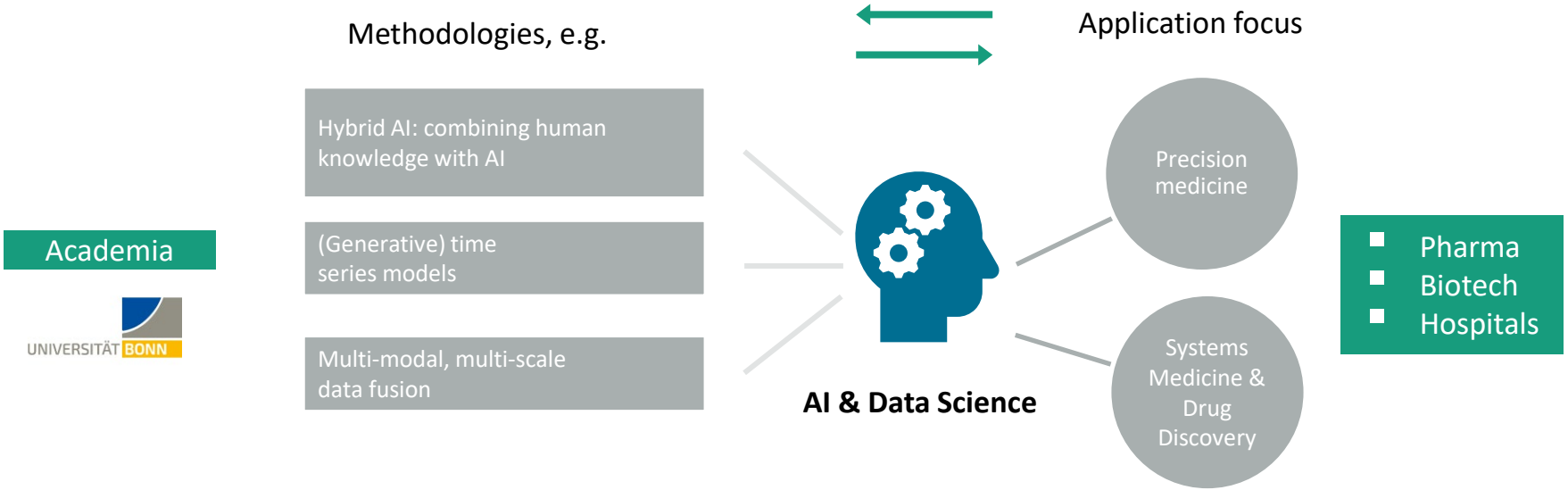
SYNTHETIC DATA FOR SUPPORTING PRIVACY PRESERVING DATA SHARING

Prof. Dr. Holger Fröhlich
Head of AI & Data Science Group, Deputy Head of Department of Bioinformatics
Fraunhofer Institute for Algorithms and Scientific Computing (SCAI)

UNIVERSITÄT BONN

AI & Data Science Group @Fraunhofer SCAI
Mission: Bringing Better Treatments to the Right Patients

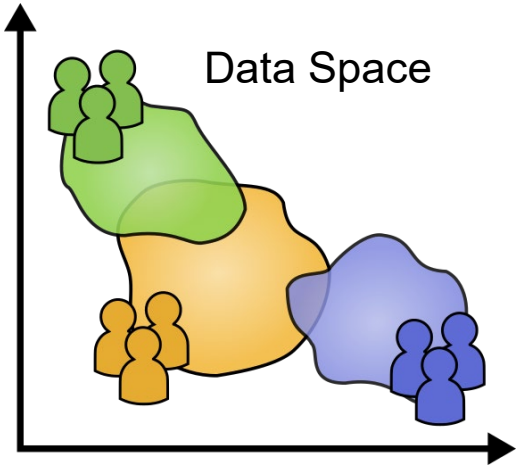


Our value chain:

https://www.scai.fraunhofer.de/de/geschaeftsfelder/bioinformatik/forschungsfelder/AI-DAS.html

Access to Data as a Key Challenge

- Access to patient-level data is highly restricted by EU and German law
- Data within a data holding organization is
  - Biased / not necessarily representative of the entire disease population
  - Often too small to develop generalizable ML models
- Is there a way to facilitate data sharing in a legally compliant manner?



