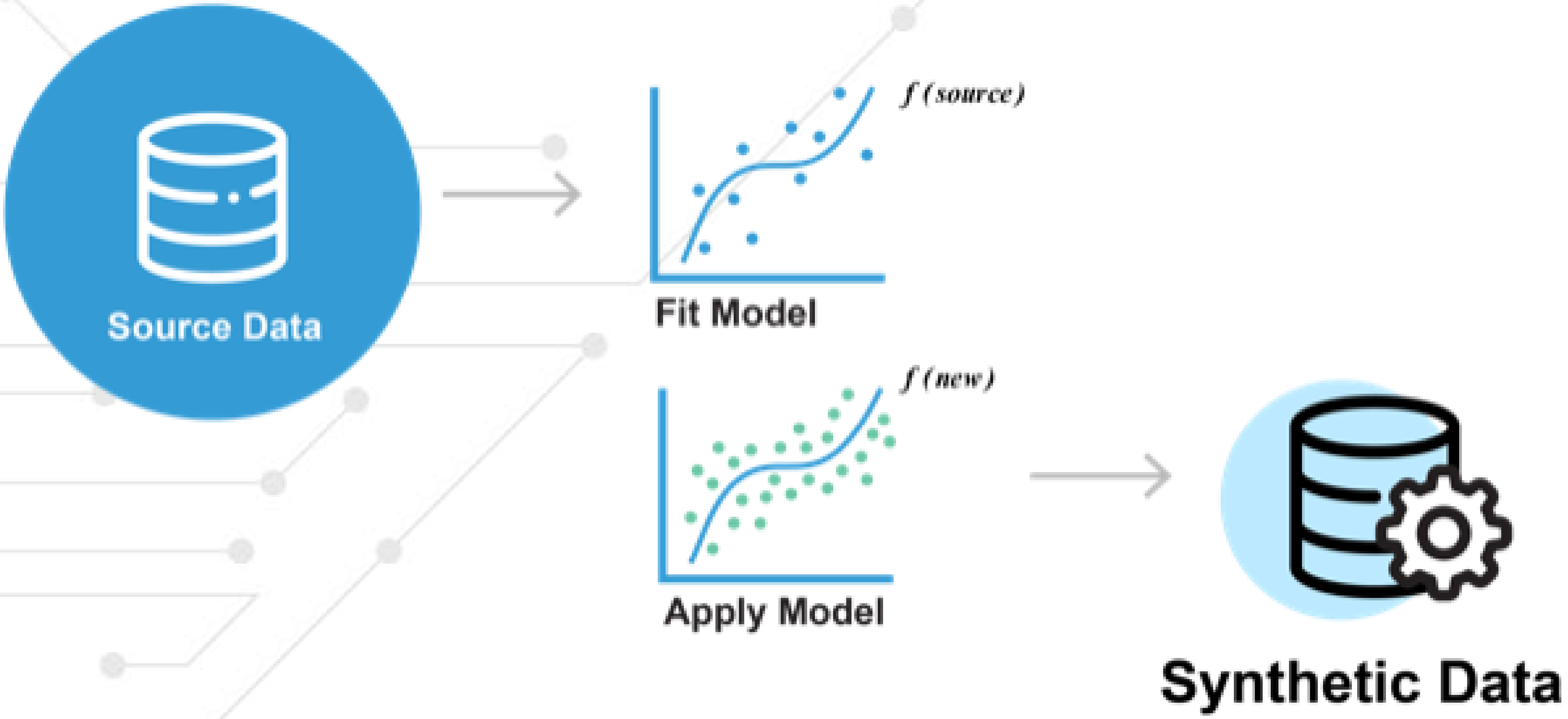
Synthetic Clinical Trial Data

Methods, Practice, and Experience

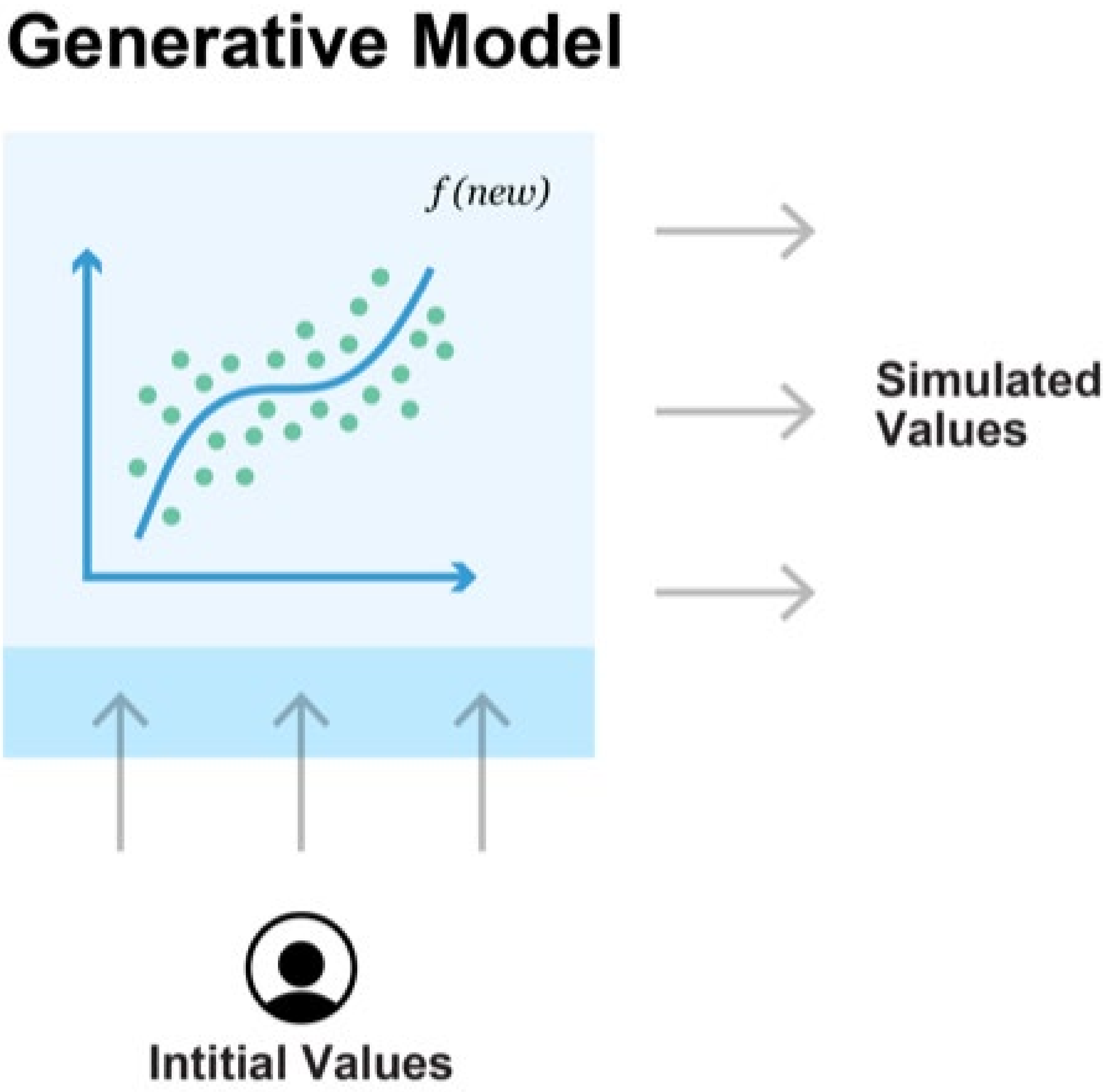
