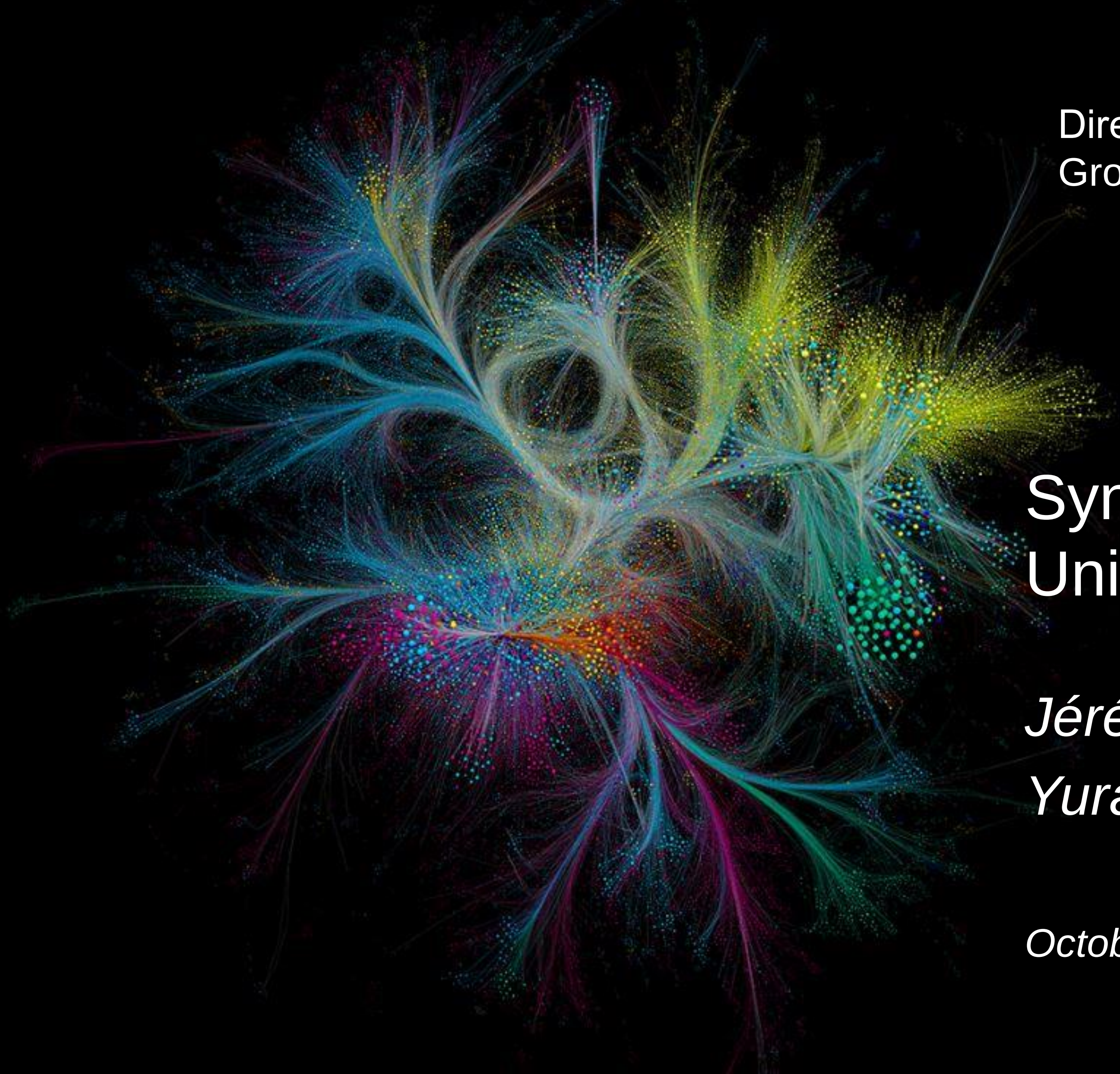


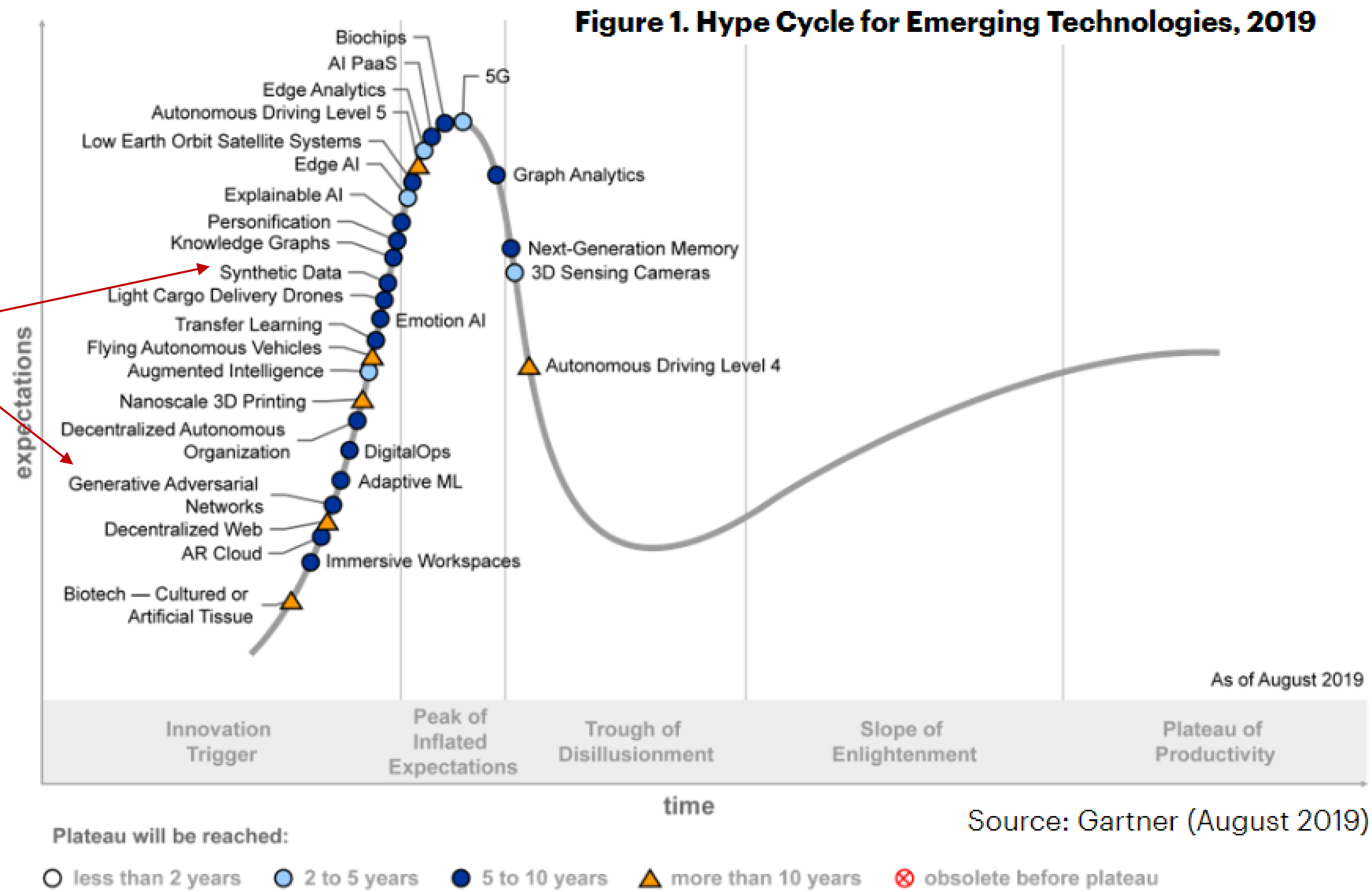
Direction des Systèmes d'Information
Groupe Data Science

Synthetic Data in a University Hospital Context

Jérémie Despraz
Yura Tak

October 2021

GANs and synthetic data : a hot topic



Source: Gartner (August 2019)

GANs and synthetic data : a hot topic

sarus

MOSTLY·AI

hazy

Statice

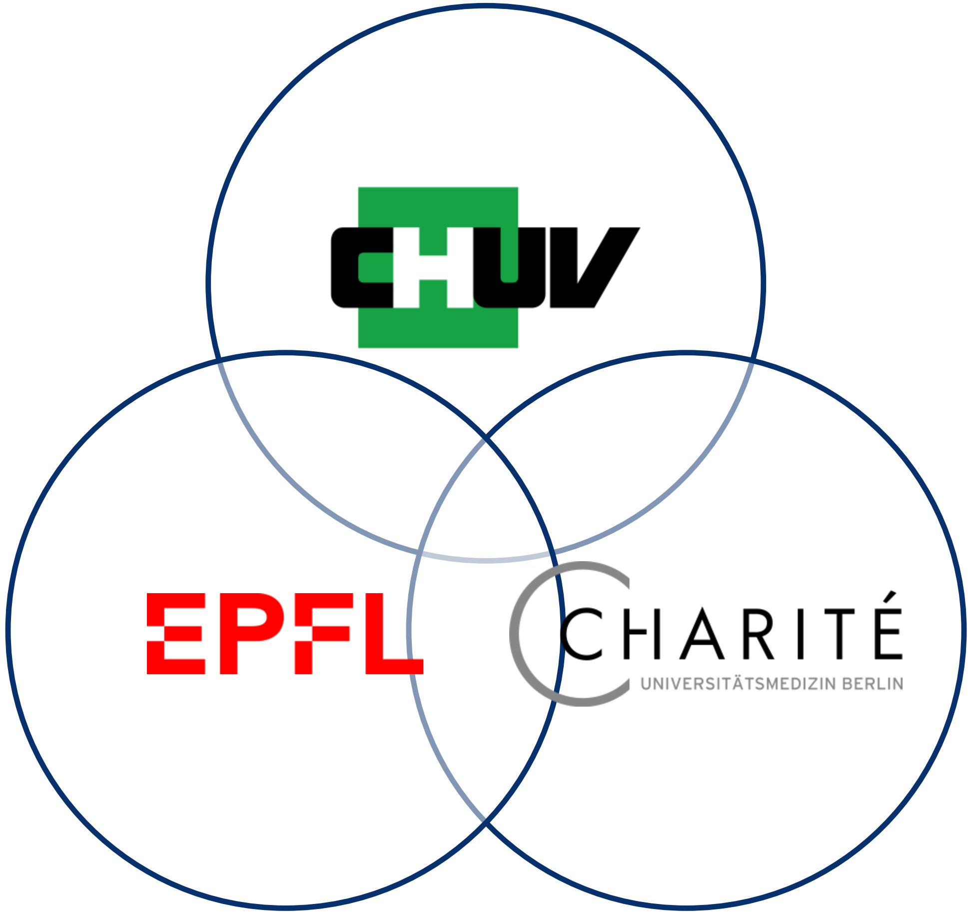
SYNTHESIZED

SYNTHEA

OSCILLATE.AI

Replica Analytics

Collaboration



Outline

1. Why do we need synthetic data
2. How do we create synthetic data
3. Use cases for synthetic data at CHUV
4. Our contributions and preliminary results
5. Take home messages