AI for Simulation of Synthetic Patient Data

Real World Clinical Study

AI-technology

VAMBN, MultiNODEs

A mathematical representation of the real world cohort based on an AI model

A synthetic, „virtual“ cohort

Quality assurance:

Good fit between observations in the real world and representation in the virtual cohort

Risk assessment for re-identification

Controlled risk algorithms such as differential privacy

Potential use cases

Simulation

- synthetic data
- "what, if" scenarios
- „Patients-like me“

Exploration

- inclusion / exclusion criteria
- extrapolation
- Statistical power

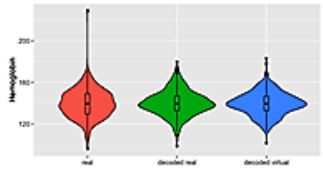
Gootjes-Dreesbach, … & Fröhlich, Frontiers in Big Data: Medicine and Public Health, 2020
Sood, … & Fröhlich, Scientific Reports, 2020

Synthetic Data are Realistic While Respecting (ε, δ)-Differential Privacy

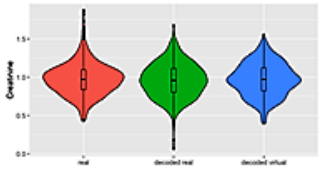
Marginal distributions

Pearson correlations

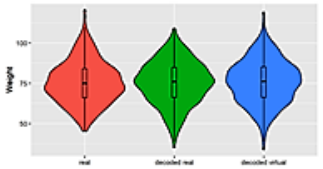
a) SP513



Type	Mean	SD	25%	Median	75%
real	139.4	13.76	129.69	139.35	149.02
decoded real	140.06	12.24	132.73	139.28	146.9
decoded virtual	139.48	11.25	132.74	139.36	146.84

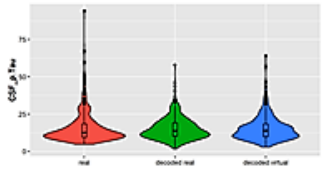


Type	Mean	SD	25%	Median	75%
real	0.98	0.21	0.84	0.97	1.1
decoded real	0.96	0.22	0.8	0.96	1.11
decoded virtual	0.97	0.22	0.82	0.97	1.12

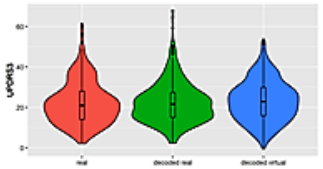


Type	Mean	SD	25%	Median	75%
real	75.65	13.41	66	75	84
decoded real	75.37	13.48	66.18	76.13	84.81
decoded virtual	76	13.64	66.19	76.25	85.17

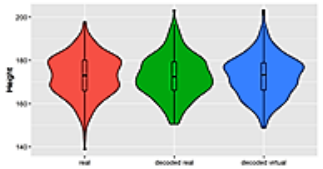
b) PPMI



Type	Mean	SD	25%	Median	75%
real	15.83	10.34	9.5	12.45	18.42
decoded real	15.4	7.59	10.15	13.75	19.09
decoded virtual	15.4	8.16	9.77	13.81	18.42

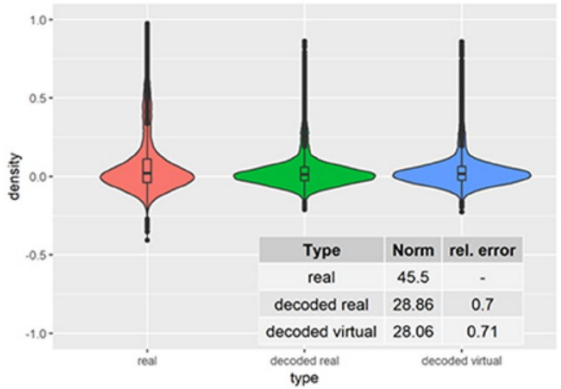


Type	Mean	SD	25%	Median	75%
real	22.35	10.65	14	21	28
decoded real	22.55	10.33	15.04	21.61	27.41
decoded virtual	22.34	10.18	15.87	22.89	29.9



Type	Mean	SD	25%	Median	75%
real	172.9	9.66	166	173	180
decoded real	172.55	9.67	166.21	172.44	179.31
decoded virtual	172.65	9.46	166.25	173.24	178.86

