



RESEARCH DATA ALLIANCE

Health Data Interest Group @P13

Data sharing challenges in biomedical Artificial Intelligence (AI)

2 April 2019, 11:30-13:00

Synthetic Data in Medicine

Davide Zaccagnini, MD

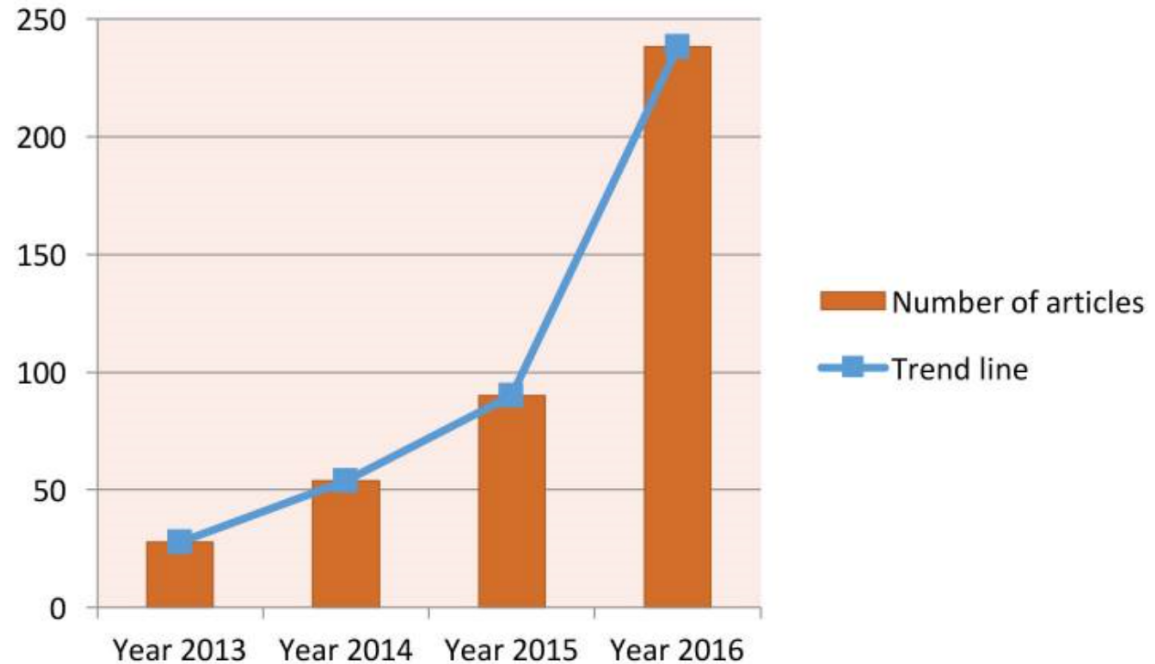
Lynkeus.eu

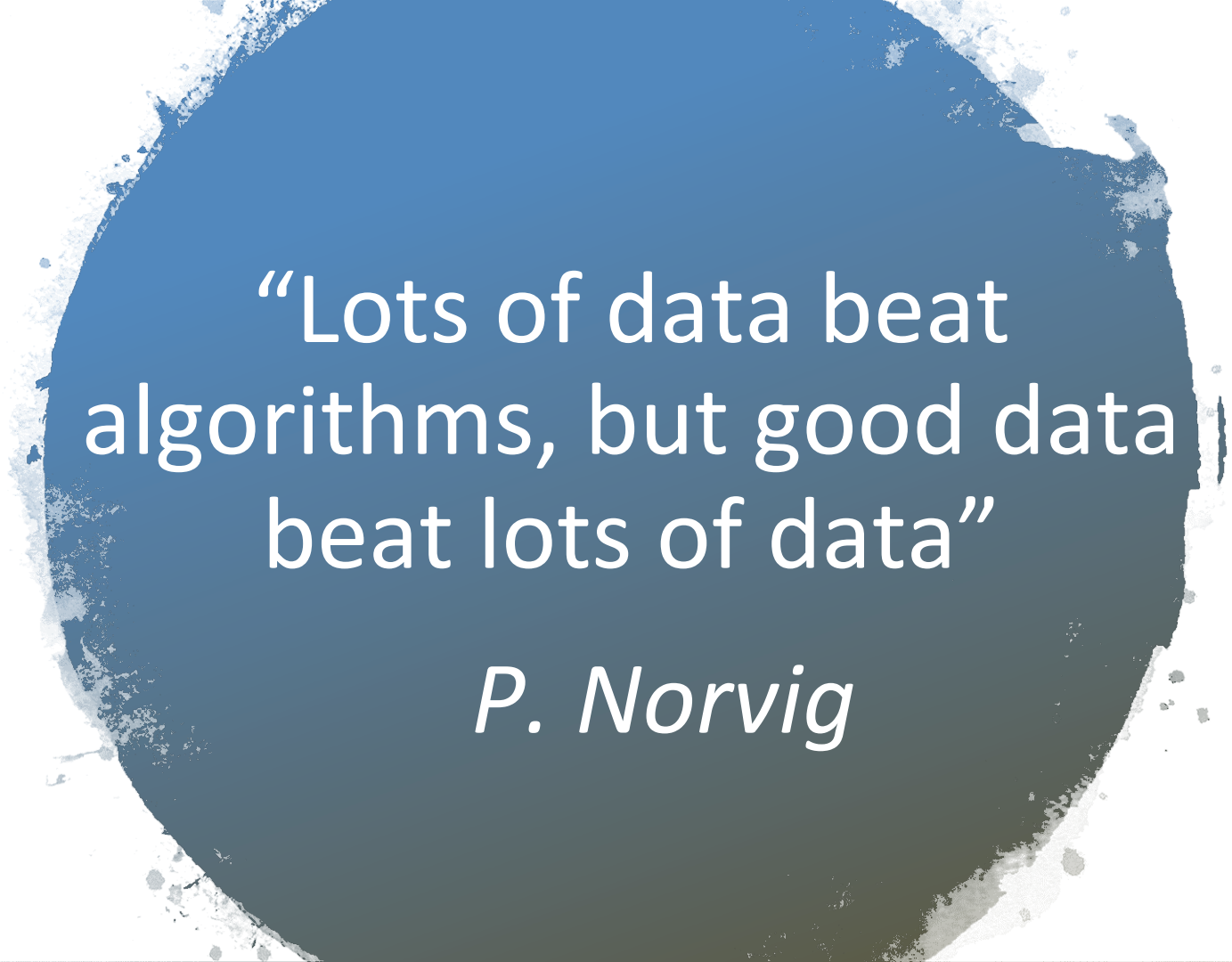
research data sharing without barriers

rd-alliance.org

AI in medicine

Publications on
medical applications
of deep learning

