# SESSION 3: APPLICATIONS OF SYNTHETIC DATA IN THE LIFE SCIENCES INDUSTRY I

## EXPLORING THE POTENTIAL OF SYNTHETIC DATA IN CLINICAL RESEARCH: APPLICATIONS, BENEFITS, AND CHALLENGES

uOttawa

Replica Analytics
AN AETION COMPANY

Presented by:

**Jan Seidel,**
Principal Methodology Statistician,
Boehringer Ingelheim Pharma

# Exploring the Potential of Synthetic Data in Clinical Research: Applications, Benefits, and Challenges

**Jan Seidel & Dooti Roy**

Boehringer Ingelheim Pharma GmbH & Co KG

Synthetic Data Summit

November 30, 2023

Boehringer Ingelheim

# Disclaimer

# What you will hear about today

- What is Synthetic Data?
- Synthetic data generation in clinical development
- Let’s look into 2 examples: GANerAid & patientGAN
- Synthetic patients for efficacy analyses: DINAMO™ trial

Boehringer Ingelheim

# What is Synthetic Data?

- Donald Rubin coined the term back in 1993 in a paper while describing simulated datasets
- Synthetic data is generated by **computer algorithms** or other **simulation methods** that resembles the original individual level data, and retains the **same characteristics of the original data**, including missing values and patterns
- In drug development, terms such as **Synthetic patients** or **Artificial patients** or **Digital Twins** are becoming very popular
- Quality controlled synthetic data can be cost-effective yet free from usual real-world challenges like privacy, logistics, re-use and lengthy recruitment time

Rubin (1993)

# What is Synthetic Data?

**None of these persons exist!**

Created by artificial intelligence

Does not only generate object **similar to** human faces (something like ☺) but considers

angle, color, ethnicity, distance, eye gaze, emotional expression, imperfections (e.g. of skin), paraphernalia

Trained by Generative Adversarial Network (GAN) with numerous images of human faces

**If we have sufficient historical input, computational power & time → we achieve amazing outcomes**

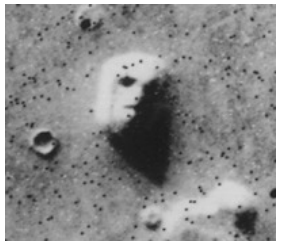
https://www.thispersondoesnotexist.com/ [Nov 30, 2023]; Goodfellow et al (2014); Karras et al (2019)

# Modern generative algorithms are great, right?!?

**If we have sufficient historical input, computational power & time**

**→ we achieve amazing outcomes**

- Evolution made us great in **holistically** identifying human faces (“Pareidolia“)



https://en.wikipedia.org/wiki/Pareidolia [Nov 30, 2023]; Thompson (1980);
https://www.thispersondoesnotexist.com/ [Nov 30, 2023]