Motivation

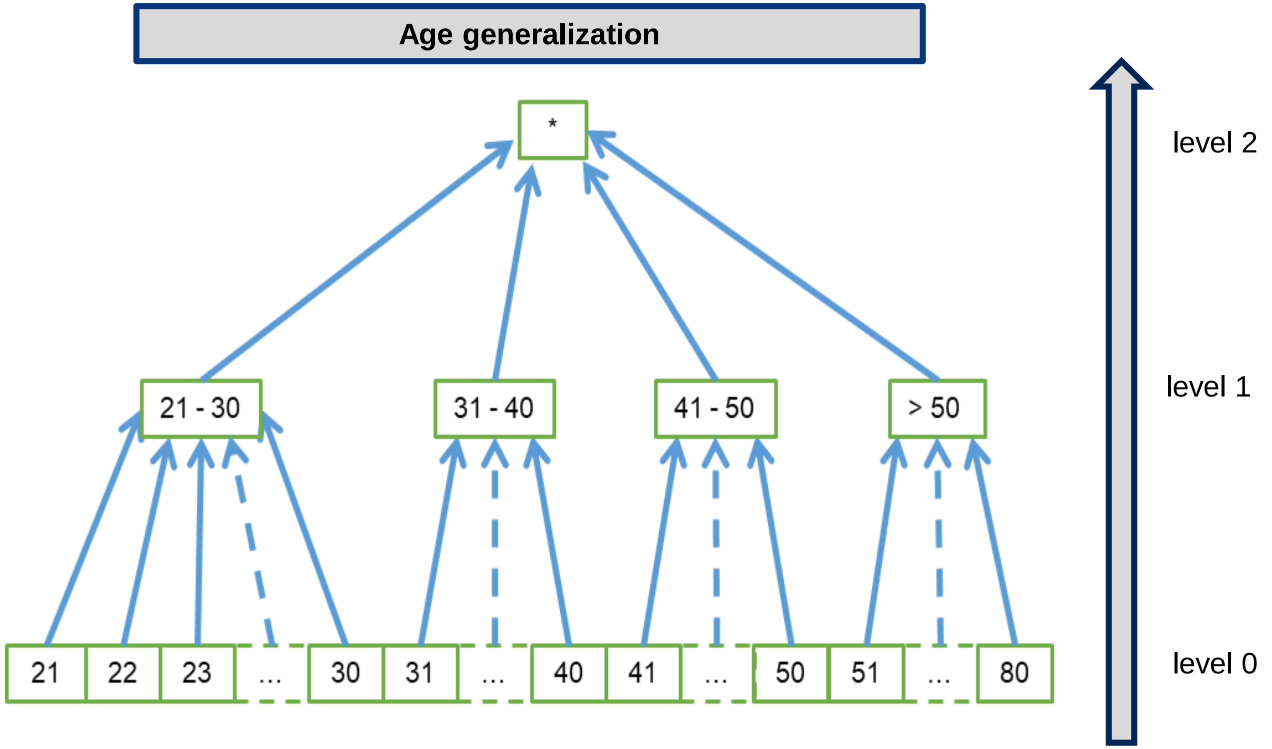
Access to medical data is strongly regulated

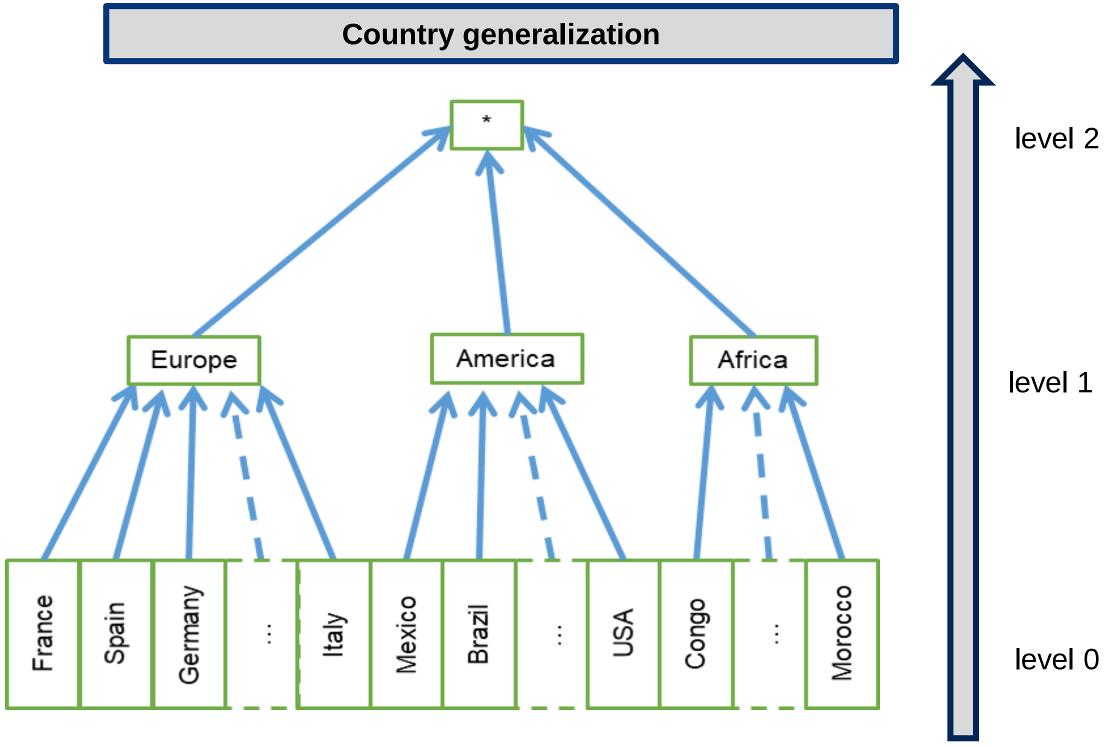
	• Loi relative à la recherche sur l'être humain (**LRH**) • Ordonnance relative à la recherche sur l'être humain (**ORH**) • Loi fédérale sur la protection des données (**LPD**) • Loi sur la protection des données personnelles (**LPrD**)
	• General Data Protection Regulation (**GDPR**)
	• Health Insurance Portability and Accountability Act (**HIPAA**) Privacy Rule

→ These regulations require health-related data to be *de-identified* to mitigate privacy risks

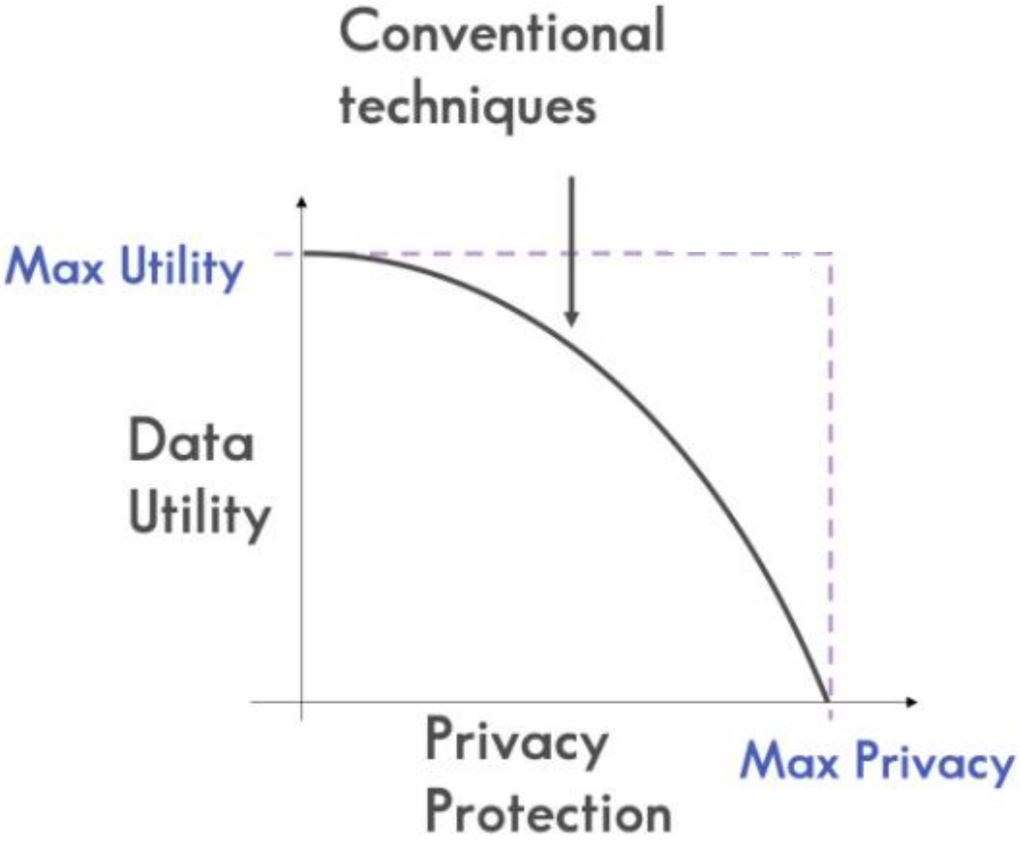
Motivation

Enforcing de-identification leads to information removal





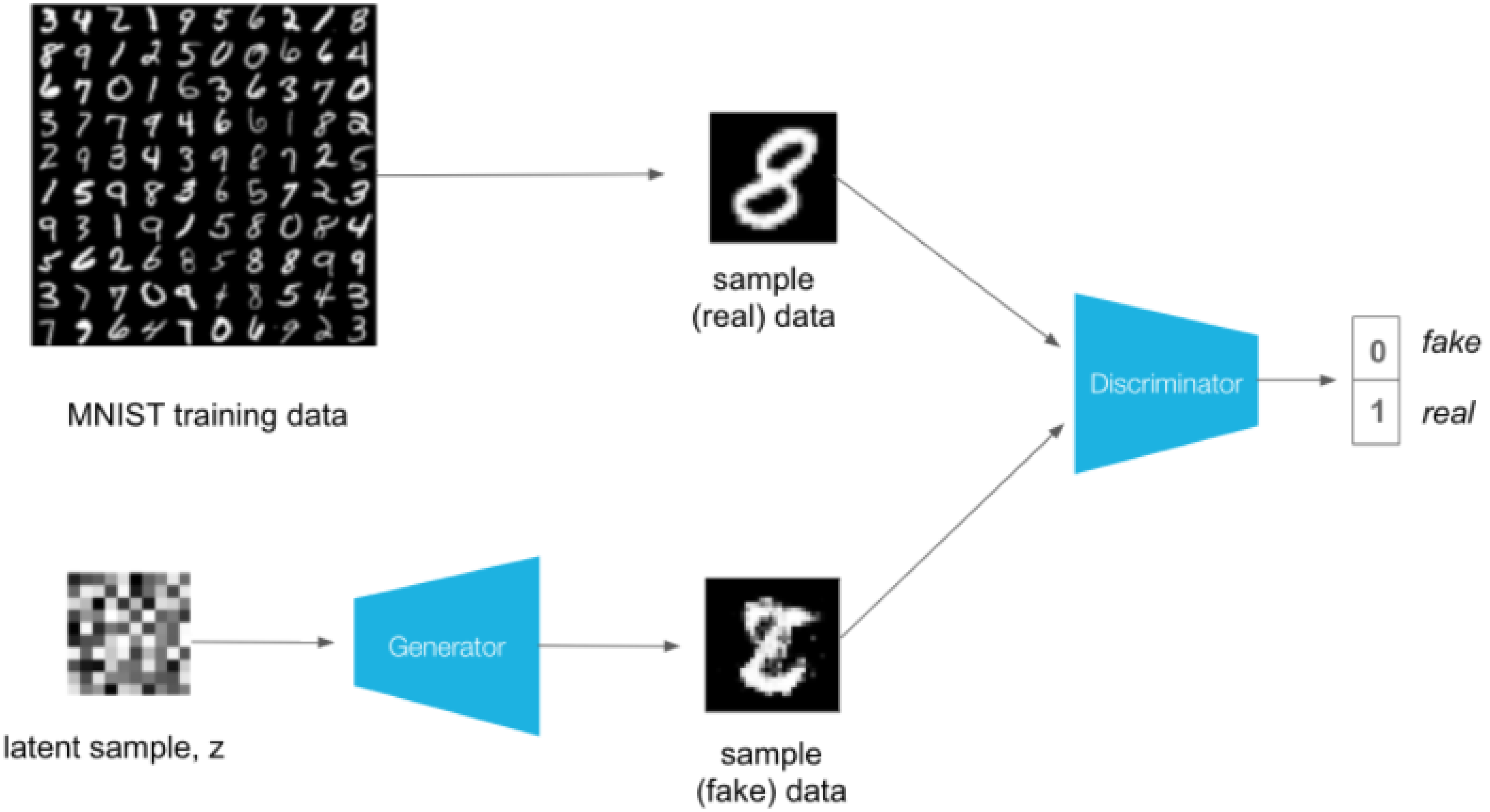
Motivation



Strong privacy requirements come at the cost of reduced data utility due to the removal of sensitive information or the addition of noise.