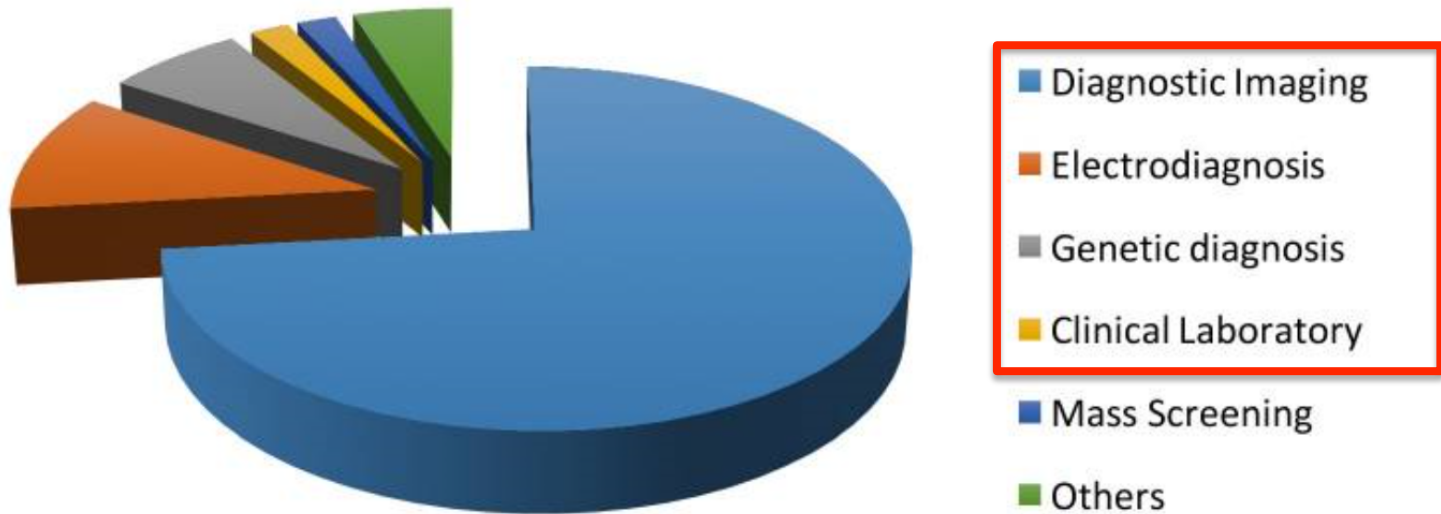




“Lots of data beat
algorithms, but good data
beat lots of data”

P. Norvig

data sources for AI development



The healthcare data landscape

- ✓ Vast amounts of typically low-quality clinical and wellbeing data
- ✓ Much fewer, good quality and increasingly expensive research data
- ✗ Both are locked in siloes by (much needed) privacy laws

Still short on data

Up to 70% of records can lack a meaningful diagnostic code