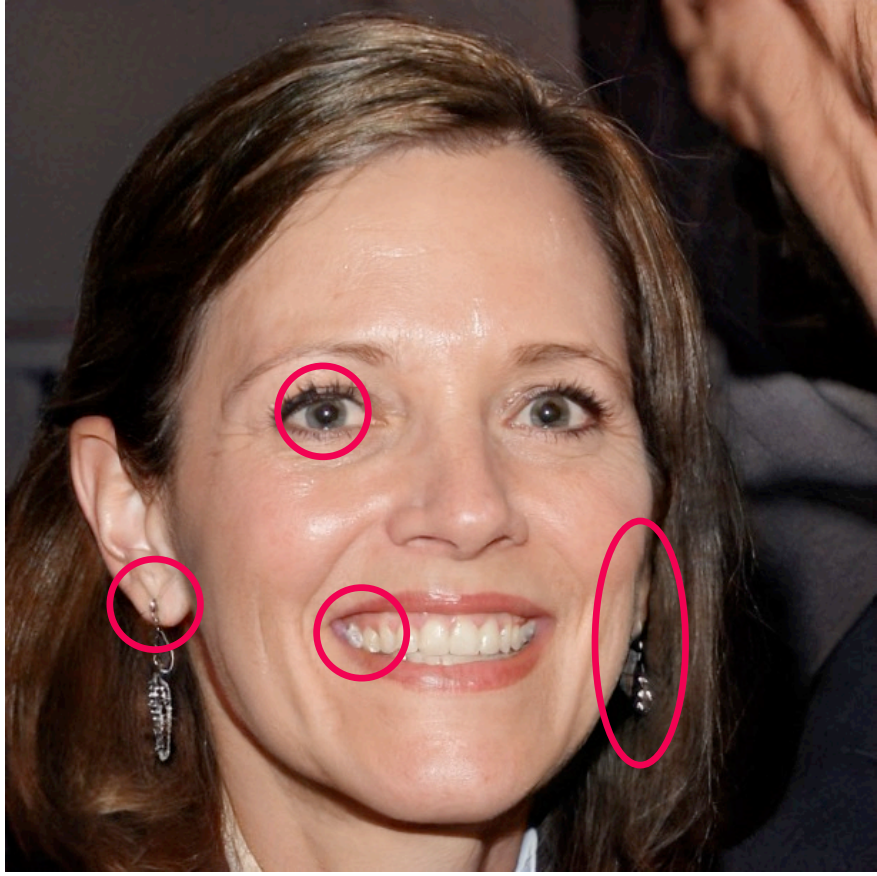
Boehringer Ingelheim

# Modern generative algorithms are great, right?!?

**If we have sufficient historical input, computational power & time**

**achieve amazing outcomes**

- Evolution made us great in **holistically** identifying human faces (“Pareidolia“)
- **What happens if we take closer look??**



→ **Generating data is one thing but evaluating quality equally important**

- In clinical development we must trust **validity of our data!**
- Difficult to tell if generated output is good enough: hard for images – really hard for **tabular data**

https://www.thispersondoesnotexist.com/ [Nov 30, 2023]

Boehringer Ingelheim

# Why should we bother?

Algorithm cannot only create images of human faces (and also cats) but also **novel molecular structures**

But synthetic data can help clinical development & analyses in different ways:

- **Data Enrichment:** improve data quantity (e.g. images, tabular instances) or quality (missing value imputation)
- **Data Enhancement:** synthetically increase amount of rare / extreme cases
- **Research optimization:** alternatives to conventional simulation approaches → create numerous realistic synthetic data
- **Trial design planning & recruitment:** keep high power while reducing real patient numbers (e.g. synthetic controls)
- **Digital Twins:** use synthetic copy of real patients as comparison or alert systems
- **Scientific communication**: publish synthetic 'raw' data to make your point, no data protection concerns!

# Synthetic Data & Artificial Intelligence
# Potentials and Pitfalls

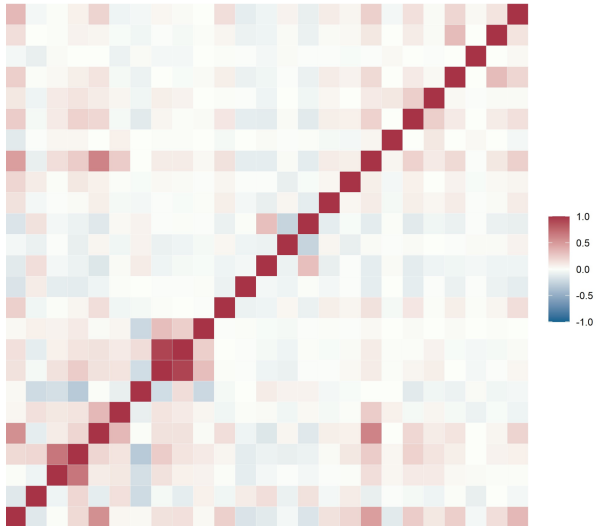
Lucas Krenmayr, Roland Frank, Christina Drobig, Michael Braungart, Jan Seidel, Daniel Schaudt, Reinhold von Schwerin, and Kathrin Stucke-Straub. „GANerAid: Realistic synthetic patient data for clinical trials“. *Informatics in Medicine Unlocked* 35 (2022): 101118.

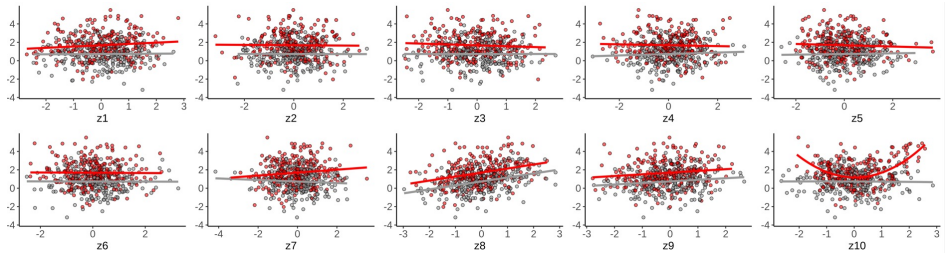
Boehringer Ingelheim

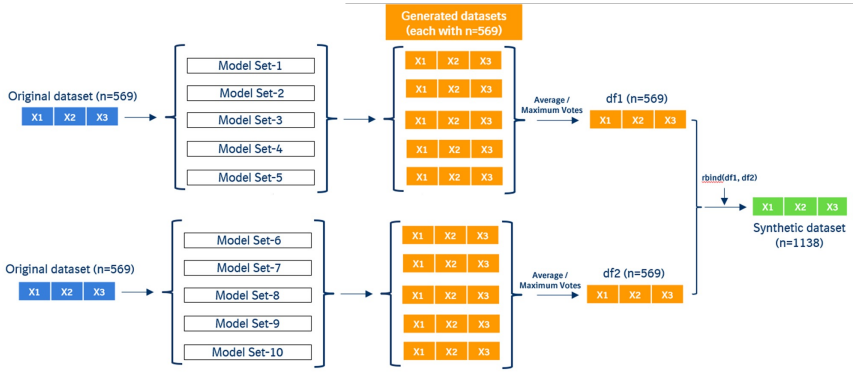
# Synthetic data generation in clinical development