Khaled El Emam

4th June 2020

The Synthesis Process



Synthesis As Simulation

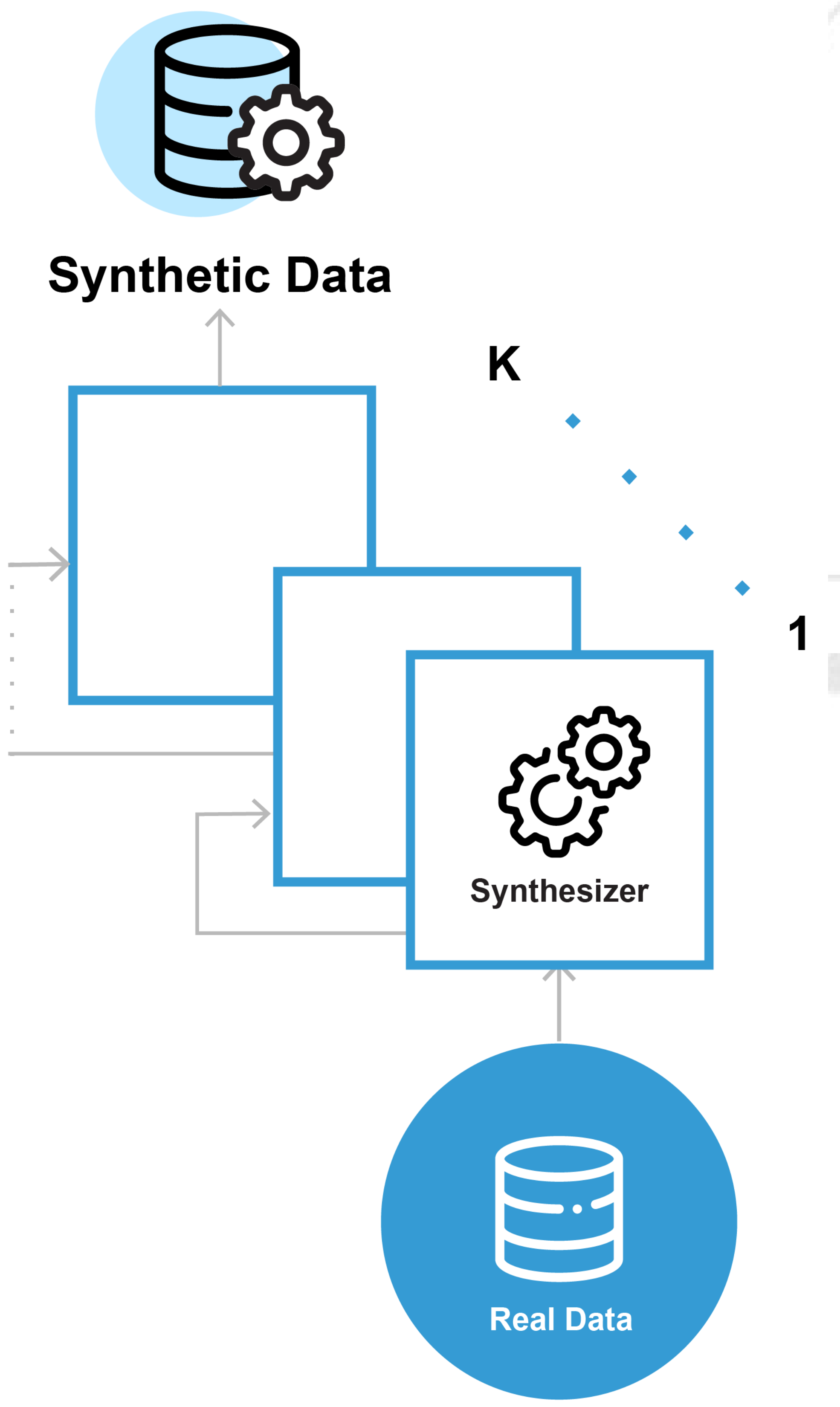