Can’t we have both privacy and utility at the same time?

Generative Adversarial Network



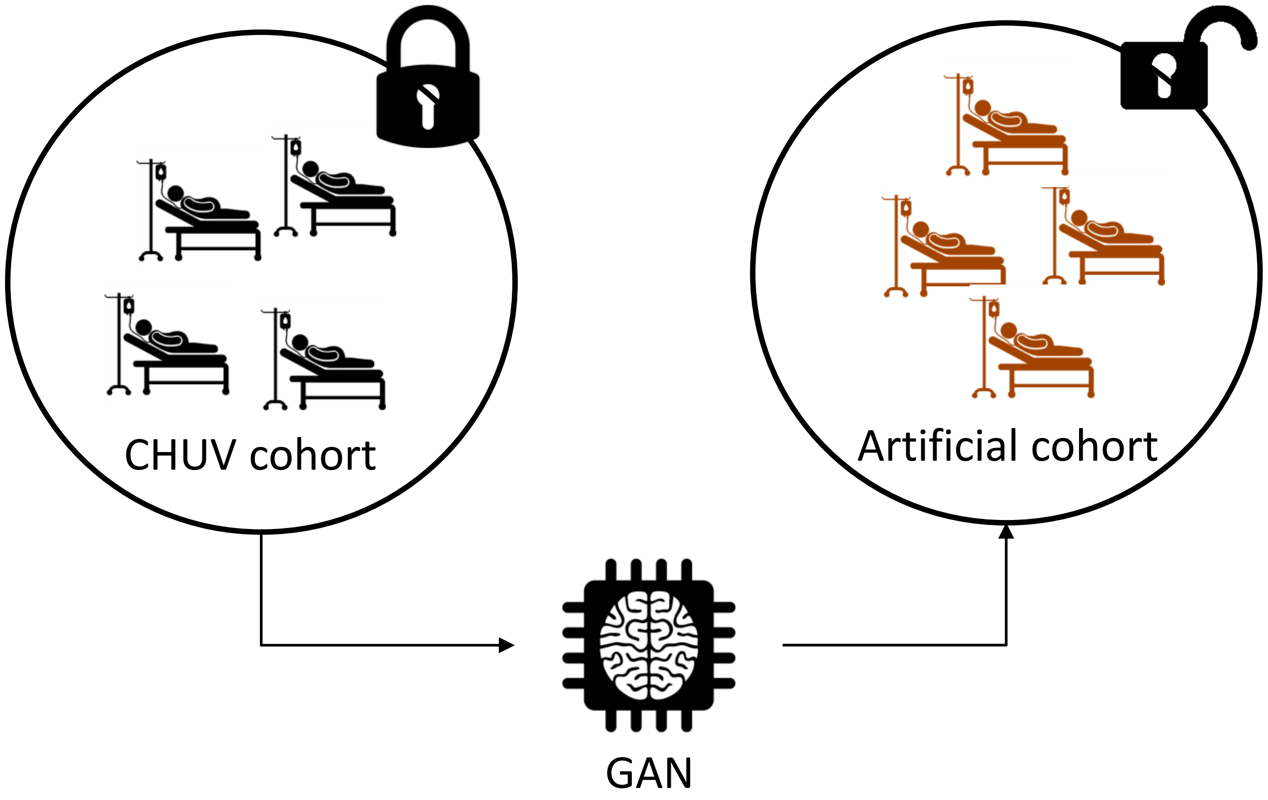
Generative Adversarial Network

This person does not exist and has been entirely created by a GAN called *StyleGAN** trained on many real portraits.

Question:
Can we do the same with data other than pixels?
Can we generate fake electronic health records based on existing data?

(*) Karras, Tero, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.

Synthetic patients



Learning from the existing CHUV electronic health records, we create artificial patients cohorts

Use-cases for synthetic data at CHUV

Standard workflow of a research project

Idea

↓

Cohort determination

↓

Ethics committee review

↓

Extraction and coding

↓

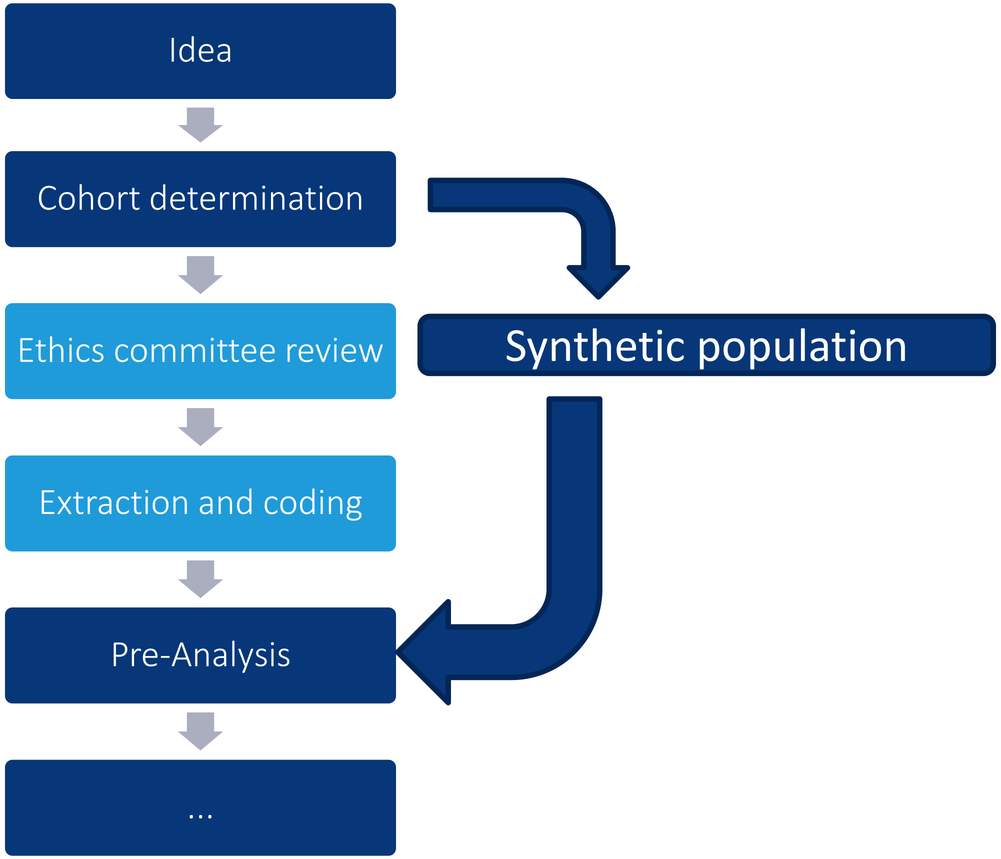
Analysis

↓

Publication

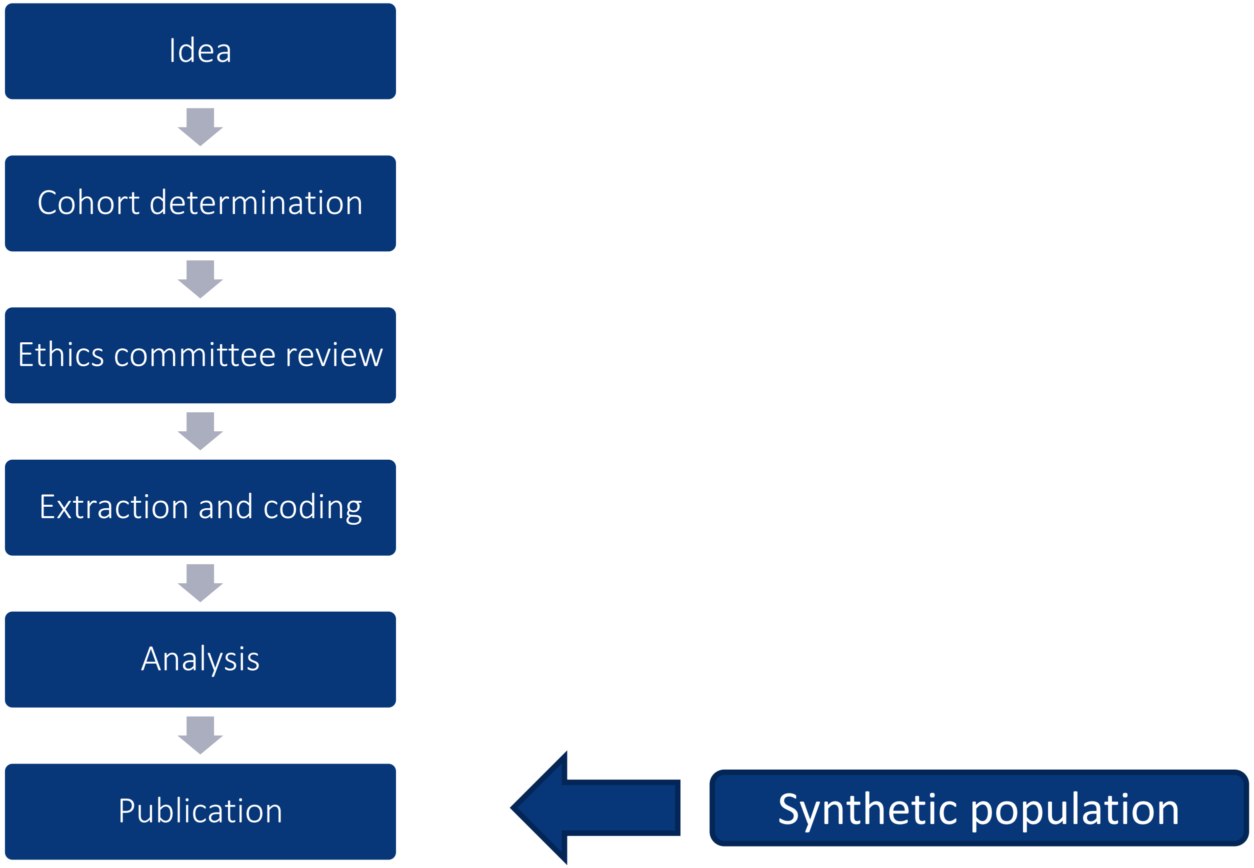
Facilitate data access

Synthetic data is used to validate the scientific hypothesis prior to ethics committee review.