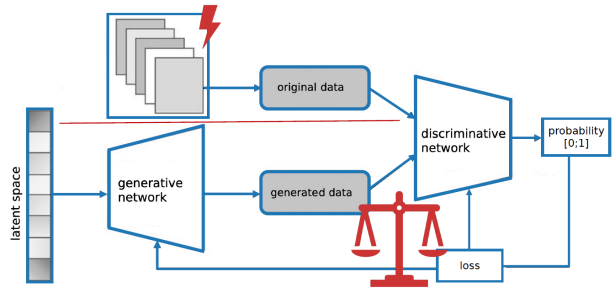
Initial questions determine **approach of synthetic data generation**

- What is the **goal** of the data generation?
- What is the **data dependency**? How many variables? What type? Intercorrelations? Nonlinearity?
- How many **real observations** (e.g. patients) do I have to be used for **training**?
- How can I evaluate whether synthetic **output is acceptable**?



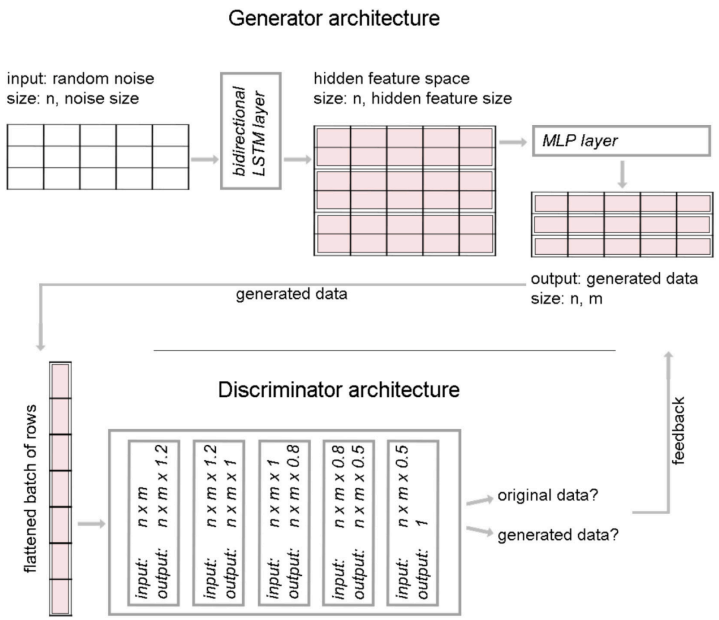

**Let's look into 2 examples**

Boehringer Ingelheim

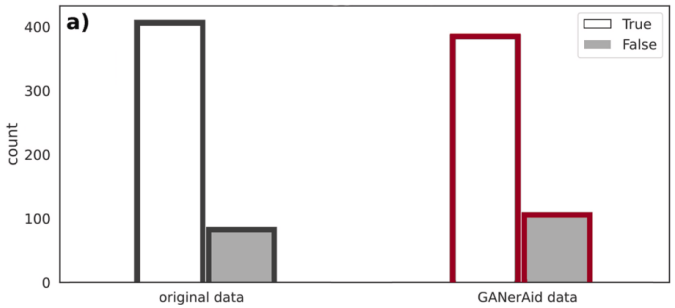
# Synthetic data generation in clinical development – GANerAid

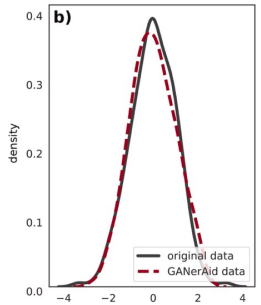
Often relying on **tabular** data with specific data relationship

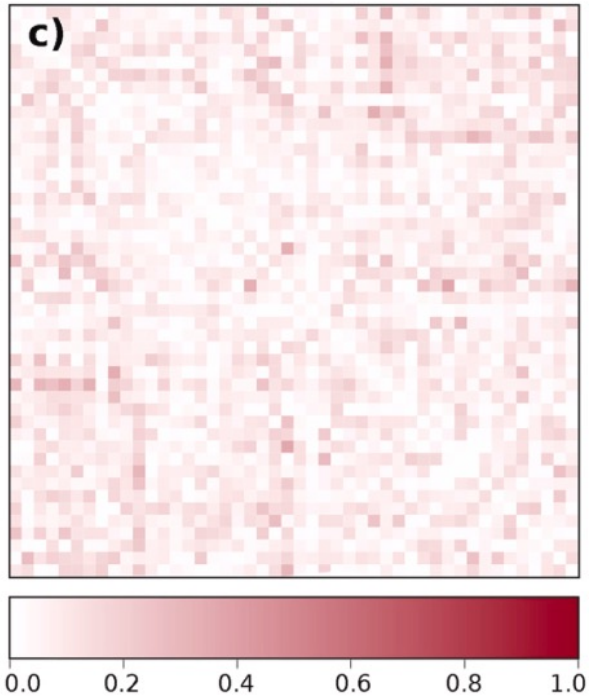
→ introduced Generative Adversarial Network based on **LSTM layers** to preserve underlying data properties (*GANerAid*)