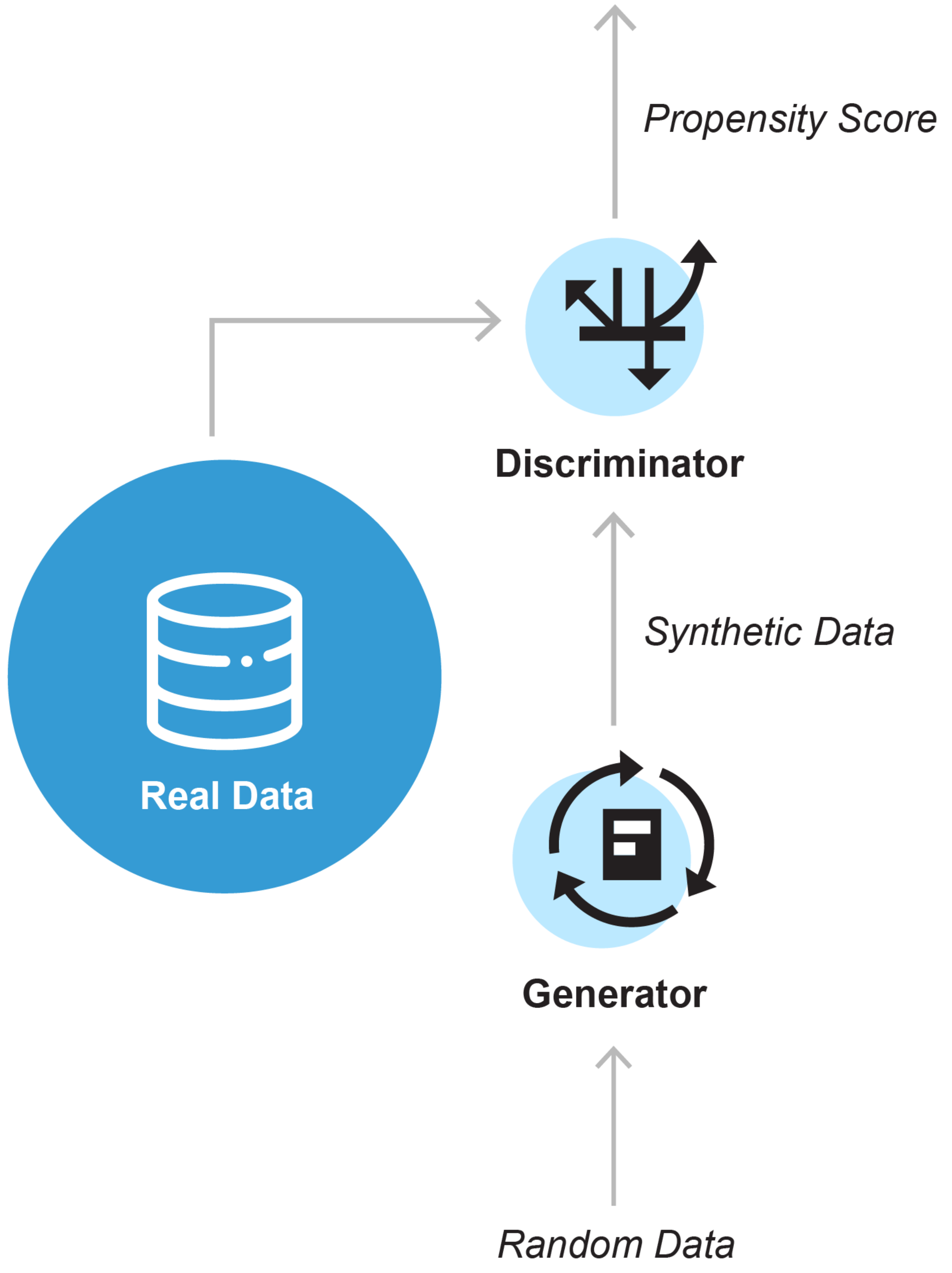
Generative Model

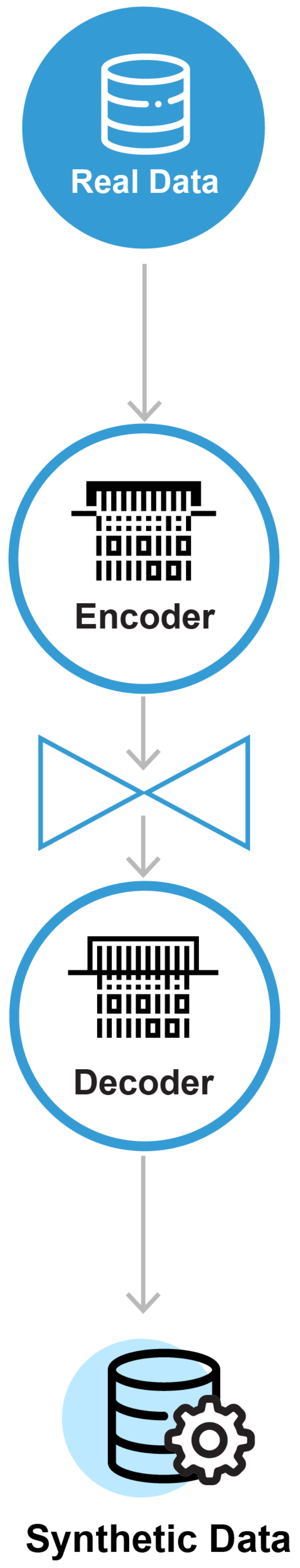
Simulated Values

Intitial Values

Synthesis Techniques



