AI-driven synthetic patient generation framework to model obesity using the Constances cohort

Authors: Diane Vincent, MSc[1], Louise Dry, MSc[1], Nina Temam, PharmD[1], Antoine Movschin, MSc[1], Sofiane Kab, PharmD PhD[2], Antoine Duclos, MD PhD[2,3], Pauline Guilmin, MSc[1], Nicolas Lambert, MSc[1]

Affiliation: [1]Quinten Health, Paris, France ; [2]INSERM, Villejuif, France ; [3]Université Paris Cité (CRESS U1153), Paris Saclay University, Université de Versailles Saint-Quentin-en-Yvelines, INSERM UMS 011, "Population-based Cohorts Unit", Paris & Université Claude Bernard Lyon 1 (HCL), France

MSR130

ISPOR Europe 2026

quinten health

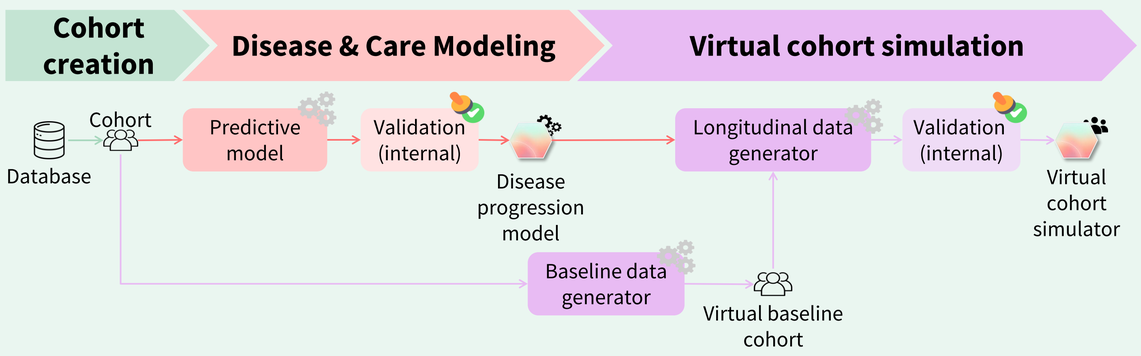
INTRODUCTION

- **Overweight or obesity** affects 39% of the global adult population and is a chronic, multifactorial disease associated with a substantial cardiovascular burden and highly heterogeneous profiles[1], creating challenges for evidence generation and healthcare decision-making.
- **Synthetic data generation (SDG)** is an innovative approach that creates **virtual patient records** mimicking the statistical properties, complex relationships, and **clinical patterns of real-world data** while strictly **preserving patient privacy**[2].
- In heterogeneous conditions like obesity, artificial intelligence (AI)-driven SDG offers a powerful toolkit to safely model and simulate complex longitudinal care pathways. More broadly, synthetic data enable **privacy-preserving data sharing**, **cohort augmentation**, **trial design optimization**, **external control arm enrichment**, **treatment effect generalization** and **disease scenario simulation**[2].

- **Objective: We aimed to develop an AI-driven framework capable of generating and evaluating realistic synthetic real-world obesity patient cohorts that preserve both patient heterogeneity and privacy.**

METHODS

Figure 1: Synthetic data generation framework



Cohort Creation

- Data were obtained from the **French Constances**[3] **cohort**, **linked to the national health claims database (SNDS)**. Adults with obesity were identified via measured or self-reported BMI ≥30 kg/m² or ≥27 kg/m² with weight-related comorbidities (diabetes, high blood pressure, abnormal levels of fats in the blood, obstructive sleep apnea, history of cardiovascular diseases)[4]. Patients who had undergone bariatric surgery or had cancer related to hormones or associated with weight change were excluded.
- The index date was defined as the first BMI record date of overweight or obesity observation in Constances.
- **The outcome was the time to occurrence of a 5-point major adverse cardiovascular event (5P-MACE) up to 6 years**, derived using the SNDS. 5P-MACE is defined as one of the following events: non-fatal myocardial infarction, non-fatal stroke, coronary revascularization, unstable angina requiring hospitalization or CV-cause death[5]. Due to unavailability of cause-specific mortality data, all-cause of death was used instead.
- **Baseline characteristics** include **socio-demographics**, **comorbidities**, **treatments**, **paraclinical data**, **quality of life** and **lifestyle**. Preprocessing included **missing values imputation** to the mode for categorical features and to the average of the 10 nearest neighbors for numerical features.
- **Clinically relevant subgroups** were defined by **sex**, **diabetes** and **cardiovascular disease** history.

