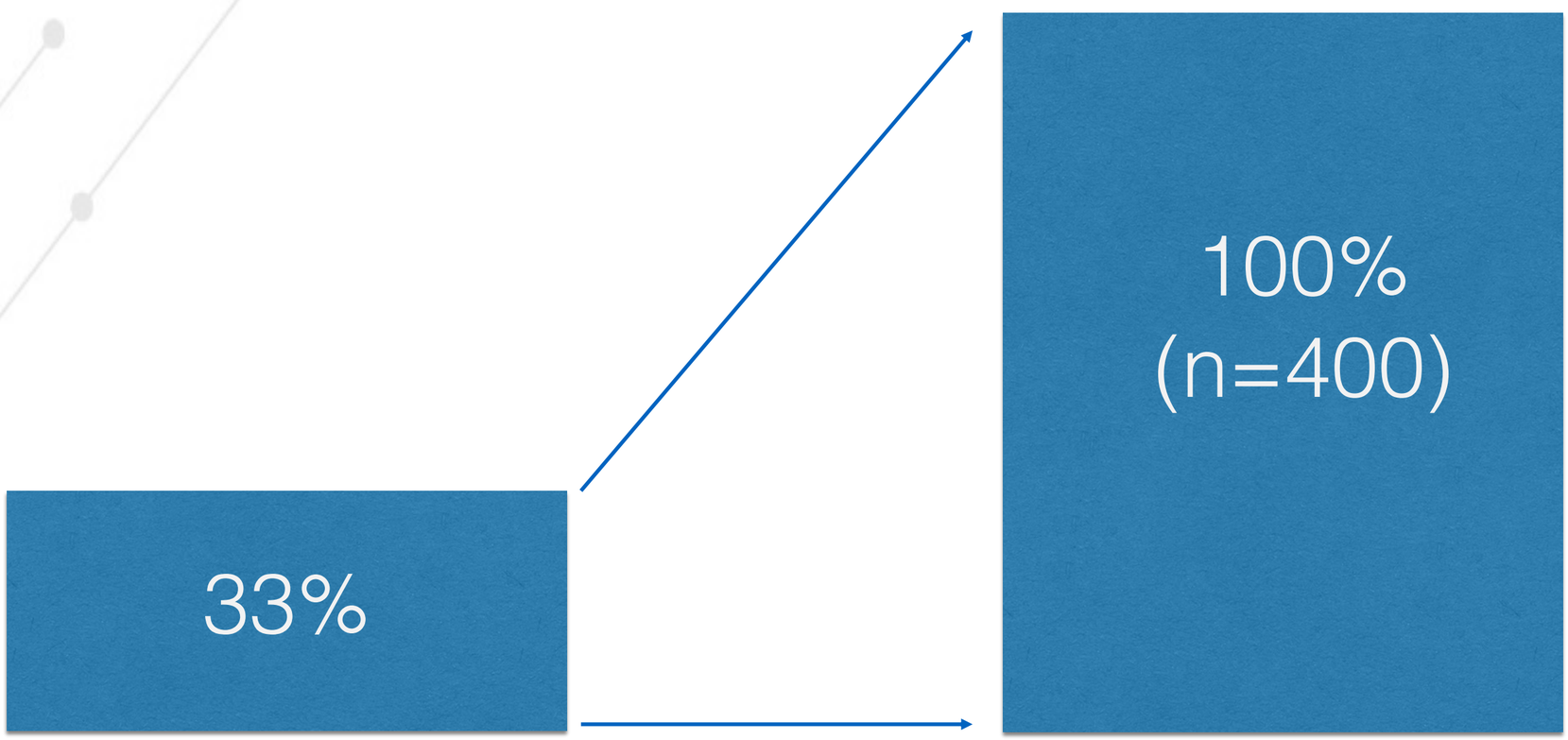
- Published in *Surgery* in 2018
- Only control arm was analyzed: N = 1,543
- Variables:
  - Predictor of interest: Obstruction
  - Outcomes: 5-year disease-free survival (DFS), 5-year overall survival (OS)
  - Covariates:
    - Cancer staging and histology (Clinical T stage, lymph node involvement, histology)
    - Baseline ECOG performance status, KRAS biomarker
    - Demographics and BMI
- Statistical methods: logistic regression models

2. Dahdaleh FS, Sherman SK, Poli EC, Vigneswaran J, Polite BN, Sharma MR, Catenacci DV, Maron SB, Turaga KK. Obstruction predicts worse long-term outcomes in stage III colon cancer: A secondary analysis of the N0147 trial. Surgery. 2018 Dec 1;164(6):1223-9.