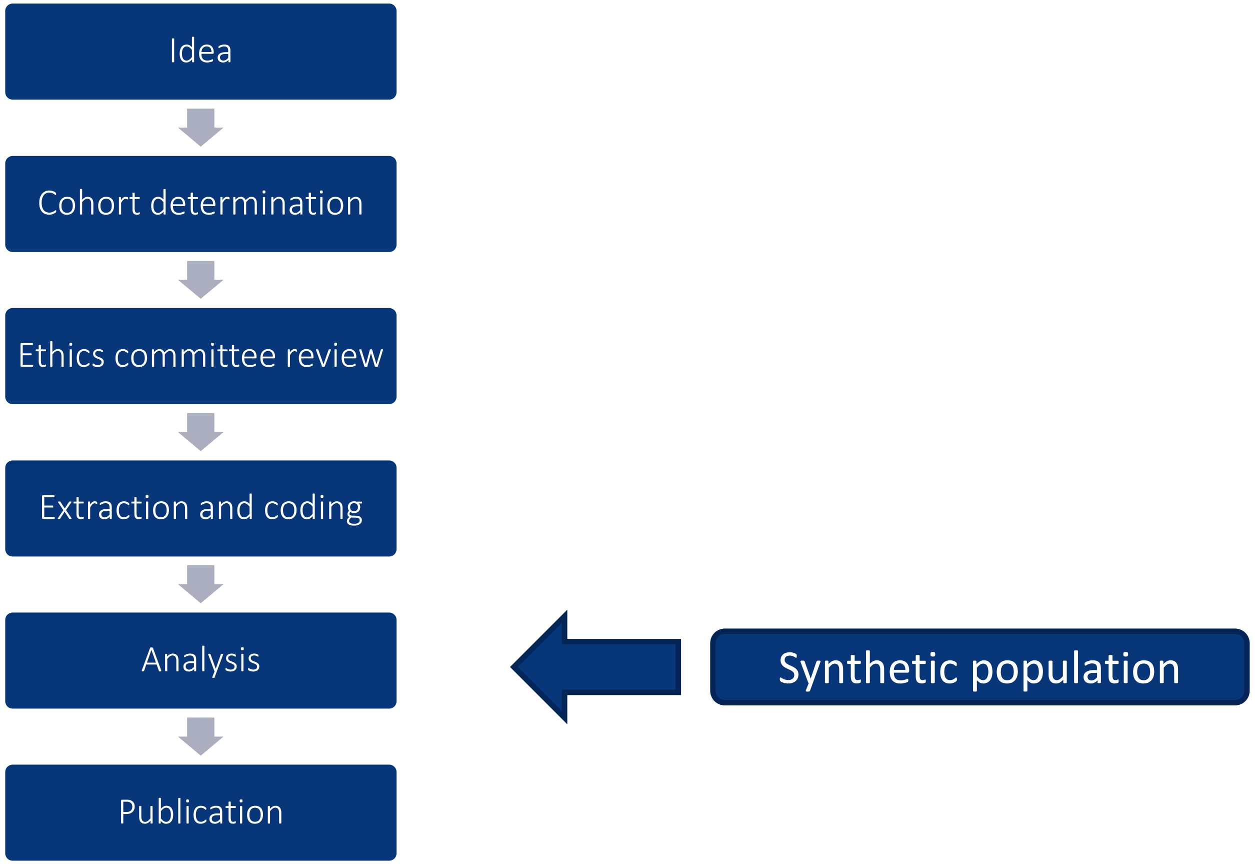
Facilitate data sharing

Numerous medical papers do not share their data. With synthetic cohorts, researchers could allow their peers to reproduce their results and perform further research.

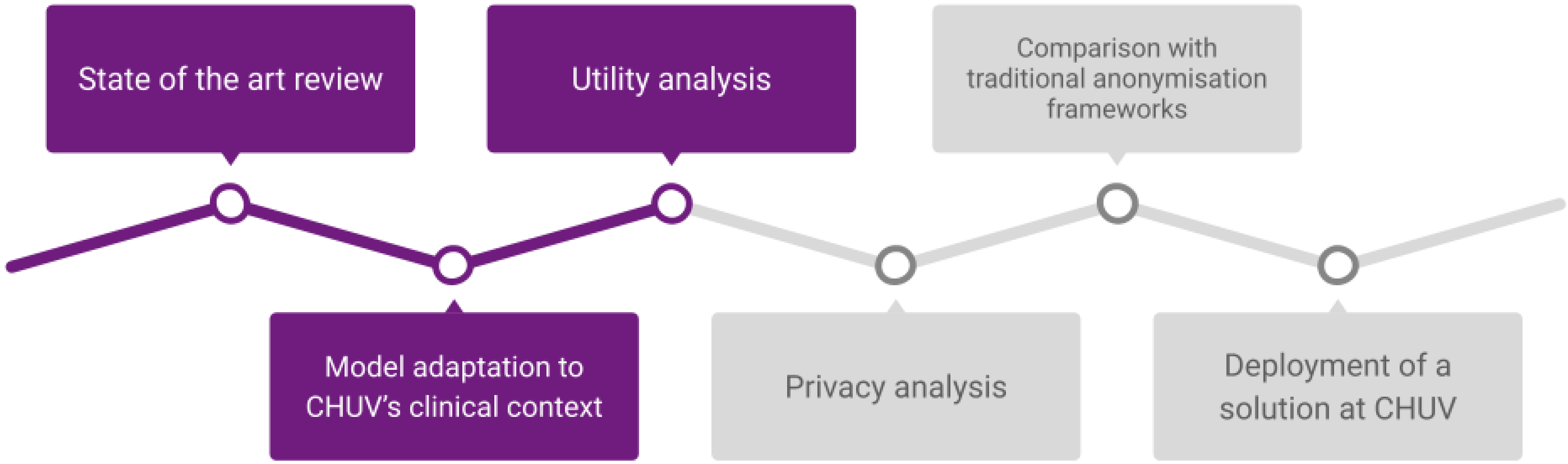


Allow for data augmentation

Synthetic datasets can augment the amount of training data to produce more robust statistical models.



Project roadmap



Our Contributions

- **Testing state-of-the-art models with real data**: From a critical literature review, two state-of-the-art models are selected and each model is tested with real clinical data.

- **Comparing different sparsity handling methods**: Implementing a sensible imputation method to fill missing values and 3 methods to reproduce data missingness in the synthetic data.

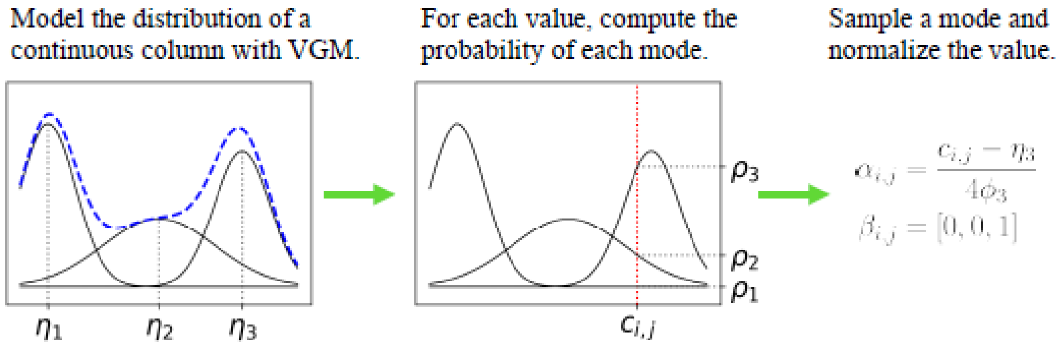
- **Model adaptation** : modification of the models to handle continuous clinical values and switch from visits to observations.

- **Assessing the similarity between real and synthetic data**: Using several similarity measures, such as dimension-wise statistics, conditional distributions or temporal evolution, we assess the similarity between the original and the generated data.

- **Reproducing medical studies on synthetic data**: Using different datasets, we construct a syntheric cohort, reproduce the entire workflow and compare the results.

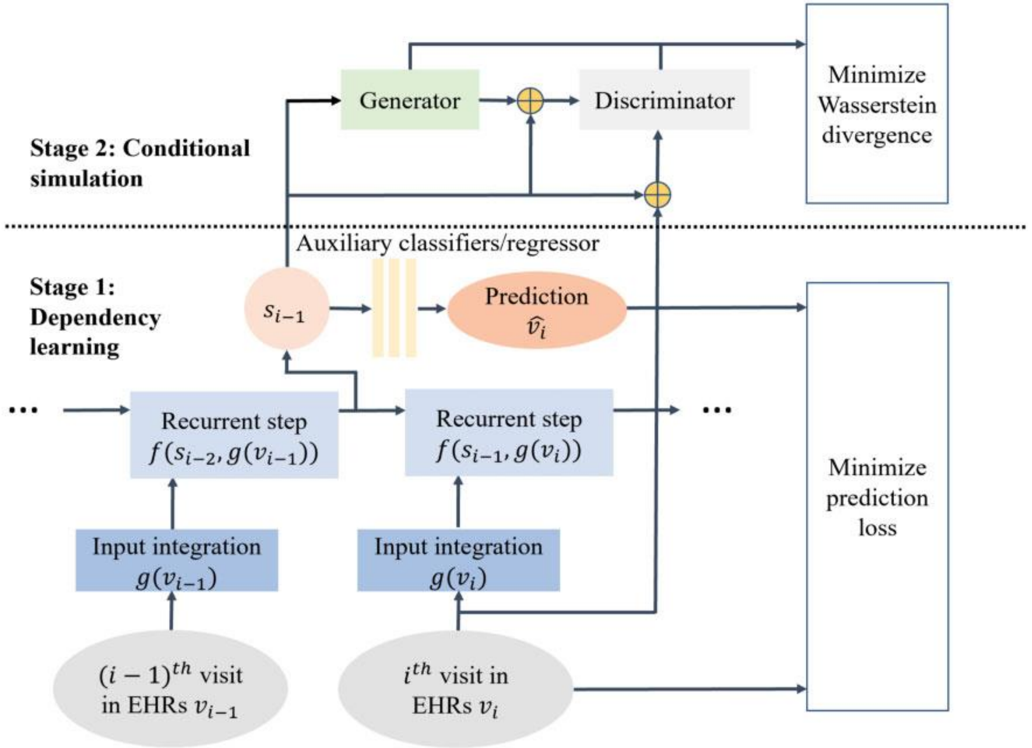
Selected Models

CT-GAN



Modeling Tabular data using Conditional GAN, Xu et al., 2019

SynTEG



SynTEG a framework for temporal structured electronic health data simulation, Zhang et al., 2020