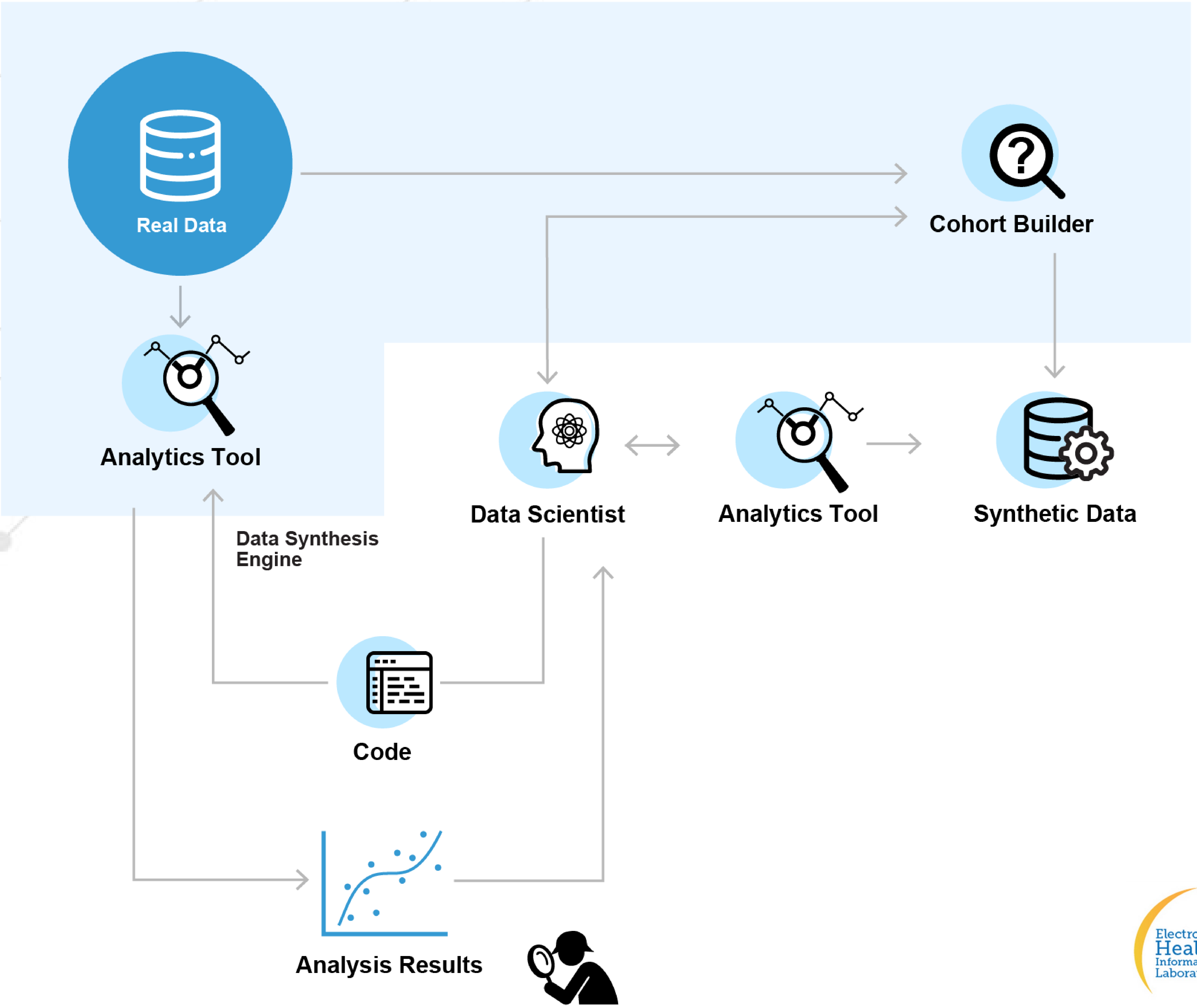




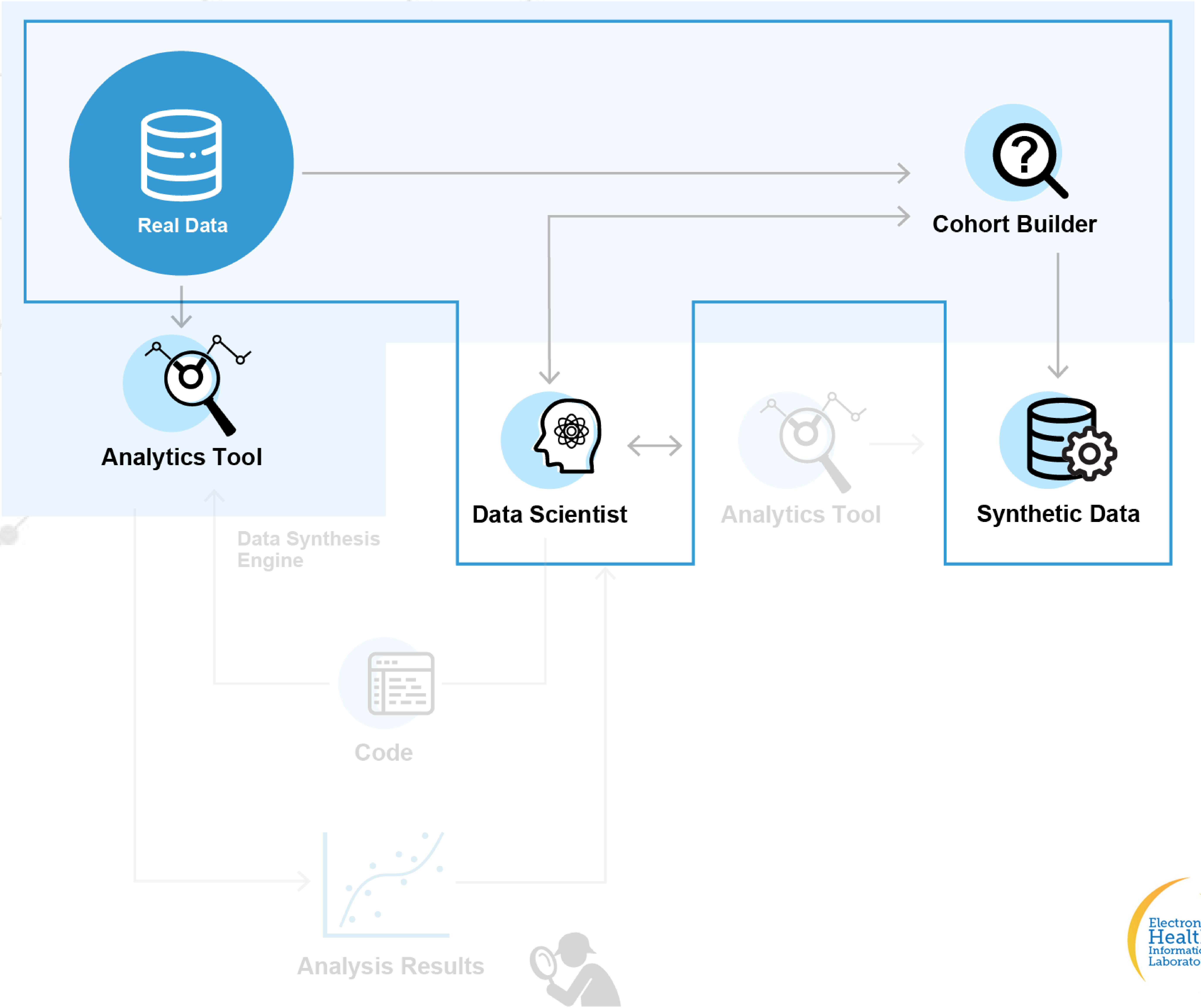
Electronic Health Information Laboratory