Quality of Data

\$7.0 to \$52.9 million
(up to \$350 million)
for Phase 2-3 trials

Cost of Data

Access to Data

HIPAA, GDPR
(penalties of up to 4%
in yearly revenue)

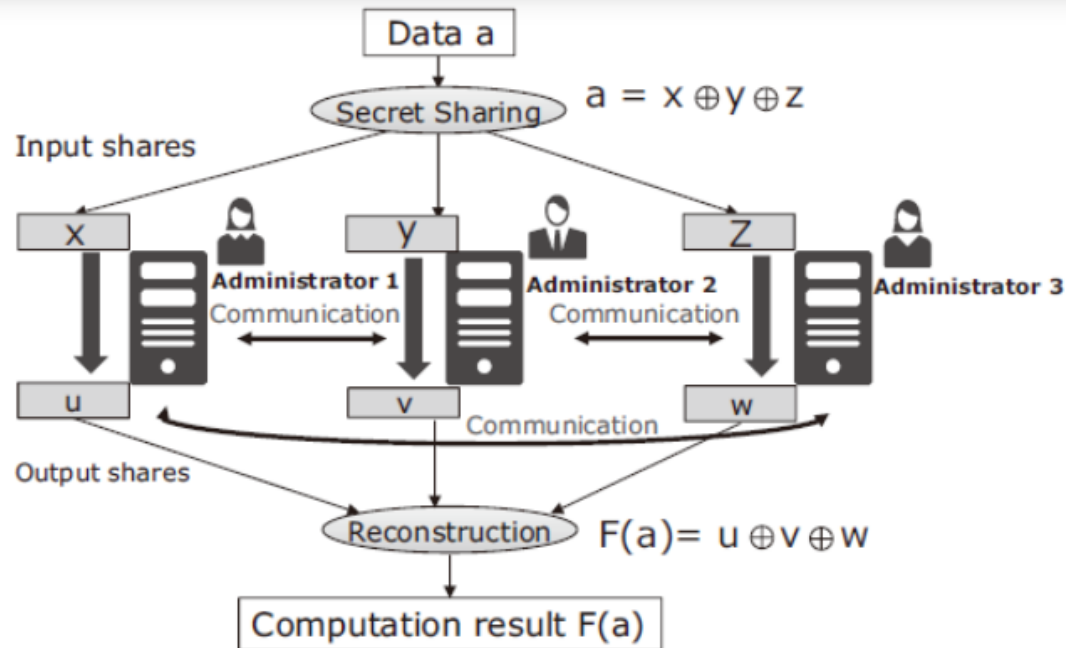


Emerging solutions

Secure multi-party computation

Synthetic data

SMPC





SMPC

- Computational costs are high, or hard to predict
- Needs strict, shared governance among data providers
- Data access is hard to quantify/monetize