Krenmayr et al (2022)



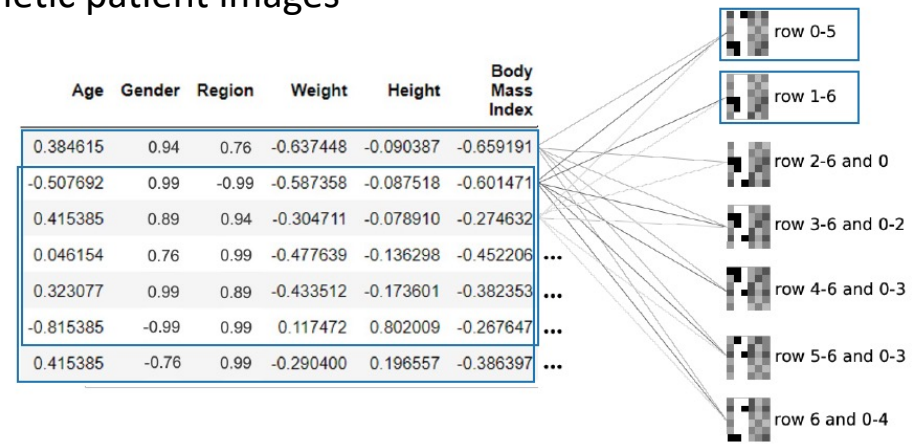




**Additional metrics!**

# Synthetic data generation in clinical development – patientGAN

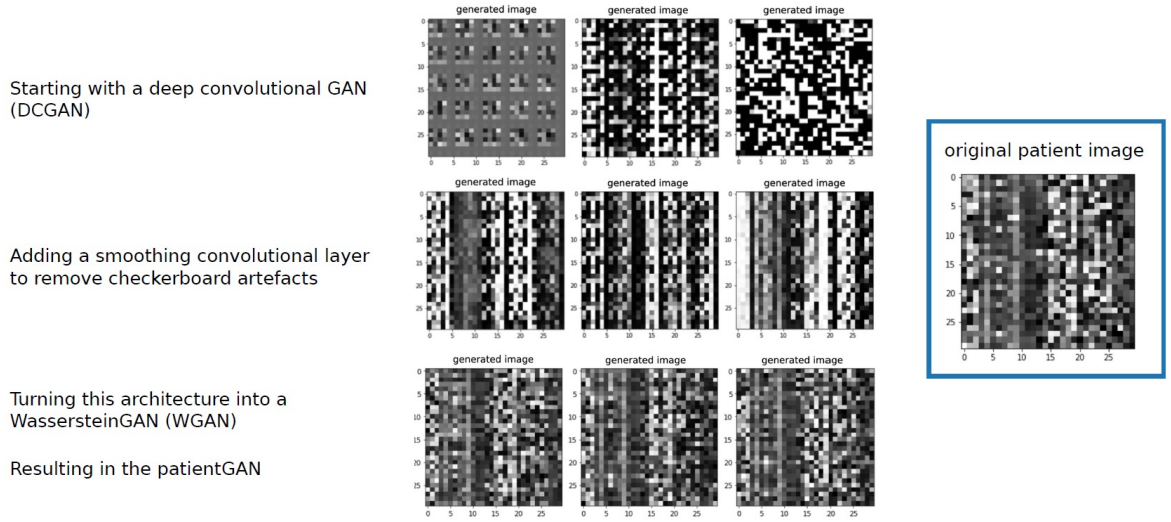
GANerAid promising approach, but needs **~500 observations minimum** → introduced the ***patientGAN***

- Turn tabular data into greyscale patient images & train *patientGAN* to generate synthetic patient images
- Combine the generated images to one tabular data set
- Evaluate the generated data set according to its statistical properties




Advantages:

- Makes use of Convolutional Neural Networks
- Epoch-wise evaluation
- Observable training
- Increase efficiency due to early stopping