Case Studies

1. **Bedsore Data** : SynTEG

2. **Visceral Surgery Data** : CTGAN

3. **Pharmacokinetics Data** : SynTEG

Case Study 1 : Bedsore Data

Data: 512 patients, 622 distinct stays, 11 attributes

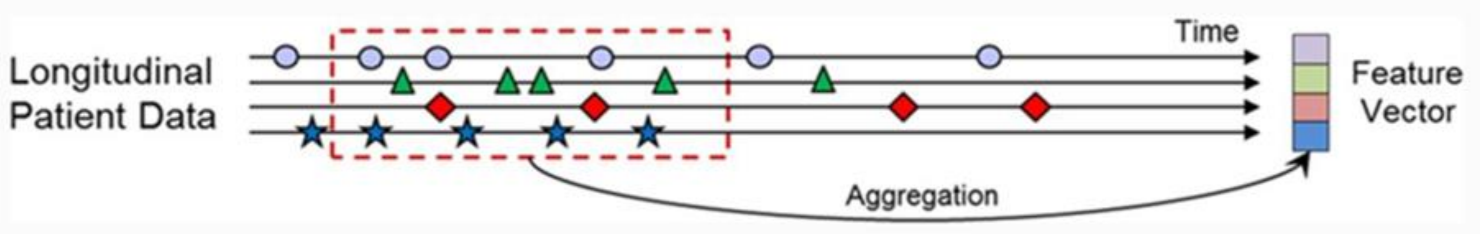
Table II: Soarian Data

LABEL	TYPE	DESCRIPTION
FND_PROCESSED	float	value of different measures
FND_CODE	string	name of the measure
FLAG_SIA	binary	intensive care flag
FLAG_URG	binary	emergency flag
LOS	float	length of the stay [day]
LABEL	binary	bedsore flag
FND_LIBELLE	string	definition of the FND_CODE
NUMERO_SEJOUR_CODE	string	unique id of each stay
IPP_CODE	string	unique id of each patient
AGE_A_LA_ADMISSION_CODE	integer	age of the patient at the moment of the admission
SOARIAN_DISPLAY_DATE_CODE	timestamp	timestamp of the moment the record has been entered

Case Study 1 : Bedsore – Pipeline

- **Input preprocessing :** with timeseries representation and sparsity handling

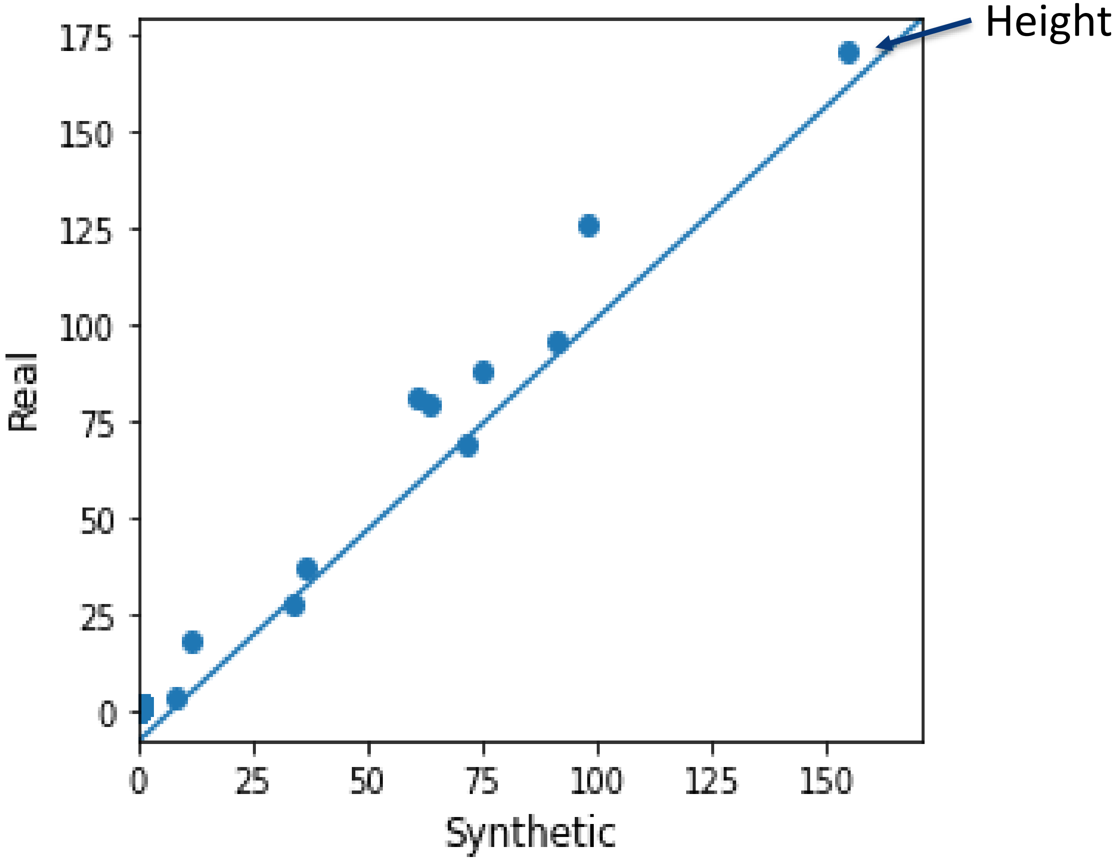
	FND_CODE	FND_PROCESSED	TIMESTAMP	AGE	LOS	SIA	URG	LABEL
(IPP_CODE, SEJOUR_CODE)								



	C1_mean	C1_std	C2_mean	C2_std	...	LOS
(IPP_CODE, SEJOUR_CODE TIME)						

Case Study 1 : Bedsore - Preliminary Results

Mean values : Plot the mean value of each finding code (temperature, height, weight, …)

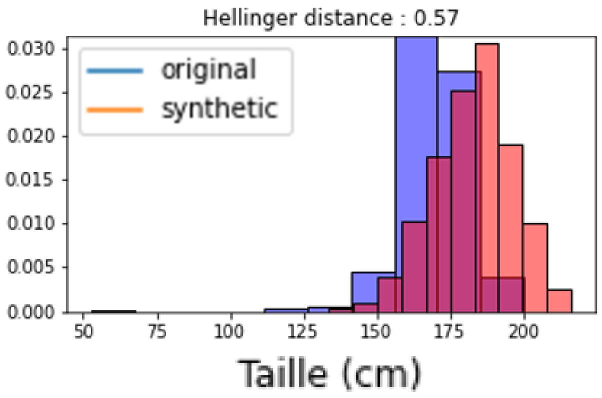


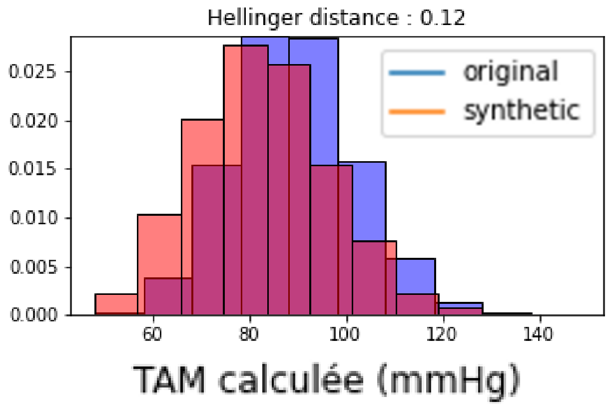
Synthetic vs Real

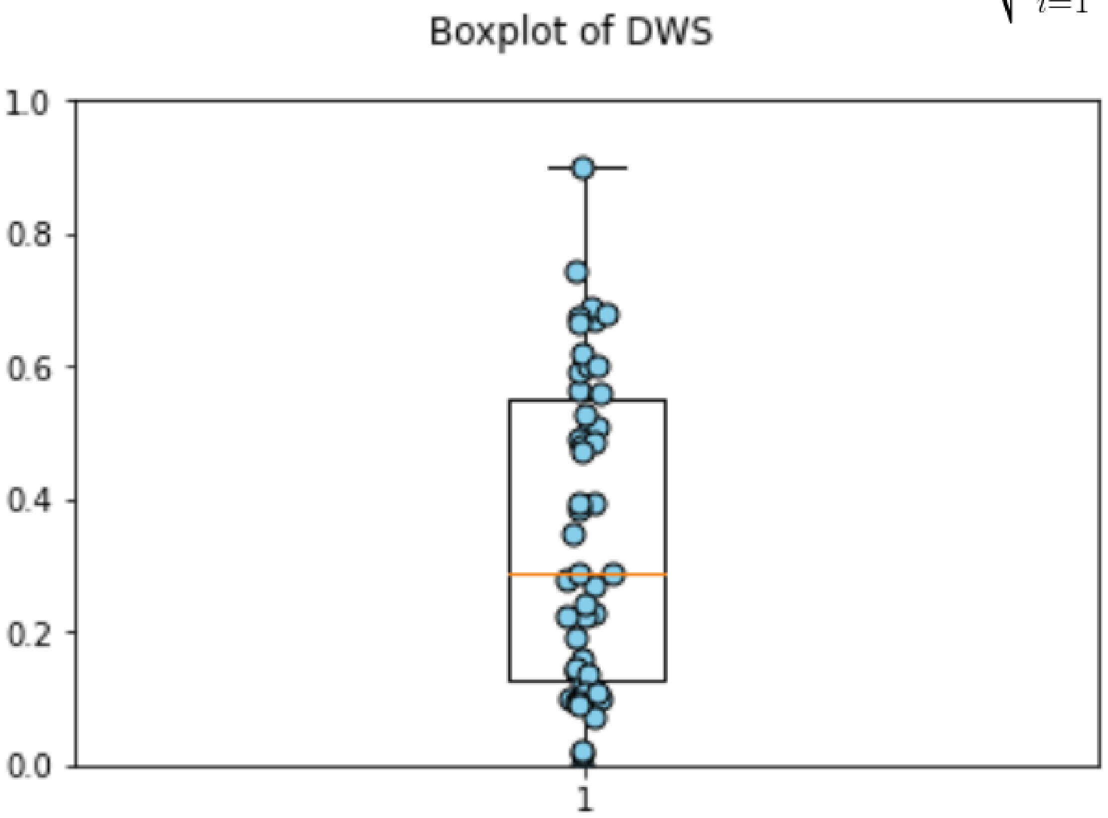
Case Study 1 : Bedsore - Preliminary Results

Dimension-wise statistics : Compare the distributions of each finding code using Hellinger distance

$$H(P,Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{k} (\sqrt{p_i} - \sqrt{q_i})^2}$$

