Data augmentation approaches for animal data in toxicology studies

- Jacobi P.[1] & Matyjasiak J.[2]

1. Syncwork AG, Berlin, Germany (peter.Jacobi@syncwork.de)
2. Department of Pharmaceutical and Pharmacological Sciences, Vrije Universiteit Brussel, Brussels, Belgium (Julia.Anna.Matyjasiak@vub.be)

FACTS & FIGURES

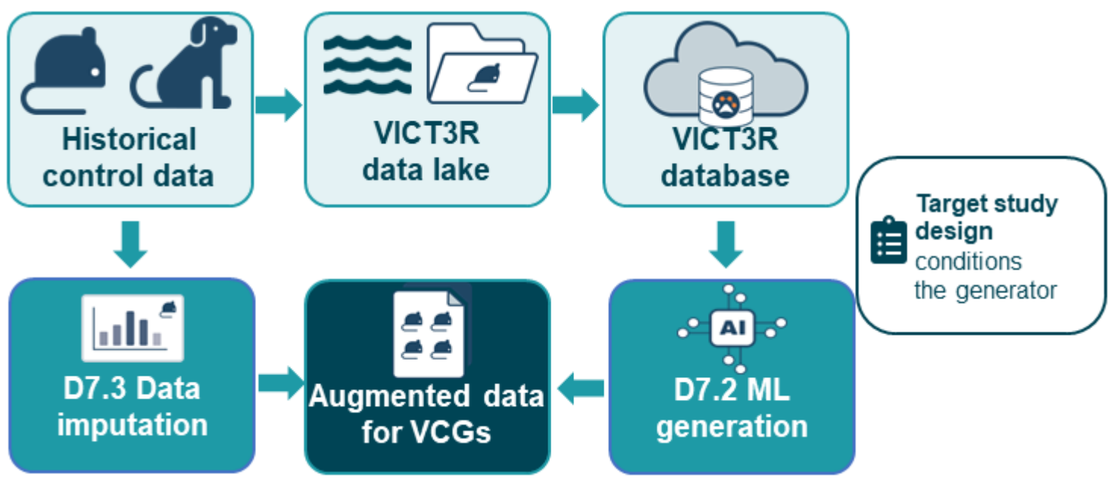
Start date:	1st September 2024
End date:	29th February 2028
IHI funding:	13,524,750 €
Industry contribution:	13,581,025 €
Contributing partners:	80,951 €
Total Cost:	27,186,726 €
Project website:	vict3r.eu
Social media:	linkedin.com/company/ihi-vict3r/

CHALLENGE

Every regulatory toxicology study includes a **concurrent control group** (CCG), animals that **receive no test article**. CCGs are still a scientific and regulatory requirement representing a large, recurring **use of laboratory animals**. **Virtual control groups (VCGs)** could **replace** or supplement CCGs with historical and synthetic control data. Two obstacles stand in the way:

- Statistical faithfulness and analytical utility of **AI generated control animals**
- **Missing data** within the VICT3R legacy control database

APPROACH & METHODOLOGY



D7.2 — Generation

A configuration-driven **pipeline** for animal data in the industry Standard of Exchange for Non-Clinical Data (**SEND**) that runs unchanged across studies.

SEND data → subject-wide table → conditional generative model → SEND data

- Four model families compared:
  - CVAE · GMM · Diffusion · CTGAN
- Optuna tuning · MLflow tracking
- Evaluated on four axes:
  - diagnostics · fidelity · utility · privacy

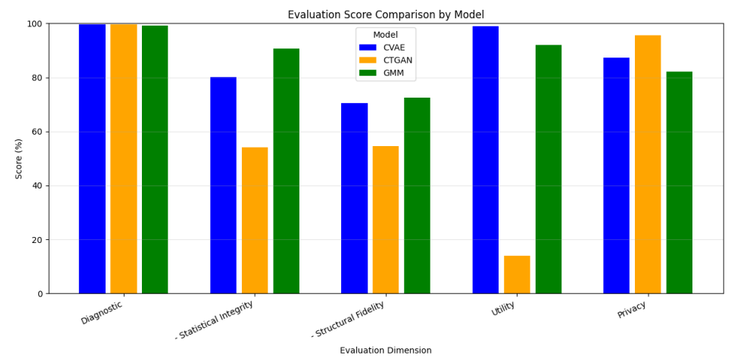
D7.3 — Imputation

Imputation on **masked ground truth**: a dataset of highly complete Wistar Han study body weights, organ weights, laboratory values.
Masking was done using **modelled real gap patterns taken from VICT3R DB**, **transferred** onto the dataset (3 approaches: **random** and proportional to VICT3R availability, **donor matching** to VICT3R subjects and **clustering** to VICT3R missing data patterns). Baseline median fill, various statistical **and machine learning** methods were compared with **generative models**.

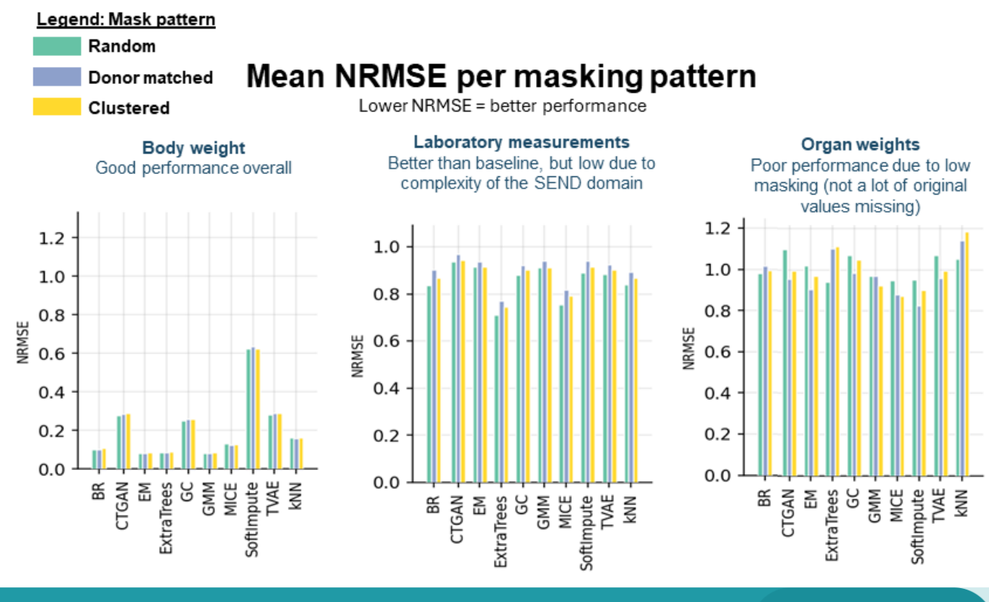
RESULTS

Impute what can be imputed, generate from there.

D7.2 — Generation



D7.3 — Imputation



VALUE OF IHI COLLABORATION

- Pooled industry control data supply the **realistic gap patterns** that make the benchmark meaningful and no single partner holds them
- **Complementary teams** per Work Package and deliverable
- **SEND as a shared standard** makes the pipeline transferable across the consortium instead of one lab
- Public deliverables and open repositories turn **method development into fewer animals used**

CAREER IMPACT

Hands-on experience in **database organization, data analysis**, and **reproducible** machine-learning workflows. Strengthened ability to **collaborate** effectively, **manage expectations** and **support the translation** of **analytical results** into **reliable solutions for non-clinical research**.

This project is supported by the Innovative Health Initiative Joint Undertaking and its members.