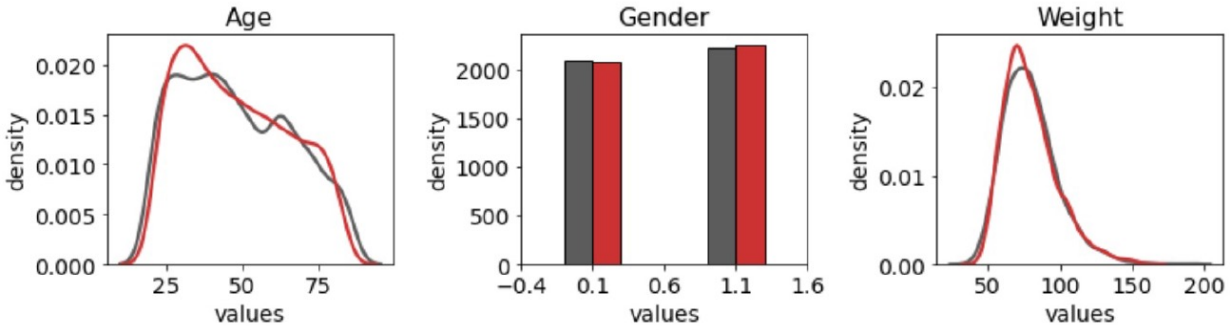


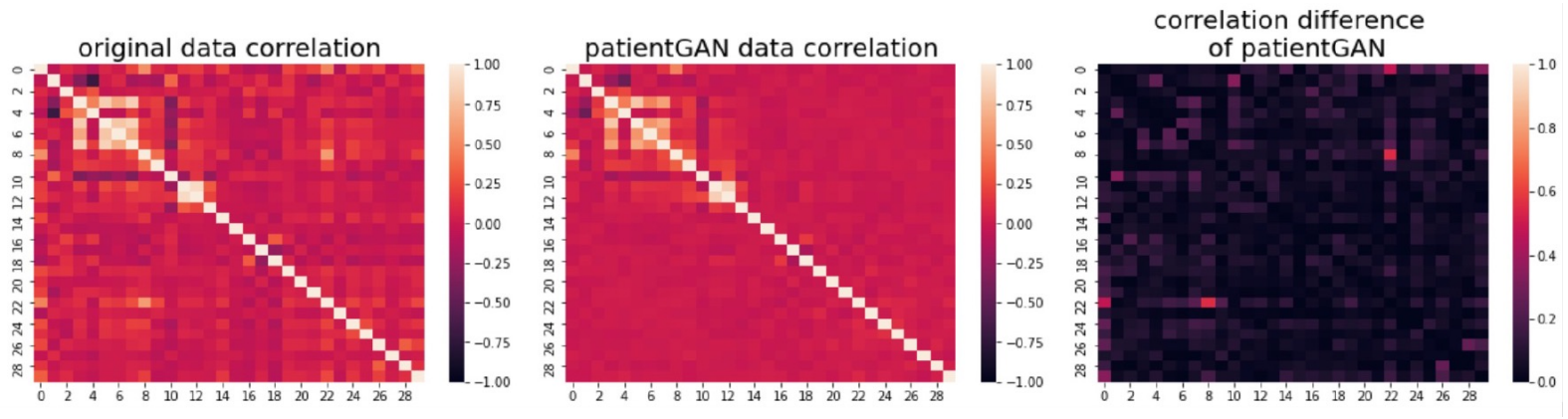

Drobig (2022)

Boehringer Ingelheim

# Synthetic data generation in clinical development – patientGAN

Results based on moderate sized tabular data set → very promising




metrical report

| mean | std | corr | replicates | duplicates | ws_bin |
|---|---|---|---|---|---|
| 21.48 | 18.54 | 0.06 | 0 | 0 | 0.05 |

| ws_con | chi2 | ks | iou | RF_F1 | BMI |
|---|---|---|---|---|---|
| 0.06 | 0.971142 | 0.992665 | 0.03 | 0.82 | 1.2 |

| ws_bin (std) | ws_con (std) | chi2 (std) | ks (std) | iou (std) | BMI (std) |
|---|---|---|---|---|---|
| 0.02 | 0.03 | 0.122595 | 0.019006 | 0.02 | 1.0 |

Drobig (2022)

Boehringer Ingelheim

# Synthetic patients for efficacy analyses: DINAMO™ trial

Lucy Fayette, Martin Oliver Sailer, and Alejandro Perez-Pitarch. "Pharmacometrics enhanced Bayesian borrowing approach to improve clinical trial efficiency: Case of empagliflozin in type 2 diabetes“. *CPT: Pharmacometrics & Systems Pharmacology* 12.10 (2023): 1386-1397.