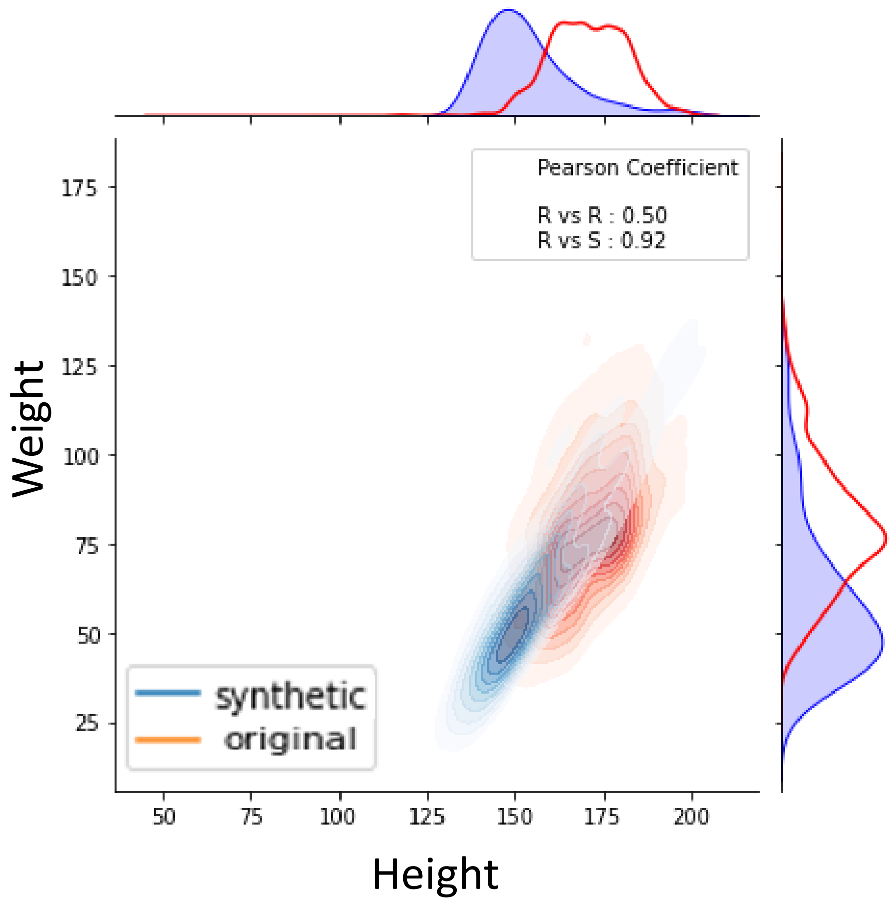






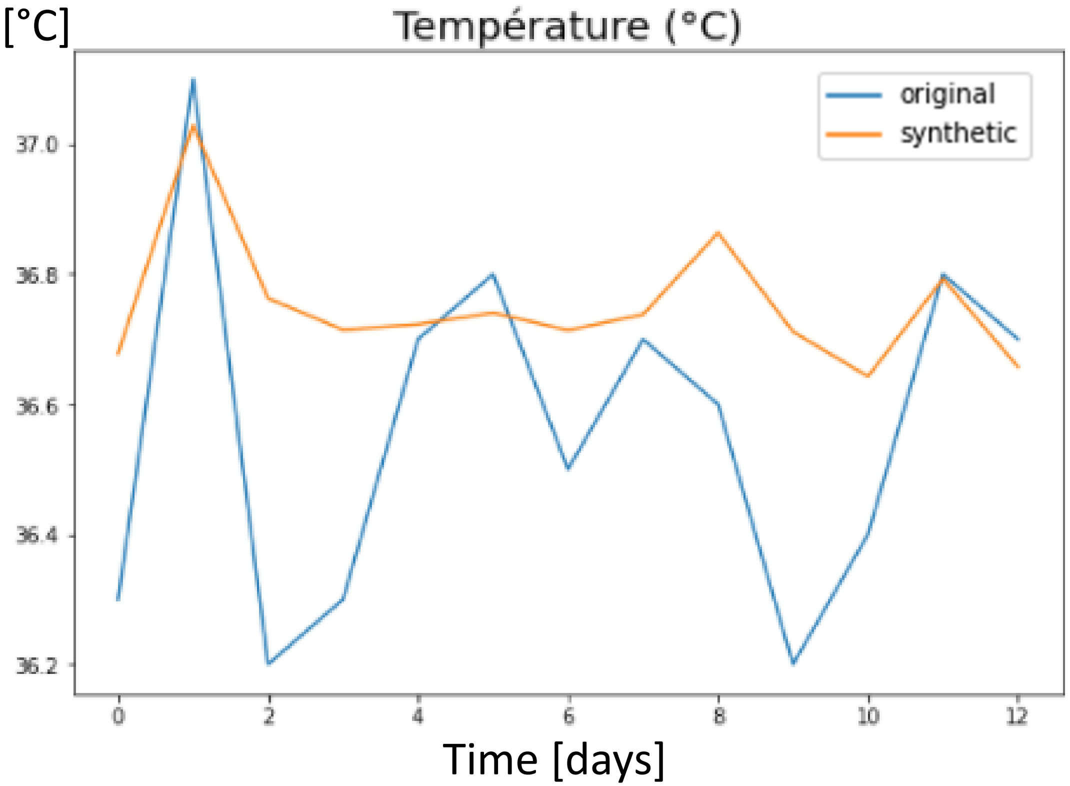
Case Study 1 : Bedsore - Preliminary Results

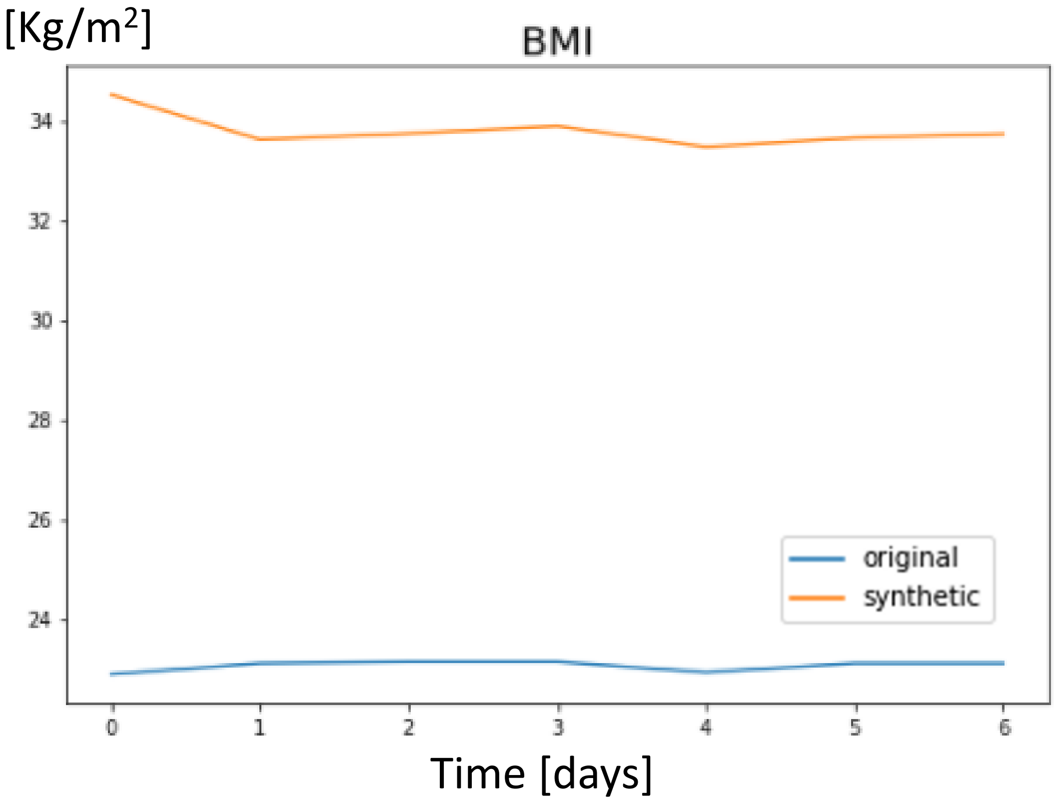
Joint distributions : Real in orange / Synthetic in blue



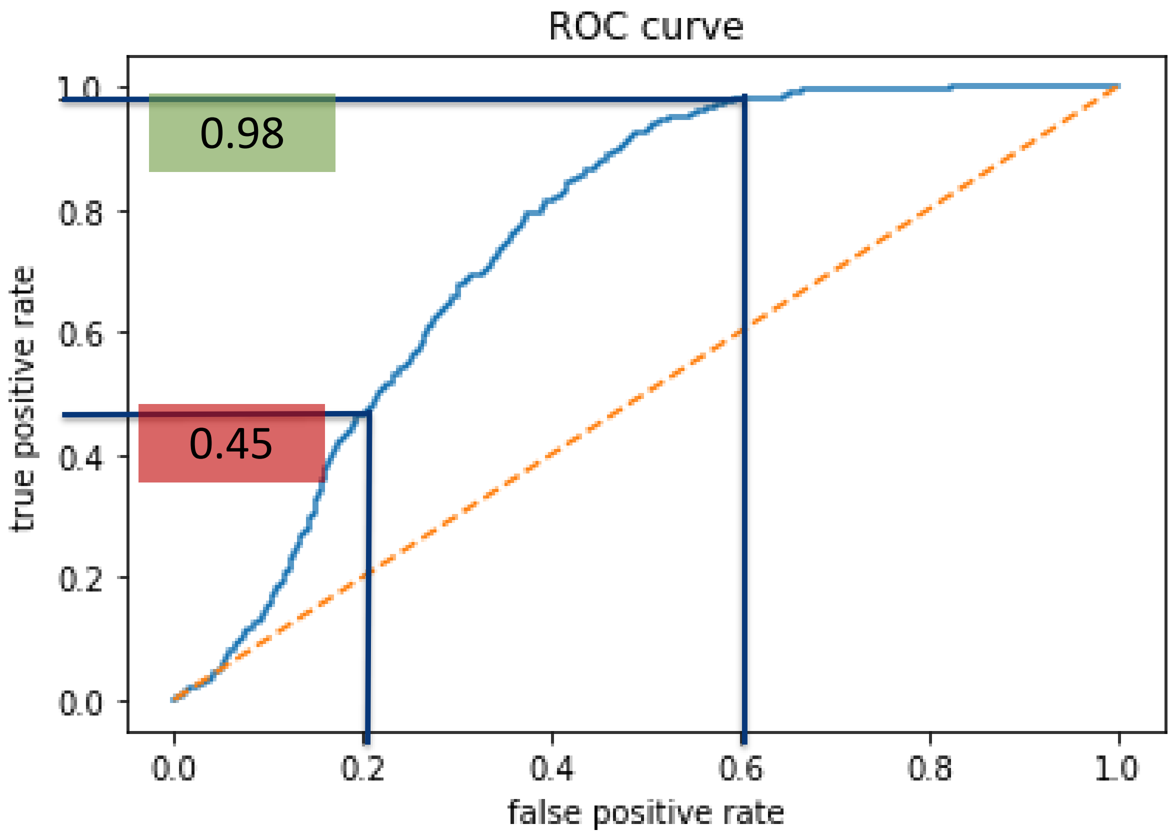
Case Study 1 : Bedsore - Preliminary Results

Temporal Evolution:

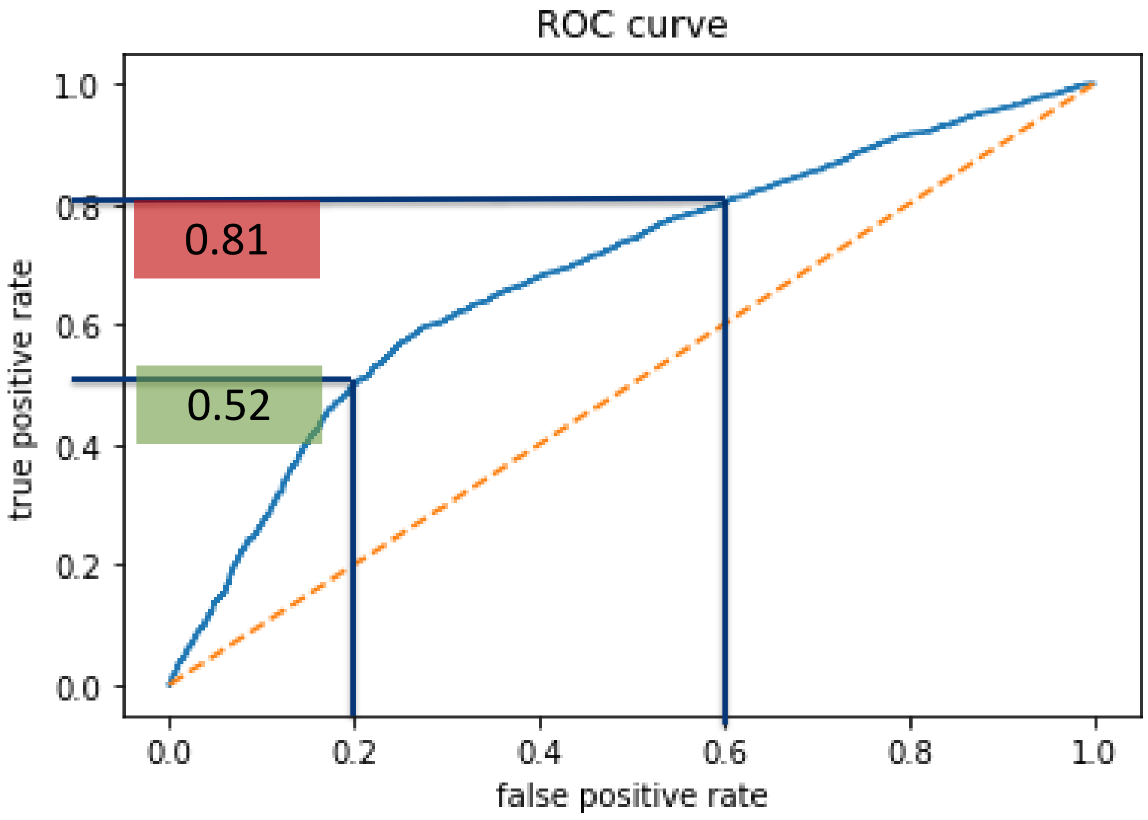




Case Study 1 : Bedsore - Results



Real



Synthetic

Case Study 1 : Bedsore – Conclusion

- Promising similarity metrics result : the SynTEG model captures the univariate distributions and the joint distributions. Time evolution of the attributes seems to be realistic.