Replica Analytics

A 33% sample was taken and then amplified to 100%

- Selected a “true” dataset size of 400 individuals
- Sampled 33% to reflect the actual number of records that could be collected in practice
- Amplified these to 400 individuals and compared to the “true” dataset

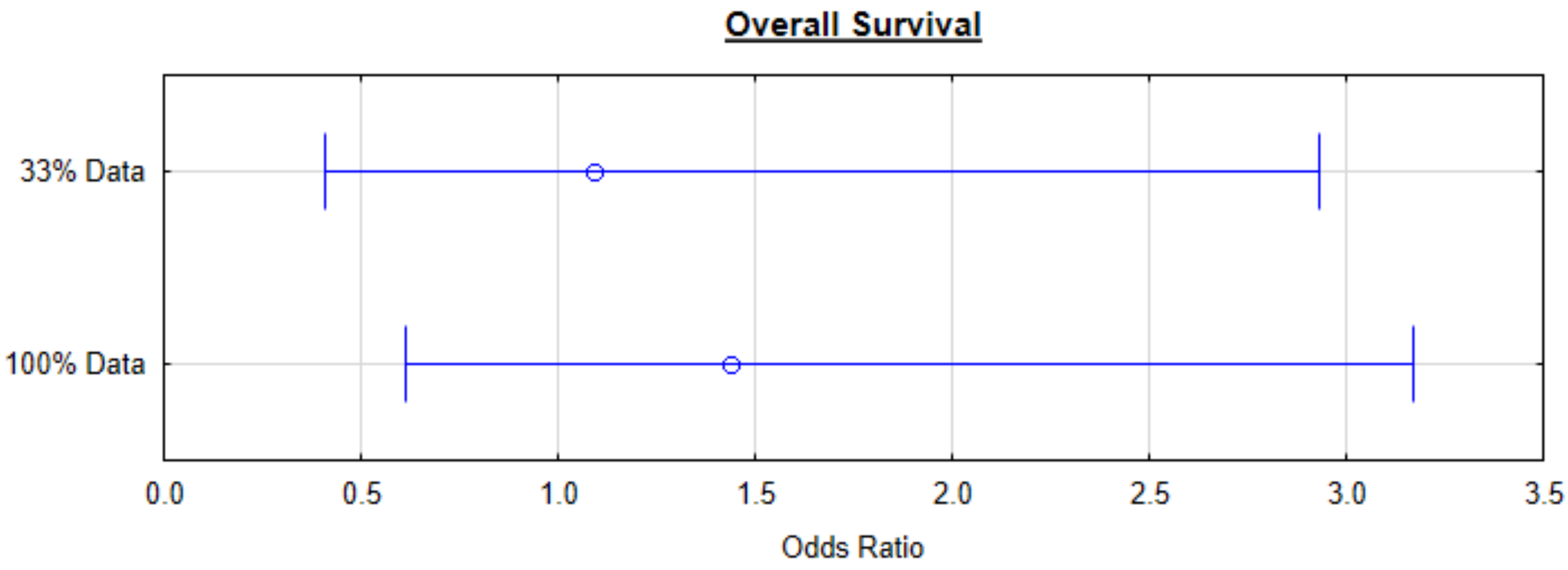


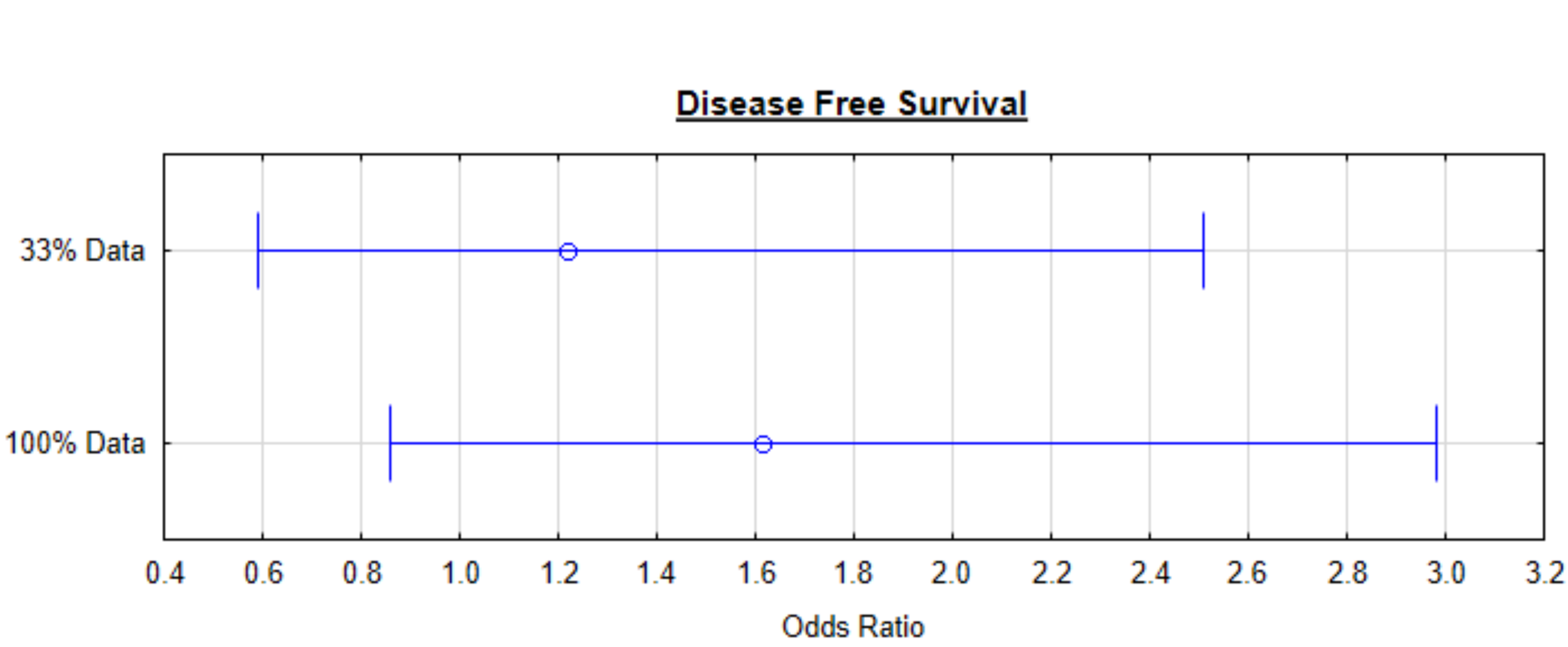
Some considerations

- The originally collected data will be small, and therefore the statistical power of any inferences will likely be low and confidence intervals will be wide
- Even if more data is collected (100% data in our example) the dataset size is still going to be relatively small
- When estimating model parameters with amplified data we still need to take into account the uncertainty introduced due to the stochastic synthesis process

Four criteria for evaluating model results

	Definition