Martin Oliver Sailer, Dietmar Neubacher, Curtis Johnston, James Rogers, Matthew Wiens, Alejandro Perez-Pitarch, Jan Marquard, and Lori Laffel. „Pharmacometrics-enhanced Bayesian borrowing for paediatric extrapolation – A case study of the DINAMOTM trial”. *PSI London, June 11-14* 2023 [link]

Boehringer Ingelheim

# DINAMO™ trial

- **SGLT2 inhibitor empagliflozin and DPP-4 inhibitor linagliptin** are well-established treatments for adults with **type 2 diabetes mellitus** (T2D)

- **Lack of oral treatments for T2D in youth**, only oral metformin and injected insulin generally approved until recent approval of GLP-1 analogues. To overcome this limitation, the **DI**abetes study of li**NA**gliptin and e**M**pagliflozin in children and ad**O**lescents (DINAMO) trial was conducted

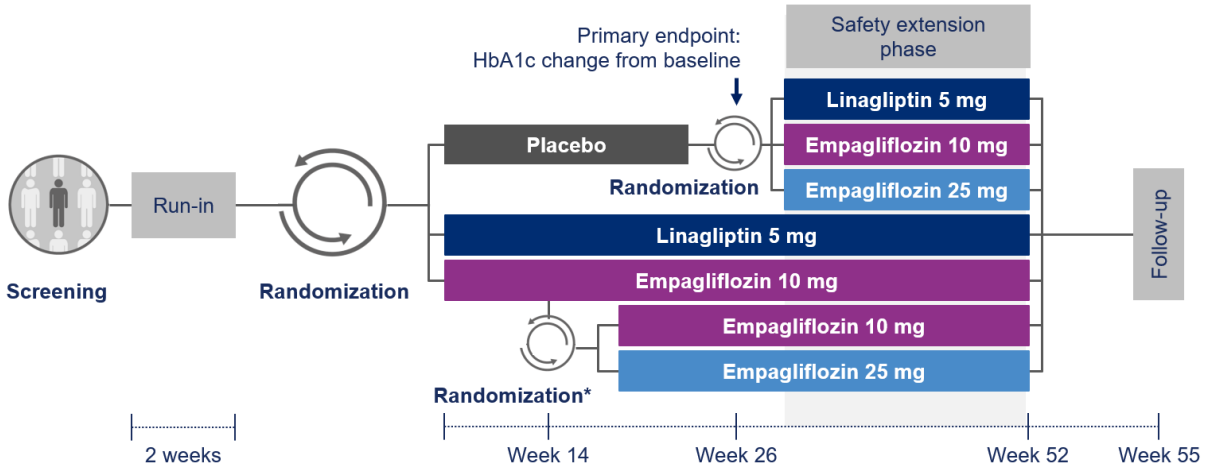
- **Main objective of the DINAMO trial:** *assess **efficacy** and **safety** of a dosing regimen with empagliflozin, with potential dose increase from 10 to 25 mg, and a single dose of linagliptin 5 mg, both compared with a shared placebo group*

Fayette, Sailer, & Perez-Pitarch (2023); Sailer et al (2023)

Boehringer Ingelheim

# DINAMO – Design and planned analyses

- Planned sample size: 150 (50 per arm). Actual sample size: **158**
- Primary endpoint: **Change in HbA1c from baseline to week 26**
- Primary comparisons:
  - Pooled empagliflozin vs placebo
  - linagliptin vs placebo
- Modified ITT analysis, using multiple imputation for missing data.
- The primary endpoint was analyzed by an ANCOVA model with baseline HbA1c as a continuous covariate, and with categorical covariates for treatment and age group
- 85% power at 5% two-sided type I error rate




Fayette, Sailer, & Perez-Pitarch (2023); Sailer et al (2023)

Boehringer Ingelheim

Laffel (2022)
HbA1c, glycated haemoglobin
* Re-randomization at week 14 for participants not achieving HbA1c <7% at week 12

# DINAMO – Sponsor proposes, reality disposes…

After recruitment was completed: high standard deviation in early blinded data → potential loss of power!

Reopening recruitment wasn‘t considered as best option

- Operational feasibility
- Substantial increase in sample size
- Substantial delay of study read-out

Study team proposed **supplementary Bayesian analysis**

- Partial extrapolation from adult data keeps original paediatric sample size
- Novel analysis method developed cross-functionally between Pharmacometrics (PMx), Statistics and Medicine:

  **Synthetic data application**

- Dedicated SAP prepared and approach discussed with FDA prior to planned read-out

Fayette, Sailer, & Perez-Pitarch (2023); Sailer et al (2023)