PPMI

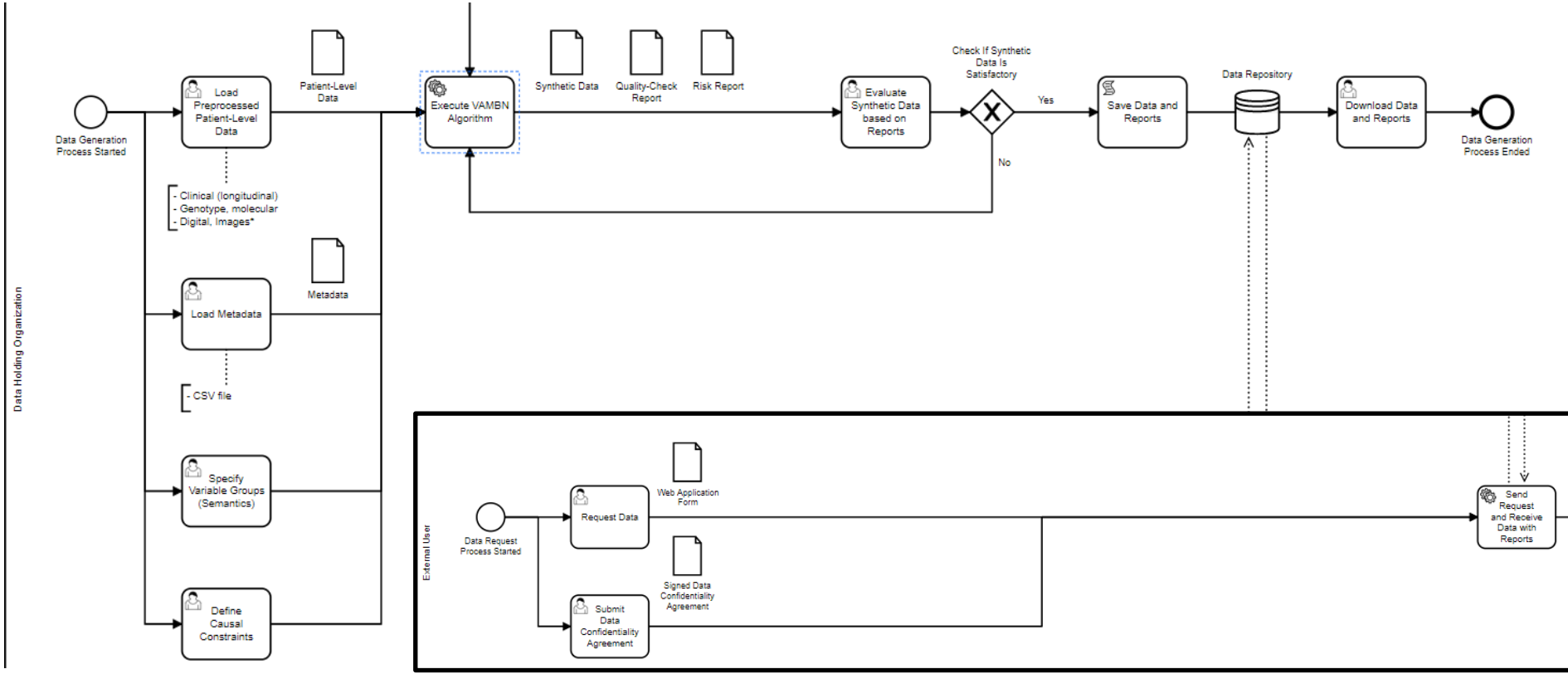


Type	Norm	rel. error
real	45.5	-
decoded real	28.86	0.7
decoded virtual	28.06	0.71

Guarantee on the probability to compromise an individual‘s privacy

Gootjes-Dreesbach, … & Fröhlich, Frontiers in Big Data: Medicine and Public Health, 2020

Envisioned Use of Synthetic Data Generator in NFDI4Health



Conclusion

Access to pseudonomyized patient-level data is one of **THE** obstacles for Biomedical Data Science, specifically in Germany

- Comparison to US: market places for data, e.g. IBM

Data silos hinder progress in medical research

- Data biases
- Data within each organization often too small

Sufficiently realistic synthetic data might be a way to facilitate data sharing and thus help data driven biomedical research