Disease and Care Modeling

- A **Cox Proportional Hazard** (CPH) model was trained on 70% of the cohort to **predict time to 5P-MACE** and **internally validated** on a 30% test set using the Concordance Index (C-Index, risk ranking accuracy) and Integrated Brier Score (IBS, risk prediction error over time).

Virtual Cohort Simulation

- The synthetic data generation (SDG) process consisted of two steps:
  1. **Synthetic baseline characteristics** were generated using a **Gaussian copula**[6] trained on the 70% training set.
  2. **The time-to-event outcome was simulated** using the **disease model** (CPH) and **inverse transform sampling** (ITS).

Virtual Cohort Evaluation

- **Fidelity** - *How closely does the synthetic data resemble the real data?* - was evaluated with distance (Wasserstein distance) and discrimination (propensity score mean-squared error (pMSE): performance of a classifier distinguishing synthetic and real patients) metrics, and comparison of descriptive statistics of synthetic and real test set characteristics.
- **Utility** - *Does the synthetic data support downstream predictive tasks?* - was evaluated by comparing synthetic and real test set Kaplan-Meier (KM) plots across entire cohorts and subgroups, and the predictive performance of a CPH model trained on synthetic versus real data.
- **Privacy** - *Is the privacy of the original data preserved?* - was evaluated on the training set using distance-based metrics: the fifth percentiles of distance to closest record (DCR, distance between any synthetic record and its closest real neighbor), nearest neighbor distance ratio (NNDR, ratio of the distance between any synthetic record and its first and second closest real neighbors) and ε-identifiability, the proportion of real records closer to a synthetic record than to another real one.

Figure 2: Attrition chart

N = 49,095	Adults with BMI ≥30 kg/m² or ≥27 kg/m² with at least one weight-related comorbidity
N = 48,504	No bariatric surgery
N = 47,475	No cancer related to hormone or associated with weight change
N = 41,495	Sufficient information in SNDS

SNDS, Système National des Données de Santé

RESULTS

- The cohort includes N=41,495 patients with obesity (Figure 2).
- **The CPH disease model shows high predictive performance**, with a C-Index of 0.76 and an IBS of 0.02 on the test set.
- ***Fidelity*** **- Synthetic cohorts closely match the statistical properties of the original test populations** (Figure 5), achieving a Wasserstein distance of 0.13 and a normalized pMSE of 0.51.
- ***Utility*** **- KM survival curves for synthetic and original test cohorts overlap closely** across the overall cohort and subgroups, except for patients with diabetes after 2 years (Figure 6). CPH model exhibits similar predictive performance when trained on synthetic data (Figure 4).
- ***Privacy*** - Distance-based privacy metrics and ε-identifiability suggest **strong privacy preservation** (Figure 3).

Figure 3: SDG privacy evaluation

Vs. real train set (N=29,046)	Synthetic (N=29,046)	Real test set (N=12,449)
NNDR	0.63	0.59
DCR	0.91	0.88
ε-Identifiability	0.04	-

SDG, Synthetic Data Generation; NNDR, Nearest Neighbor Distance Ratio; DCR, Distance to Closest Record

Figure 4: SDG utility evaluation

Metrics (test set)	Model trained with Synthetic	Model trained with Real train set
C-Index	0.76	0.76
IBS	0.02	0.02

SDG, Synthetic data generation; C-Index, Concordance Index; IBS, Integrated Brier Score

Figure 5: SDG fidelity evaluation – Descriptive statistics of real test set and synthetic cohorts