Replica Analytics

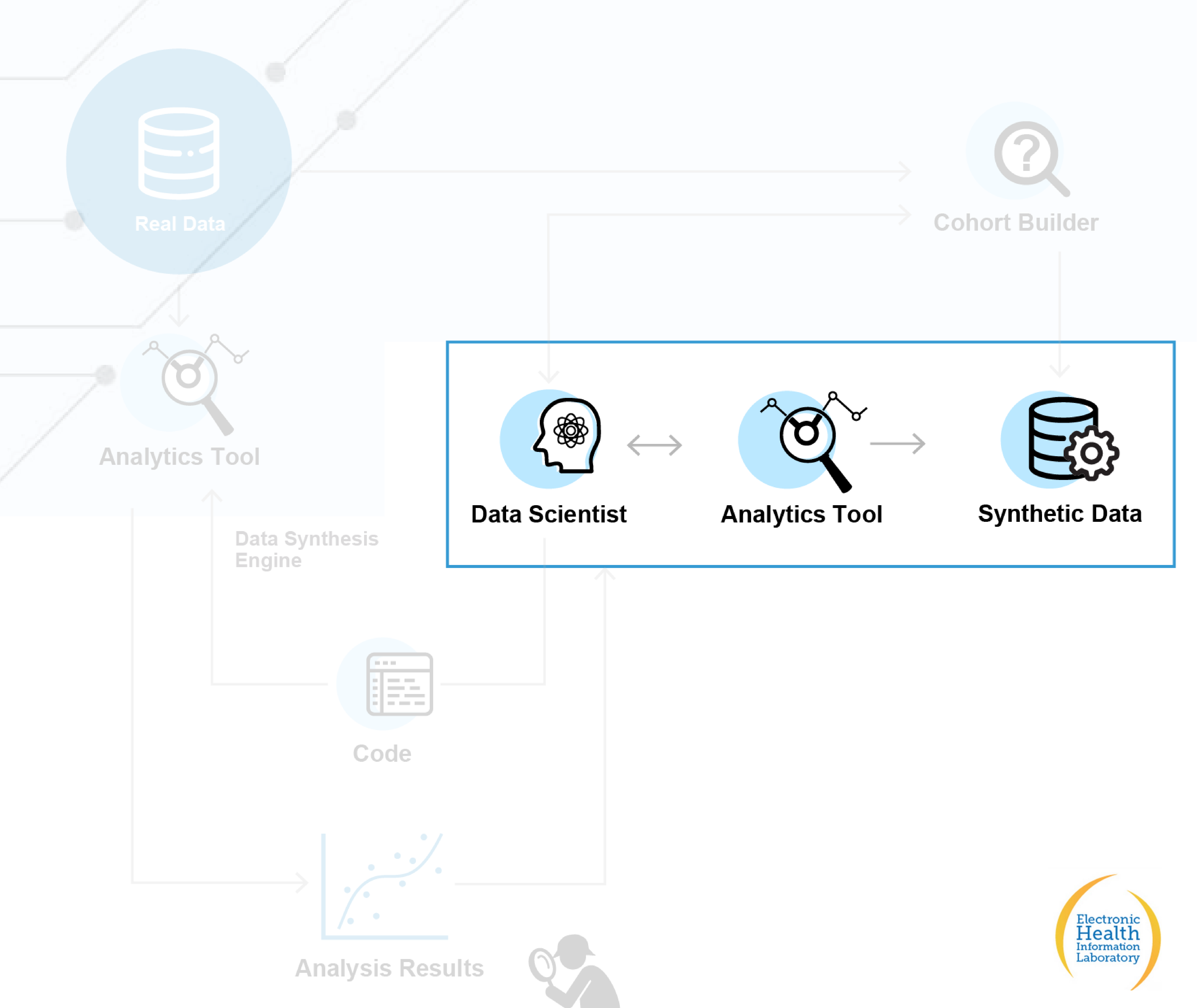
Synthesis as a Platform



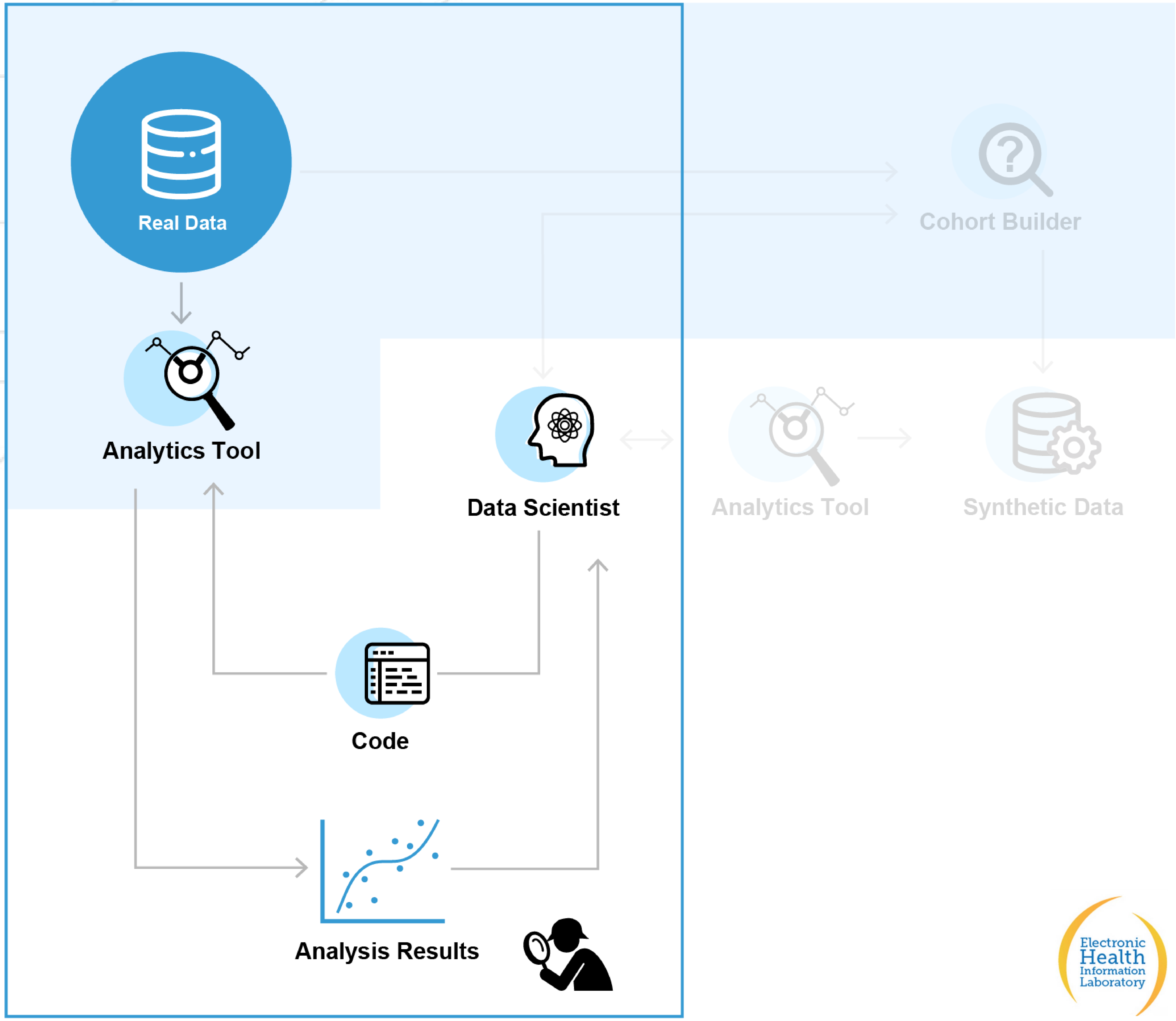
Synthetic Cohort Builder



Analysis on Synthetic Data



Validation Server



Complex Data – Clinical Trials

- Small datasets – is there enough signal to capture the patterns in the data
- Complex relational data models
- Unique patient relationships rather than tabular relationships – the RELREC problem
- Large number of heterogeneous events over an extended period of time

Replicated Analysis

- N0147 trial: Effect of cetuximab on survival among patients with resected stage III colon cancer
- Randomized trial run between 2004 and 2009
- In total, 2,686 adult patients with stage III colon cancer, with two arms:
- Control: adjuvant regimens of folinic acid, fluorouracil, and oxaliplatin / fluorouracil, leucovorin, and irinotecan
- Treatment: cetuximab + control regimens
- Replication was published in 2018 in *Surgery*