Boehringer Ingelheim

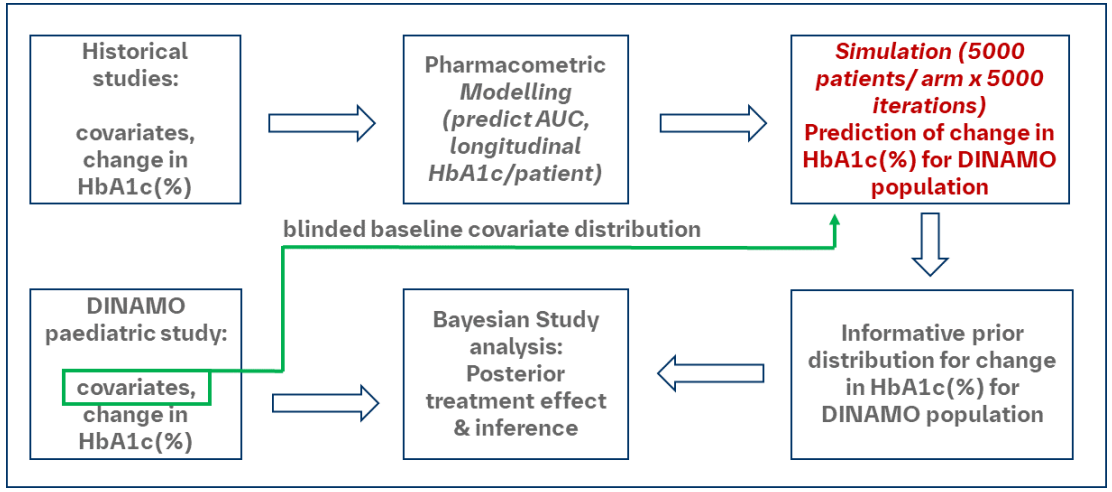
# DINAMO – What was done?

Direct borrowing from adult data not possible

- Exchangeability assumption violated between adults / children

- Typical regression models based on age does not reflect the mechanistic knowledge about the PK and PD differences between adults and children

- PMx model for change in HbA1c(%) in empagliflozin and linagliptin exists
- **PMx enhanced Bayesian borrowing** (Fayette et al. 2023)

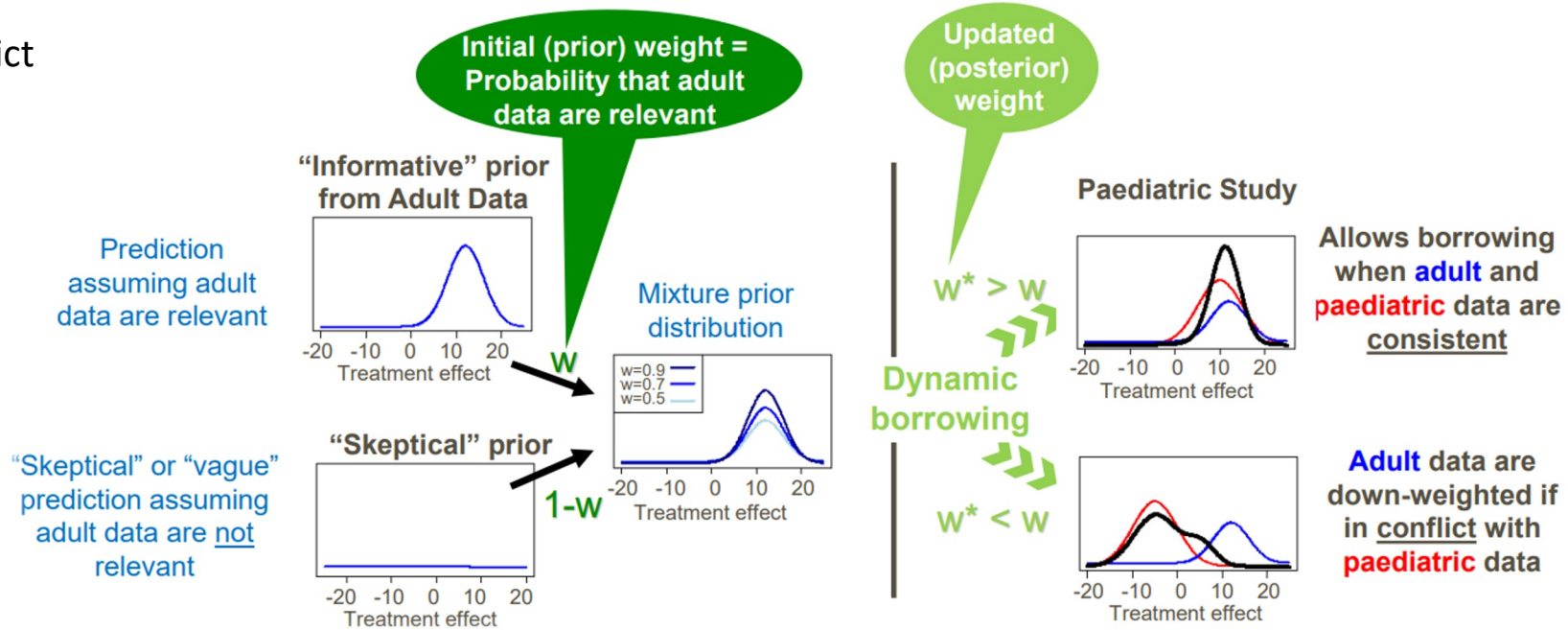



Fayette, Sailer, & Perez-Pitarch (2023); Sailer et al (2023)