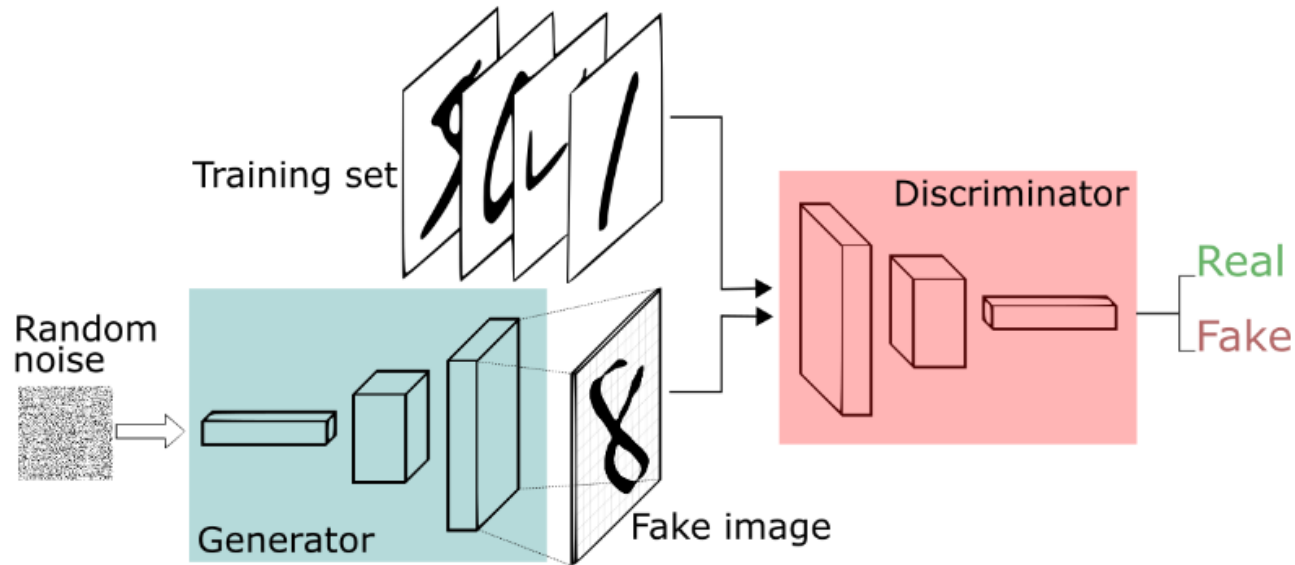
What is synthetic data?

Data generated to recreate pre-defined characteristics of a target population for one or more clinical modalities.

Invented outside of medicine



Understanding synthetic data



Generative Adversarial Network framework.



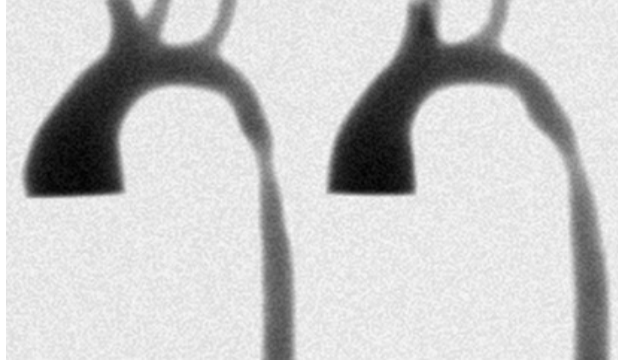
Understanding Synthetic Data

- **Generative methods**
 - GAN, InfoGAN, Monte Carlo simulations
- **Coupled with quality control systems**
 - Discriminators, including human experts
 - Risks: mode collapse, leakage
- **And INTERPRETABILITY!**
 - Ex. Mutual information algorithm