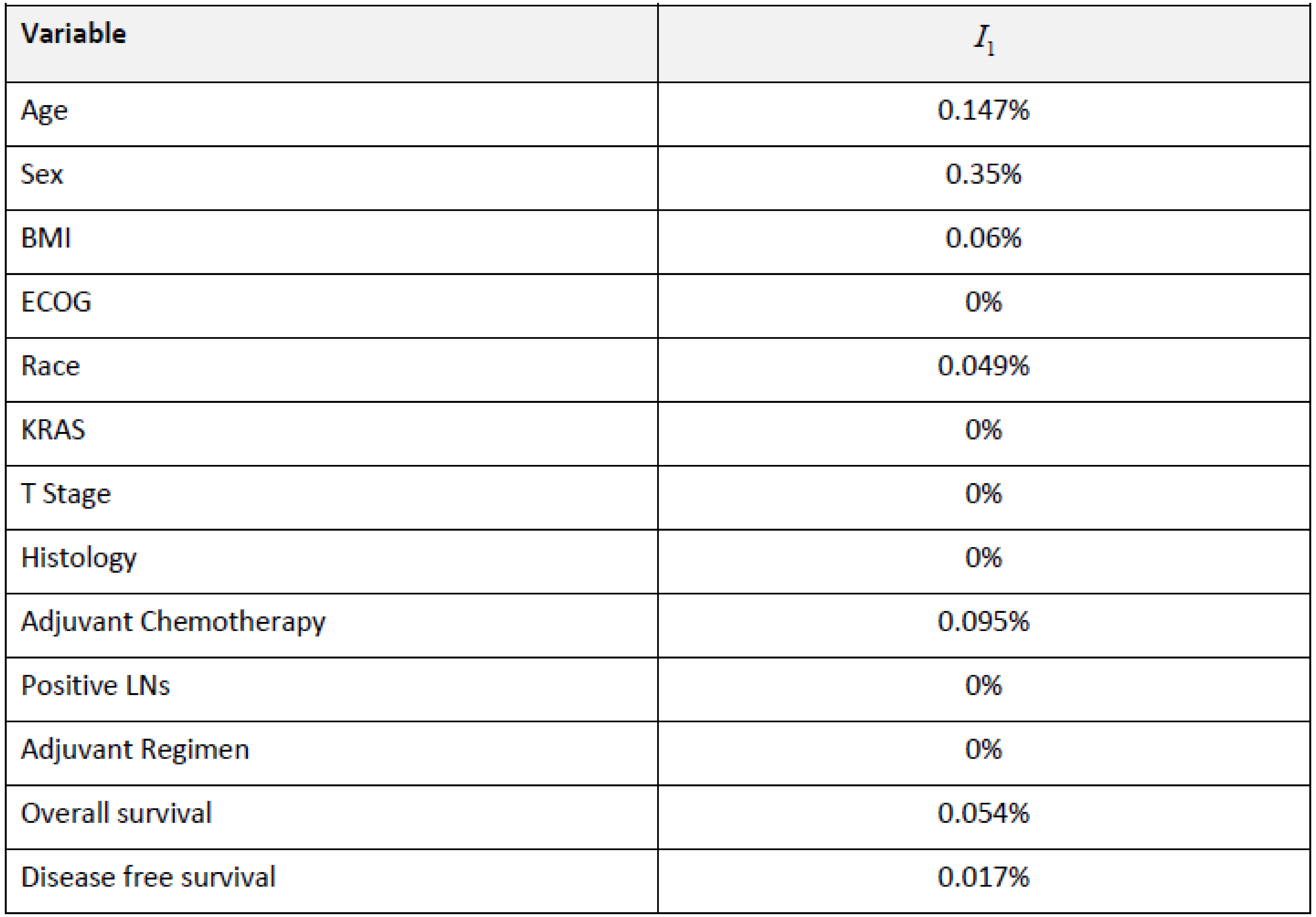
Electronic Health Information Laboratory

Replica Analytics

Replicated Analysis

- Only the control arm, N=1,543
- **Objective:** understand the impact of obstruction on overall survival and disease-free survival
- **Covariates:** cancer staging, lymph node involvement, histology, baseline ECOG performance status, KRAS biomarker, demographics and BMI
- **Statistics:** Descriptive statistics, bivariate relationships, cox model

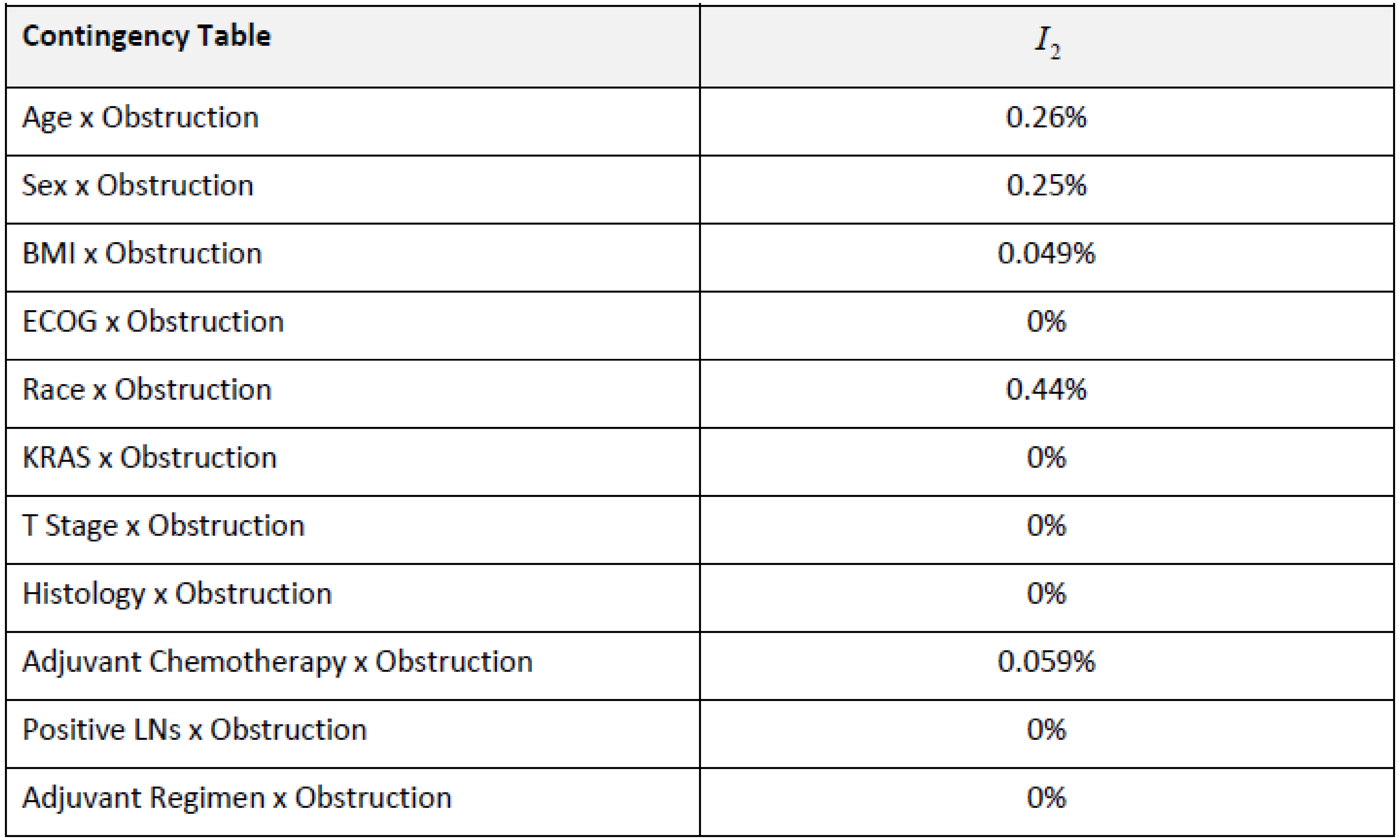
Descriptive Statistics



Variable	I_1
Age	0.147%
Sex	0.35%
BMI	0.06%
ECOG	0%
Race	0.049%
KRAS	0%
T Stage	0%
Histology	0%
Adjuvant Chemotherapy	0.095%
Positive LNs	0%
Adjuvant Regimen	0%
Overall survival	0.054%
Disease free survival	0.017%