- Bedsore prediction model: with the synthetic data, the prediction quality, in terms of AUROC, is not as good as with real data but a non-trivial model can still be constructed.

Take-home message

- **No plug and play model** :
    - On real clinical data, heavy input preprocessing is needed

- **Considerable contribution to model adaptation**:
    - SynTEG Model transformation to transition from discrete to continuous data

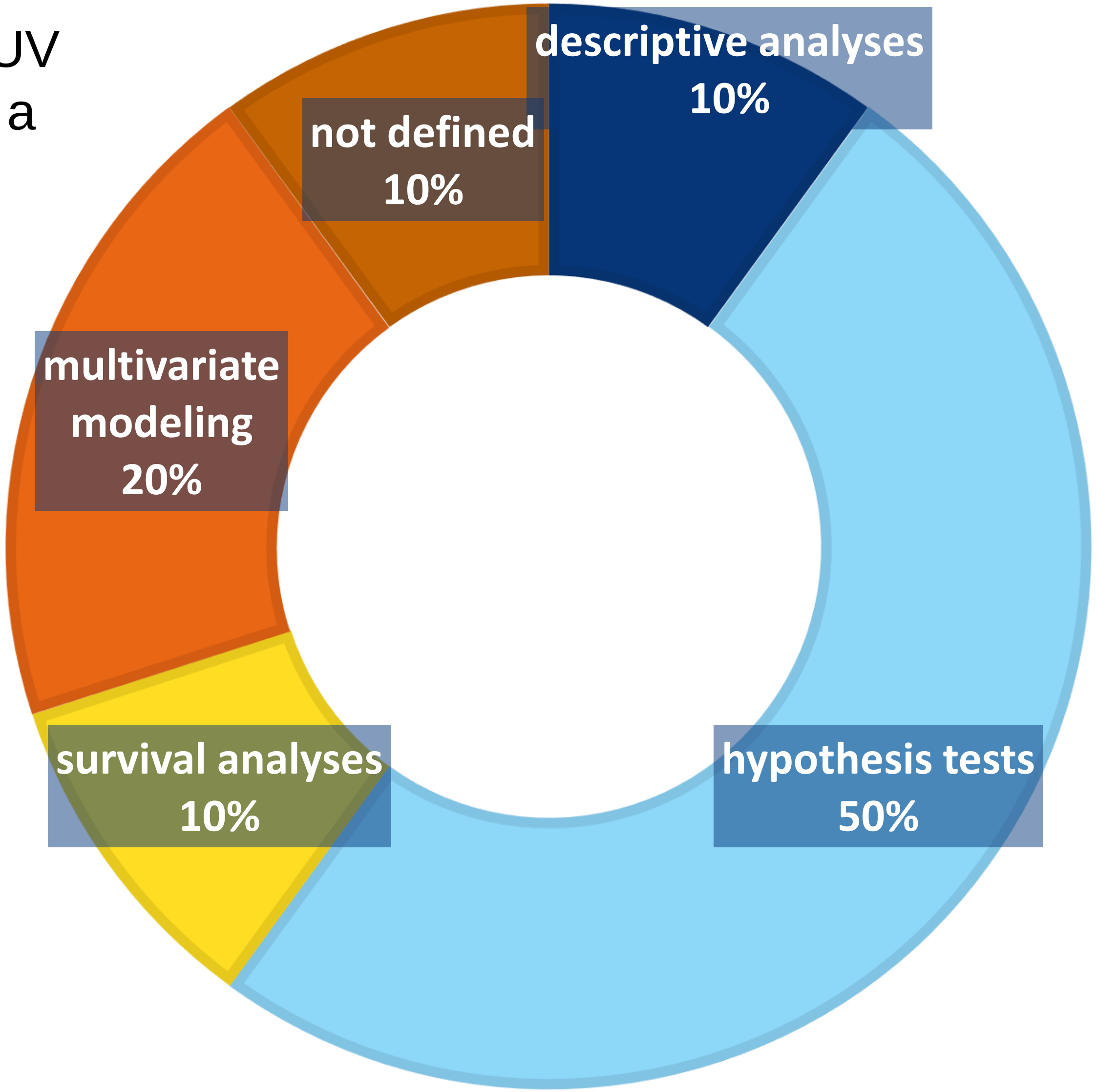
- **Preliminary results**:
    - Initial results are promising on real data but require a further, thorough analysis

Thank you for your attention

Supplementary Material

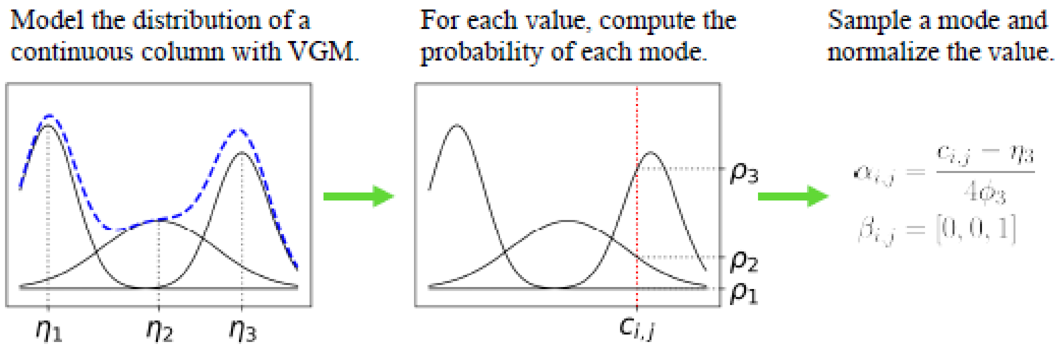
Research at CHUV

Breakdown of 150 CHUV research projects over a 7 month period:

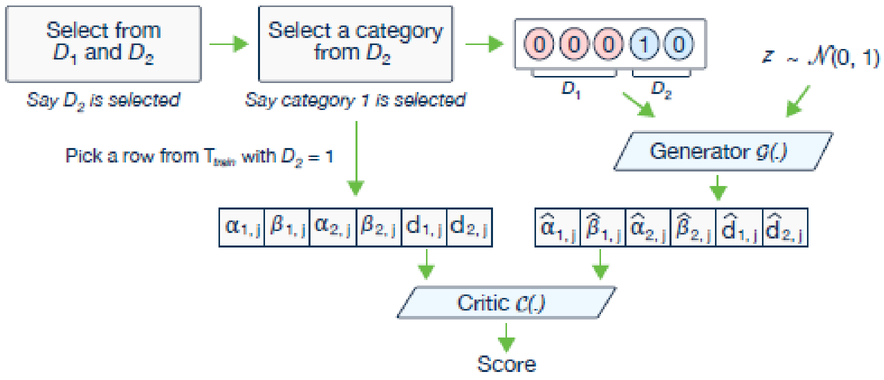


CTGAN

- **Heterogeneous Tabular Data**
  - Continuous : non-Gaussian / multiple modes
  - Discrete : imbalance
- **Mode-specific Normalization**