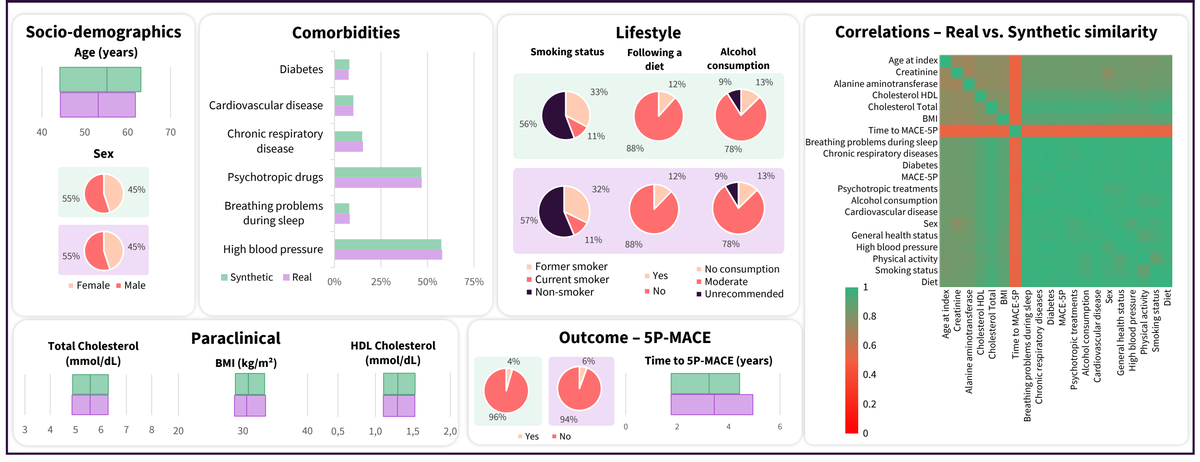
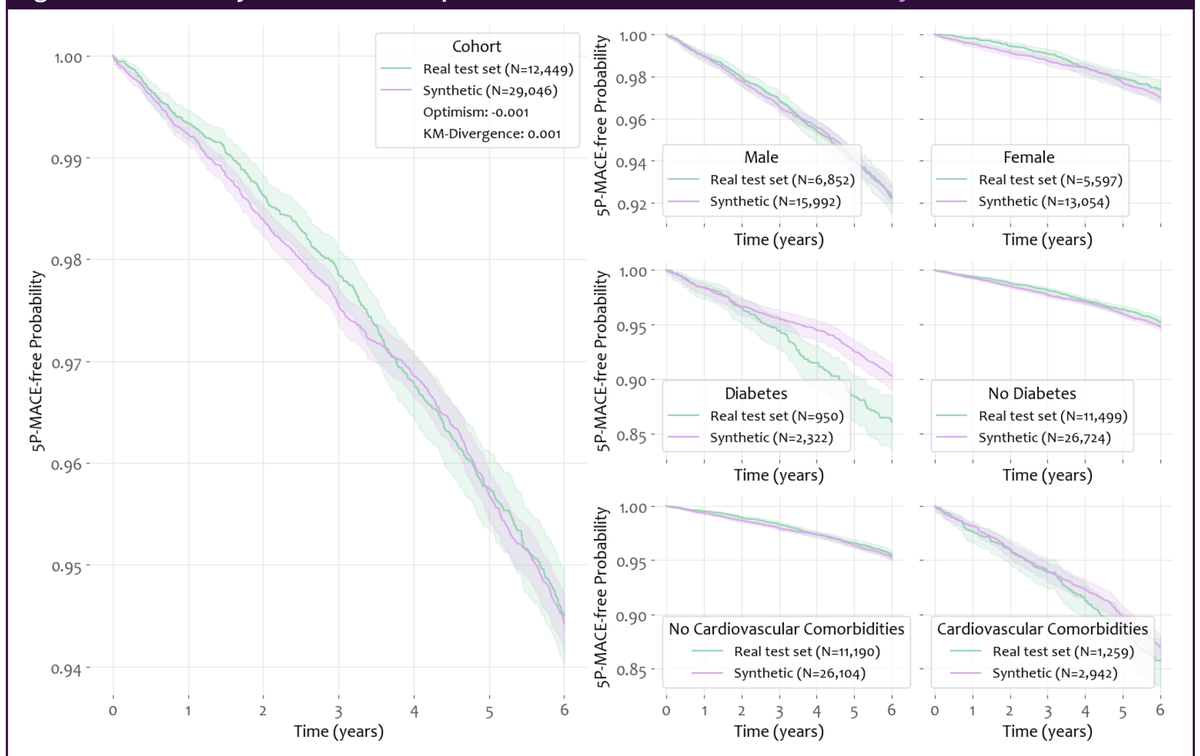


Figure 6: SDG utility evaluation – Kaplan Meier curves of real test set and synthetic cohorts



Optimism is the mean difference and KM-divergence is the absolute mean difference between real and synthetic KM curves.