% information loss due to using synthetic data as opposed to the real data

Bivariate Statistics



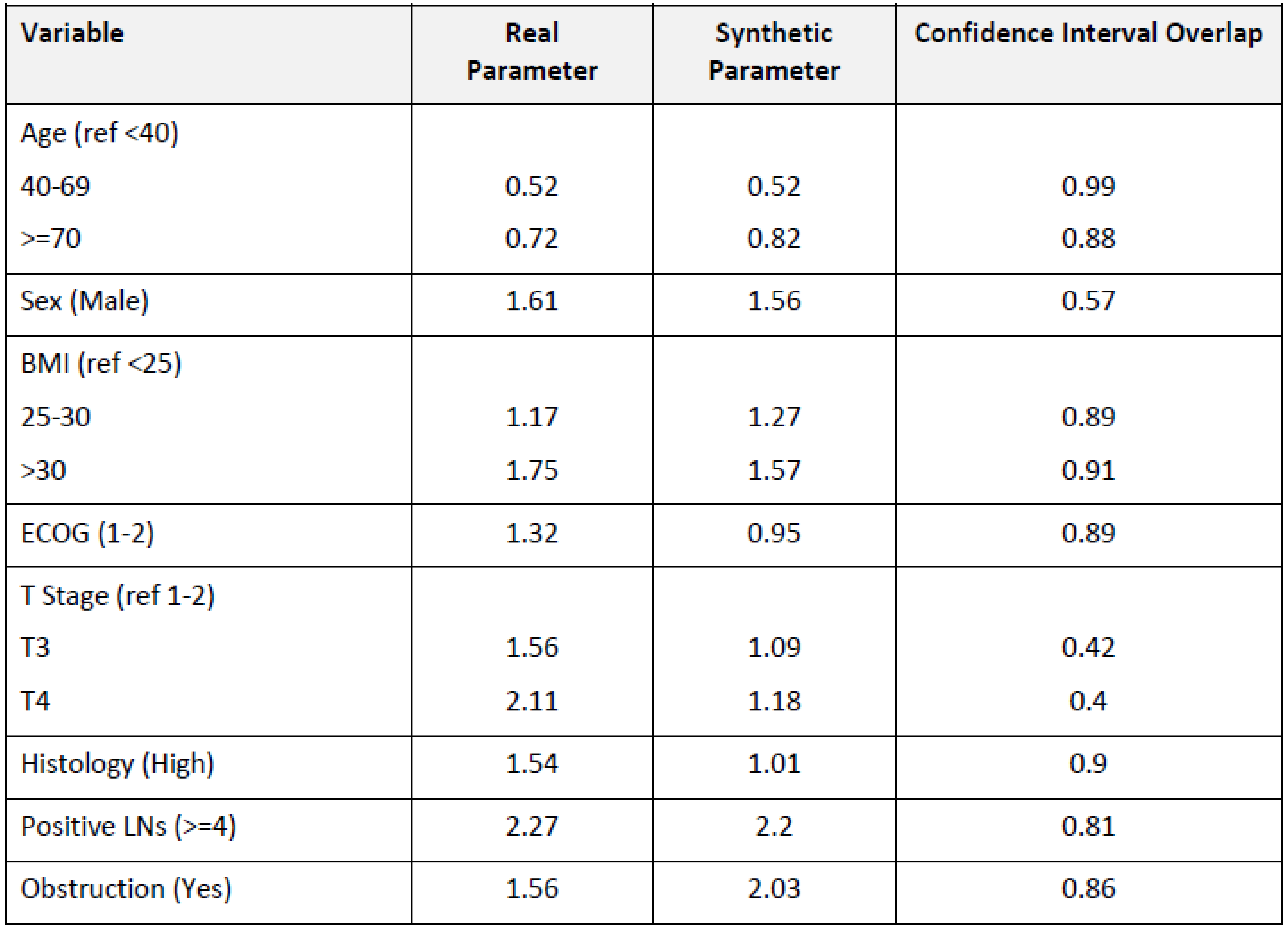
Contingency Table	I_2
Age x Obstruction	0.26%
Sex x Obstruction	0.25%
BMI x Obstruction	0.049%
ECOG x Obstruction	0%
Race x Obstruction	0.44%
KRAS x Obstruction	0%
T Stage x Obstruction	0%
Histology x Obstruction	0%
Adjuvant Chemotherapy x Obstruction	0.059%
Positive LNs x Obstruction	0%
Adjuvant Regimen x Obstruction	0%