Boehringer Ingelheim

# DINAMO – Robust mixture prior

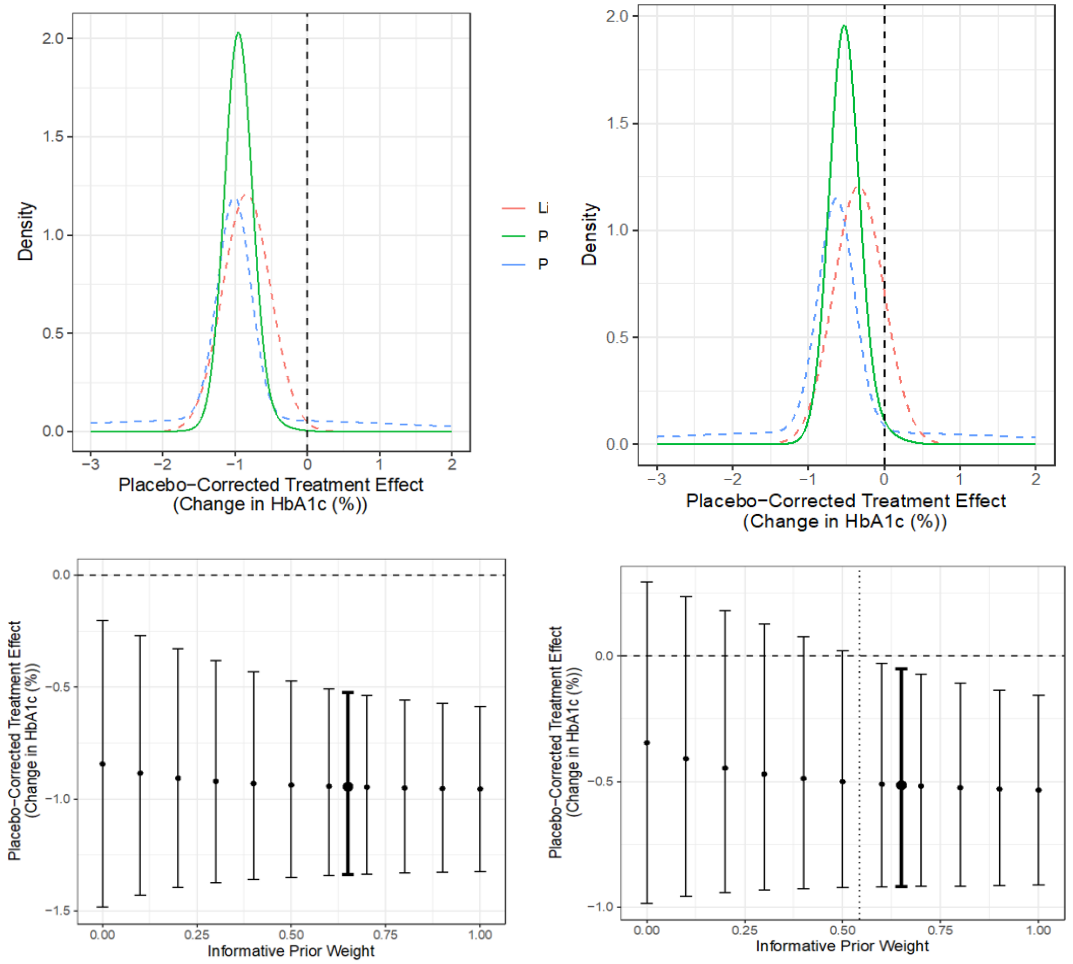
- Weights for the mixture prior chosen such that $ESS_{ELIR}$ (Neuenschwander et al. 2020) equals planned sample size
- Informative prior distributions are derived from the synthetic data: mean (of means) and variance (of means)
- Protects against prior-data conflict




Fayette, Sailer, & Perez-Pitarch (2023); Sailer et al (2023)

# DINAMO – Outcome

- For empagliflozin, Bayesian analyses confirmed evidence for **meaningful efficacy** (in line with primary analysis)
- For linagliptin, Bayesian analyses provided evidence for **superiority** (whereas primary analysis did not)
- Regulatory agencies FDA and EMEA both provided positive comments on the application
- DINAMO showed that empagliflozin dosing regimen provided **clinically and statistically meaningful reductions** in HbA1c in youth with T2D!




Fayette, Sailer & Perez-Pitarch (2023); Sailer et al (2023)

# Thank you for your attention

- Synthetic data can be a **valuable** and **useful tool** not only in clinical drug development, but brings **possibilities for numerous challenges** in clinical contexts – including tabular data enrichment

- Regulatory authorities **encourage early engagement** and discourse when such innovative approaches are considered

- Use of such approaches is **dependent on strong scientific merit** and **practical rationale**

- We must **systematically ensure** that our data generation approaches lead to trustworthy results

Boehringer Ingelheim