Practical uses

Images, free text, EMR or genetic data

- Multi-modal generation



From MRI to Angiography
From 3D to multiple 2Ds views

From structured data to free text
(ex. for NLP applications or
patient communications)

Synthesis of images of the spine

 **frontiers**
in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 03 May 2018
doi: 10.3389/fbioe.2018.00053

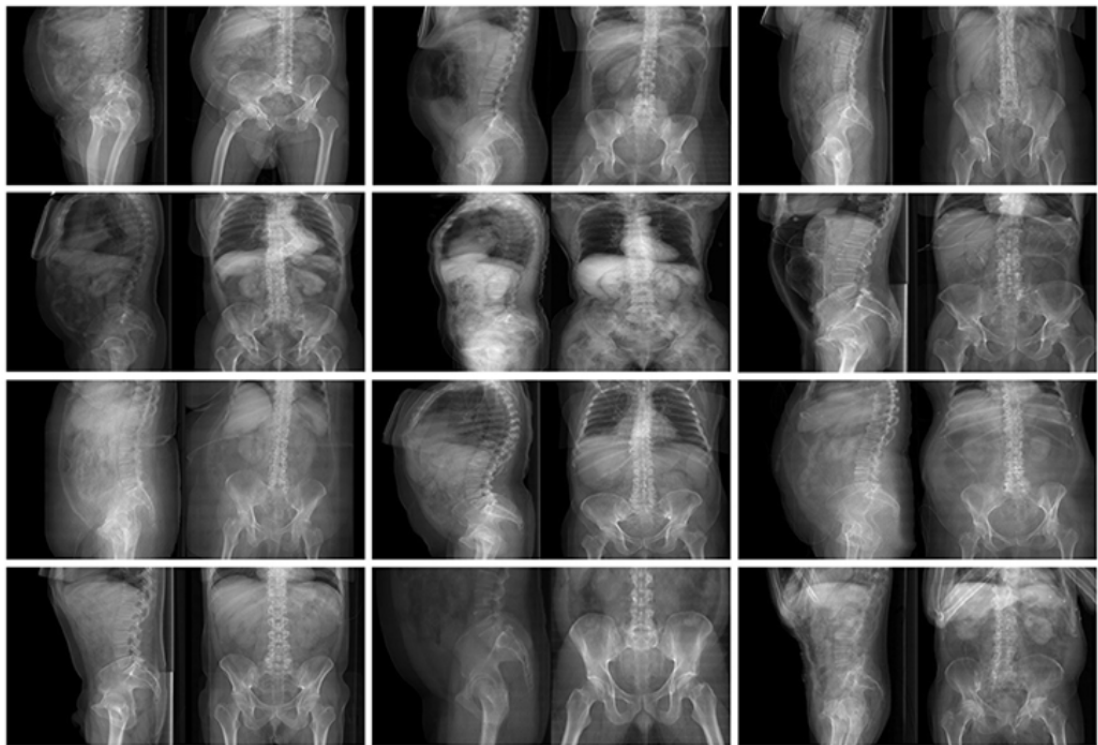


Exploring the Potential of Generative Adversarial Networks for Synthesizing Radiological Images of the Spine to be Used in *In Silico* Trials

Fabio Galbusera^{1}, Frank Niemeyer², Maike Seyfried³, Tito Bassani¹, Gloria Casaroli¹, Annette Kienle³ and Hans-Joachim Wilke²*

¹ IRCCS Istituto Ortopedico Galeazzi, Milan, Italy, ² Center for Trauma Research Ulm, Institute of Orthopedic Research and Biomechanics, Ulm University, Ulm, Germany, ³ SpineServ GmbH & Co. KG, Ulm, Germany

Generating sagittal views of the spine XR images





Value and Limitations

- **If you didn't build it in, it isn't there**
 - Information replication, not information generation
- **But if it was not there, you may be able to build it**
 - Correcting data gaps and biases
 - Skewed distributions, underrepresented populations
 - Errors (ex. Blood pressure measurements of “0” instead of “missing”)
- **GDPR compliant**
 - “Reasonable efforts to protect data”
 - Risk vs. cost can be assessed