% information loss due to using synthetic data as opposed to the real data

Electronic Health Information Laboratory

Replica Analytics

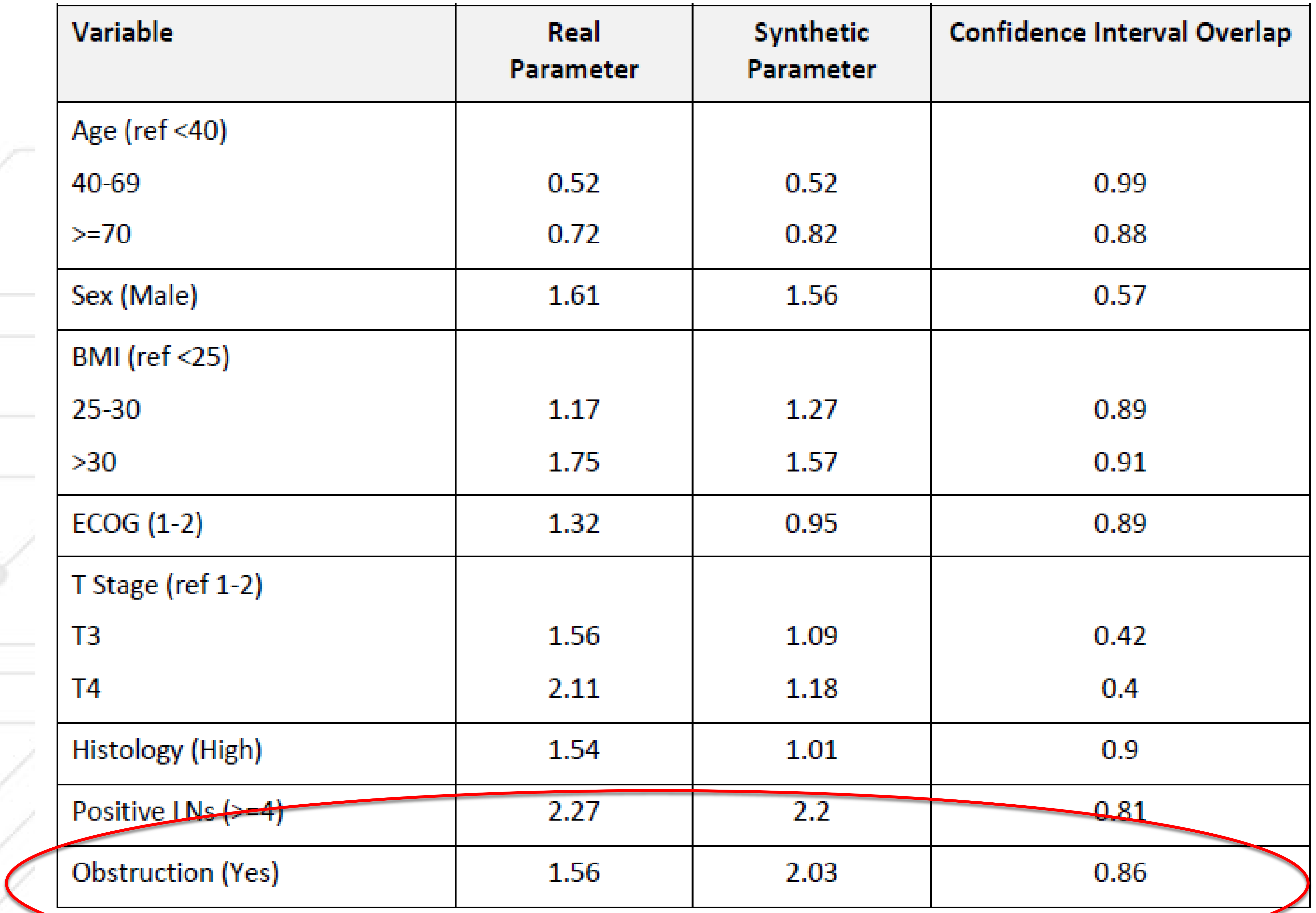
Impact of Obstruction on Overall Survival



Variable	Real Parameter	Synthetic Parameter	Confidence Interval Overlap
Age (ref <40)			
40-69	0.52	0.52	0.99
>=70	0.72	0.82	0.88
Sex (Male)	1.61	1.56	0.57
BMI (ref <25)			
25-30	1.17	1.27	0.89
>30	1.75	1.57	0.91
ECOG (1-2)	1.32	0.95	0.89
T Stage (ref 1-2)			
T3	1.56	1.09	0.42
T4	2.11	1.18	0.4
Histology (High)	1.54	1.01	0.9
Positive LNs (>=4)	2.27	2.2	0.81
Obstruction (Yes)	1.56	2.03	0.86

Proportion of confidence interval overlap

Impact of Obstruction on Overall Survival



Variable	Real Parameter	Synthetic Parameter	Confidence Interval Overlap
Age (ref <40)			
40-69	0.52	0.52	0.99
>=70	0.72	0.82	0.88
Sex (Male)	1.61	1.56	0.57
BMI (ref <25)			
25-30	1.17	1.27	0.89
>30	1.75	1.57	0.91
ECOG (1-2)	1.32	0.95	0.89
T Stage (ref 1-2)			
T3	1.56	1.09	0.42
T4	2.11	1.18	0.4
Histology (High)	1.54	1.01	0.9
Positive LNs (>=4)	2.27	2.2	0.81
Obstruction (Yes)	1.56	2.03	0.86

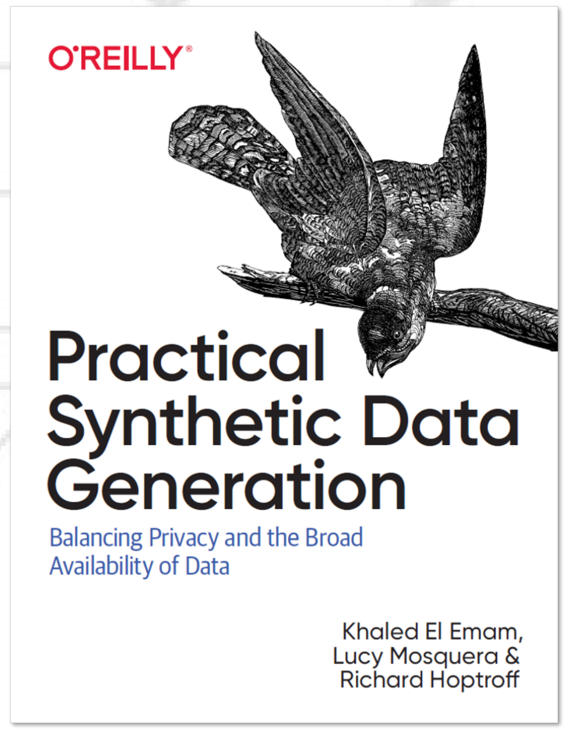
Proportion of confidence interval overlap

Conclusions

- It was possible to replicate the analysis and draw the same conclusions given the objectives of the study
- The combination of cohort builder and validation server enable reliable analytics with synthetic data, as well as faster access to data

If You Want 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: https://bit.ly/2TXI0Jy
- Read our introductory report and book on the topic

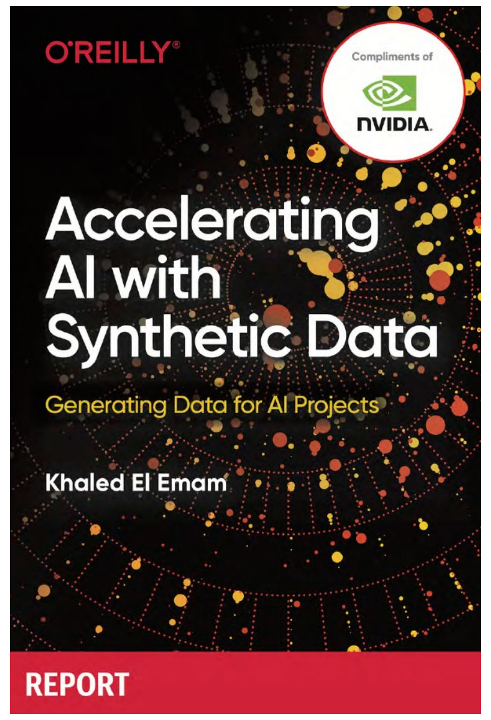


O'REILLY

Practical Synthetic Data Generation

Balancing Privacy and the Broad Availability of Data

Khaled El Emam, Lucy Mosquera & Richard Hoptroff



O'REILLY

Compliments of NVIDIA

Accelerating AI with Synthetic Data

Generating Data for AI Projects

Khaled El Emam

REPORT

Electronic Health Information Laboratory

Replica Analytics