Criterion 1	• Parameter direction is the same
Criterion 2	• Statistical significance the same
Criterion 3	• Synthetic parameter within real 95% CI
Criterion 4	• CI overlap greater than 37.5%


Overall Survival
33% Data
100% Data
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Odds Ratio


Disease Free Survival
33% Data
100% Data
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0
2.2
2.4
2.6
2.8
3.0
3.2
Odds Ratio

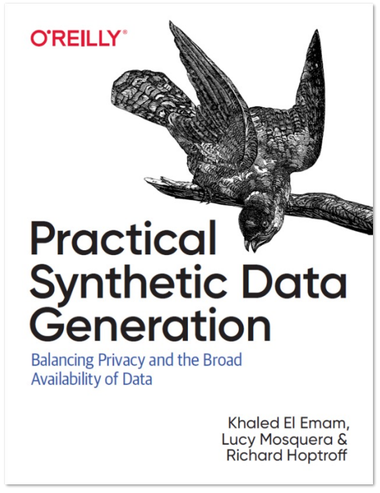
Prediction accuracy

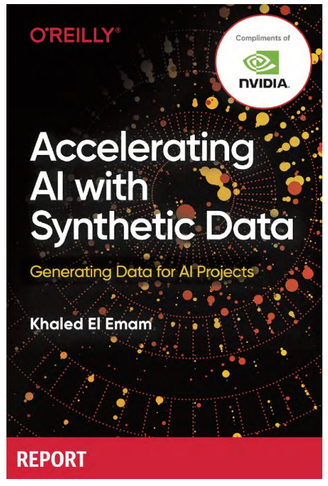
- For the 400 observations we used 5-fold cross-validation to compute the AUROC
- For the synthetic data we used a 35% hold-out from the original

	Real (100%) AUROC	33% Amplified AUROC
OS	0.65	0.624
DFS	0.57	0.57

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
- Read our introductory report and book on the topic





Replica Analytics

QUESTIONS