DISCUSSION

- The strong performance of the disease model combined with the high fidelity of generated baseline characteristics demonstrate that this **AI-driven framework successfully generates realistic, privacy-preserving patient trajectories**.
- This approach provides immediate value for **augmenting underrepresented patient subgroups** and **simulating alternative care pathways** without exposing sensitive data.
- **Future work** will **expand the evaluation framework** with additional privacy and cross-dimensional metrics to better quantify privacy - utility - fidelity trade-offs.
- Additionally, exploring **advanced deep generative architectures** (such as VAEs, GANs, or Transformers) or **Bayesian networks,** and integrating explicit **clinical rule** engines will allow the expansion of this framework toward **multi-timepoint, multi-output longitudinal trajectories while guaranteeing strict medical consistency**.

- **This AI-driven framework generated realistic synthetic obesity cohorts that may enable augmentation of under-represented subgroups or simulation of alternative therapeutic trajectories.**
- **Thereby it can support research on exacerbation patterns, treatment sequencing, progression, cardiovascular outcomes, long-term prevention strategies and healthcare resource use to inform clinical and healthcare decision-making in obesity.**

METRICS INTERPRETATION:

- Survival: C-Index ≥ 0.7 and IBS ≤ 0.1 indicate good performance.
- Privacy: NNDR < 0.50 or a synthetic DCR *lower* than the real baseline suggests potential data leakage; ε ≤ 0.05 indicates strong privacy.
- Fidelity: Wasserstein distance near 0 and normalized pMSE near 1 indicate strong statistical fidelity.

ABBREVIATIONS:

BMI, Body Mass Index; C-Index, Concordance Index; CPH, Cox Proportional Hazard; DCR, Distance to Closest Record; HDL, High-Density Lipoprotein; IBS, Integrated Brier Score; ITS, Inverse Transform Sampling; KM, Kaplan-Meier; NNDR, Nearest Neighbor Distance Ratio; pMSE, Propensity Score Mean-Squared Error; SDG, Synthetic Data Generation; SNDS: Système National des Données de Santé

REFERENCES:

1.Fava, S. et al. (2019). Obesity and cardio-metabolic health.
2.Gonzales, A. et al. (2023). Synthetic data in health care: A narrative review.
3.Zins, M. et al. (2015). The French CONSTANCES population-based cohort: design, inclusion and follow-up.
4.European Medicines Agency. (2016). Guideline on clinical evaluation of medicinal products used in weight management (EMA/CHMP/311805/2014).
5.Bosco, E. et al. (2021). Major adverse cardiovascular event definitions used in observational analysis of administrative databases: a systematic review.
6.Patki, N. et al. (2016). The synthetic data vault.

ACKNOWLEDGEMENT: The authors thank the team of the Epidemiological Population Cohorts Unit in France (UMS 011 – Inserm, Universités Paris Cité, Paris Saclay, Versailles Saint-Quentin), which designed and manages the Constances French national cohort ; doi.org/10.13143/inserm_constances. They also acknowledge the French national health insurance fund (Caisse nationale d'assurance maladie, CNAM) and its Health examination centers for collecting a large part of the data, as well as the French national old-age insurance fund (Caisse nationale d'assurance vieillesse, CNAV) for their contribution to the establishment of the cohort, and ClinSearch for performing data quality control. Access to confidential data, on which is based this work, has been made possible within a secure environment offered by CASD – Centre d'accès sécurisé aux données (Ref. 10.34724/CASD).

CONFLICT OF INTEREST: The Constances cohort is supported and funded by the French national health insurance fund (Caisse nationale d'assurance maladie, CNAM). Constances is recognized as a "National infrastructure in biology and health" and receives funding from the French national research agency (ANR-11-INBS-0002 and ANR-23-INBS-0001). The cohort is also marginally funded by private sector companies, notably from the healthcare industry, within the framework of public-private partnerships. None of these funding sources played any role in the study design, data collection and analysis, or the decision to publish the findings.

FUNDING: This work was supported by Quinten Health.

CONTACT INFO: **contact@quinten-health.com**