Value and Limitations 2

- **Low cost**
 - Infinitely scalable once the generative pipeline is set
 - Images can be pre-annotated
 - Modifying parameters allows a range of generations
- **Require substantial statistical and data management skills**
 - But it's getting better
- **For some biomedical products they are becoming main-stream**

Phase	Goal	Population	Success rate
Phase I	Testing for safety	20–100 healthy volunteers	approximately 70%
Phase II	Testing for efficacy and side effects	100–300 patients with diseases	approximately 33%
Phase III	Testing for efficacy, effectiveness and safety	300–3,000 patients with] diseases	25–30%

The problem with clinical trials

The promise of in silico trials



Reduction of testing in animals and humans

-50%



Reduction of clinical trial cost by 50%

-30%



Reduction of the duration of clinical development by up to 30%



Reduction of failure rate of concept drugs in the pipeline

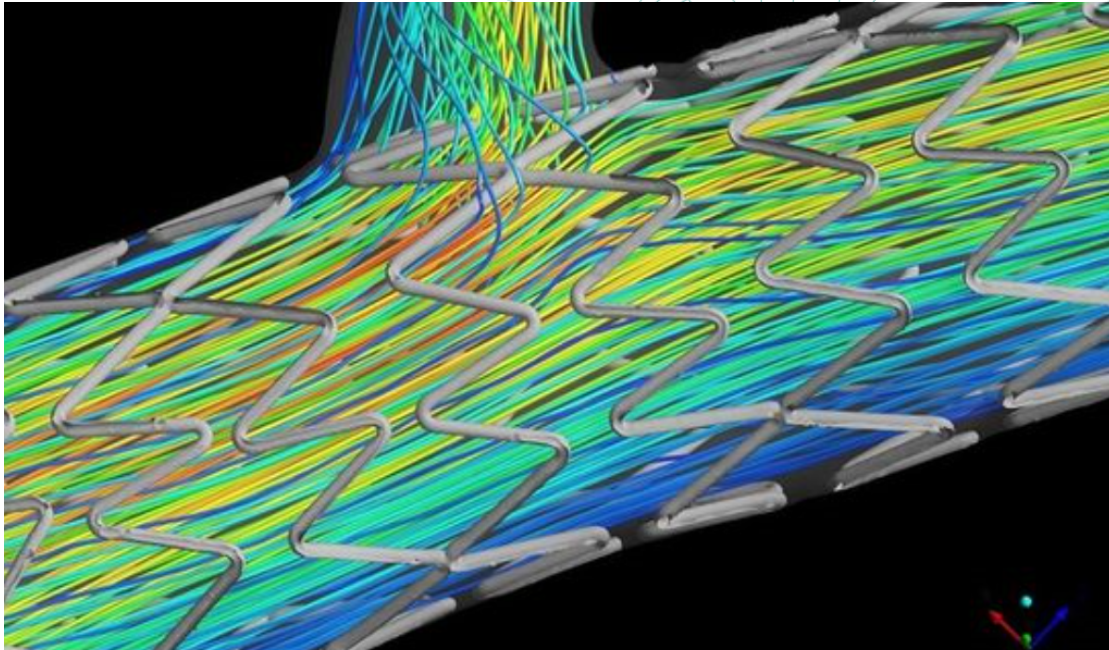


There are examples of successful ISCT, reaching up to a 93% alignment, but it is still minimally used in the pharma industry



15 *in silico* trials reported on clinicaltrial.gov by 2016
6 completed

‘Dynamic’ synthetic data:
Models of anatomy, physiology
and pathology to perform
simulations



Adaptive, in-silico trials

- Using good prior information in a Bayesian approach for the statistical analysis of a trial
- Good simulations produce good priors

The way forward

- SD can
 - Effectively protect patient privacy
 - Reduce costs of biomedical data access, at scale
 - Support real-world applications in research and AI development

What is missing?



Data marketplaces



Thank-you