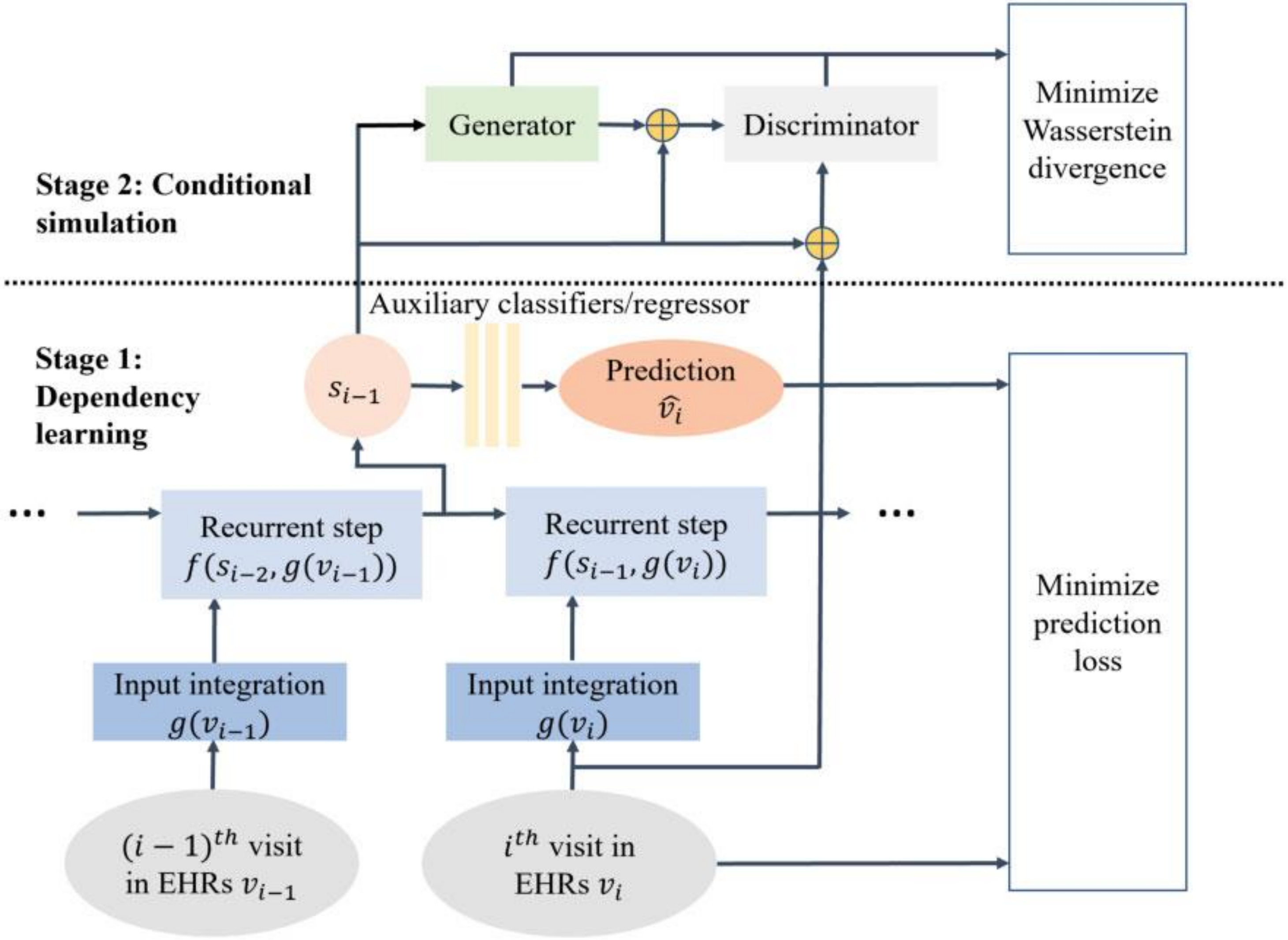


- **Conditional Generator and Training-by-Sampling**



Modeling Tabular data using Conditional GAN, Xu et al., 2019

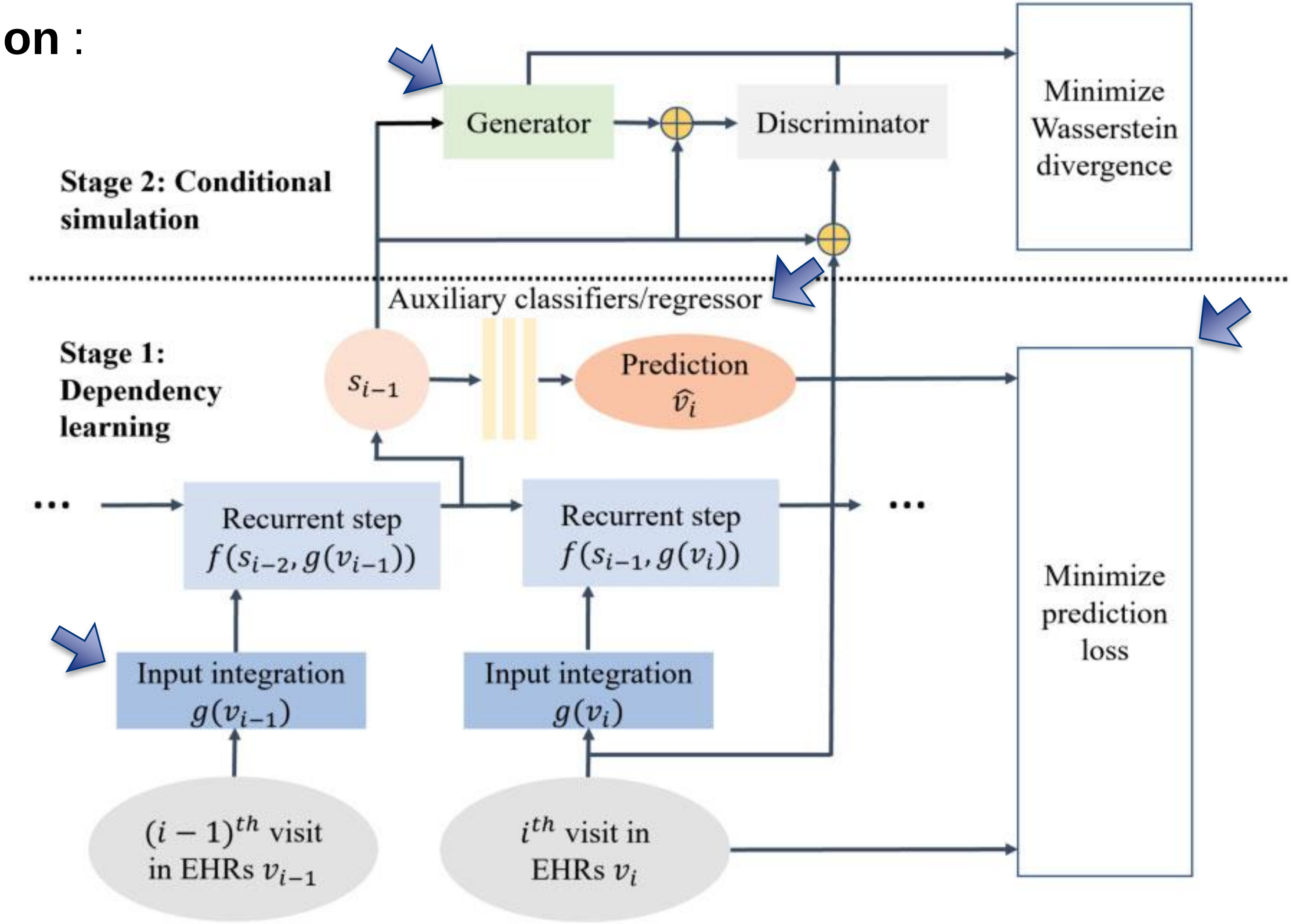
SynTEG



SynTEG a framework for temporal structured electronic health data simulation, Zhang et al., 2020

Case Study 1 : Bedsore – Pipeline

- **SynTEG Model Adaptation** :
    - Input integration
    - Auxiliary classifier
      & Prediction loss
    - GAN



Case Study 2 : Visceral Surgery : Pipeline

52 attributes : (10 + 42) attributes

Table XI: Visceral Surgery Data

LABEL	TYPE	DESCRIPTION
MNPPID	integer	Patient's Identifier
DATE_INTERVENTION_CODE	timestamp	timestamp of the intervention
DATE_SUIVI_OCC_CODE	timestamp	timestamp of the patient's visit
DATE_SUVI_KM_CODE	tiemstamp	same as DATE_SUIVI_OCC_CODE
AMPUTATION	binary	1 for amputation / 0 otherwise
TYPE_AMPUTATION	integer	2 for major / 1 for minor / 0 for no amputation
DECES	binary	1 for patient's death / 0 otherwise
REINTERVENTION	binary	1 for reintervention / 0 otherwise
JOURS_DELAIS	integer	Number of days from the first intervention (days)
TYPE_INTERVENTION	binary	1 for open / 0 for endovascular

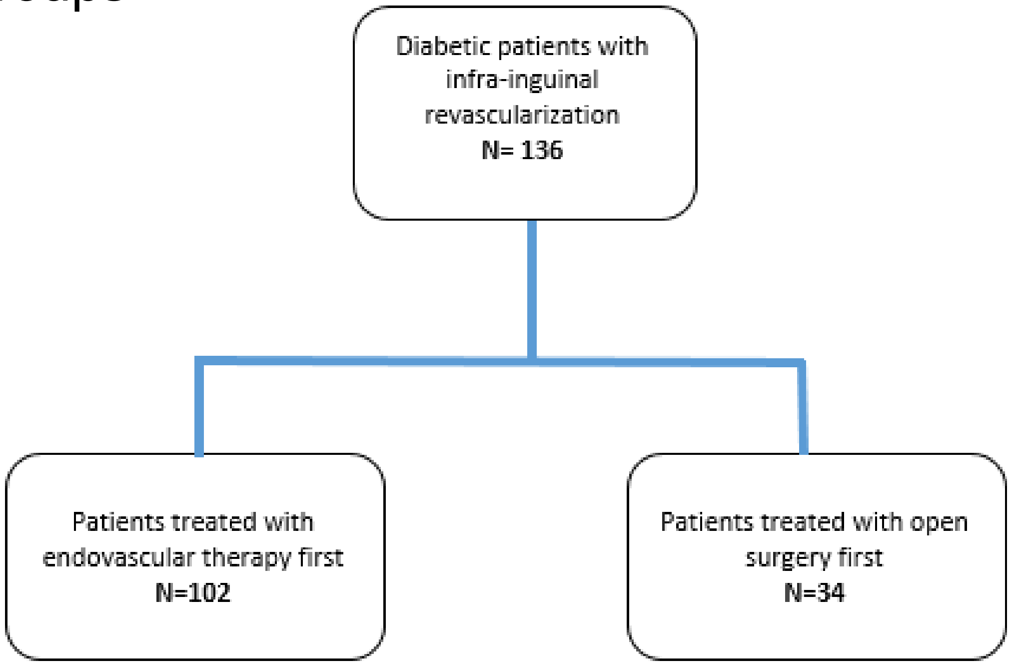
Table XII: Visceral Surgery Demographic Data

LABEL	TYPE	DESCRIPTION
MNPPID	integer	Patient's Identifier
DATEOP_CODE	timestamp	timestamp of the operation
DSEJOUR	integer	length of stay
POSTOP_CODE	timestamp	timestamp of the post-operation
SUIVI_CODE	timestamp	timestamp of the patient's visit
AGE	integer	age of the patient
SEXE	string	'F' or 'M'
POIDS	integer	weight of the patient
TAILLE	integer	height of the patient
BMI	float	BMI of the patient
DIABETE	binary	1 if the patient is diabetic in this diabetic cohort, always 1
DIABETE_INSULINO_REQUERANT	binary	1 if the patient requires insulin
HYPERCHOLESTEROLEMIE	binary	1 if the patient has hypercholesterolemia
HYPERTENSION	binary	1 if the patient has hypertension
ATCENT_PATHOL_CARDIAQUE	binary	1 if antecedent in cardiac pathology
ATCD_CEREBROVASC	binary	1 if antecedent in cerebrovascular disease
INSUFF_RENALE	binary	1 if renal insufficiency
ATCD_FAMILIAUX_VASC_NUM	integer	1 / 0 / -1 for family antecedent of vascular
ATCD_FAMILIAUX_VASC_LIB	string	Label of ATCD_FAMILIAUX_VASC_NUM 'Oui' for 1 / 'Non' for 0 / 'Inconnu' for -1
ATCD_AMPUT2	binary	1 if antecedent in amputation / 0 otherwise
ATCD_AMPUT	string	Label of ATCD_AMPUT2 'Oui' for 1 / 'Non' for 0
TABAC_NUM	integer	2 / 1 / 0 for tobacco use
TABAC_LIB	string	2 and 1 for 'Tabagisme Actuel and Ancien' 0 for 'Pas de tabagisme'
ASA_SCORE	integer	ASA score
MED_ANTICOAGULANTS	binary	use of anticoagulant
MED_ANTIAGREGANTS	binary	use of antigregant
MED_STATINES	binary	use of statin
MED_ANTIHYPERTENS	binary	use of antihypertensive
LOCAL_COTE_OP_NUM	binary	operation side 0 for left / 1 for right
LOCAL_COTE_OP_LIB	string	'G' for 0 / 'D' for 1
STADE_FONTAINE_G_VAL	integer	left fontaine stage
STADE_FONTAINE_G_LIB	string	label of STADE_FONTAINE_G_VAL
STADE_FONTAINE_D_VAL	integer	right fontaine stage
STADE_FONTAINE_D_LIB	string	label of STADE_FONTAINE_D_VAL
STADE_FONTAINE_GLOBAL	integer	global fontaine stage
INTERVENTION	timestamp	timestamp of the operation
INTER_CLASSIFICATION_NUM	binary	1 if 'Urgent' / 0 if 'Electif'
INTER_CLASSIFICATION_LIB	string	label of INTER_CLASSIFICATION_NUM
CHIR	binary	1 if chirurgical intervention
ENDOVASC	binary	1 if endovascular intervention
DUREE_INTERVENTION	integer	length of intervention
INTERV_TYPE	integer	type of intervention

No model adaptation for CTGAN

Case Study 2 : Visceral Surgery

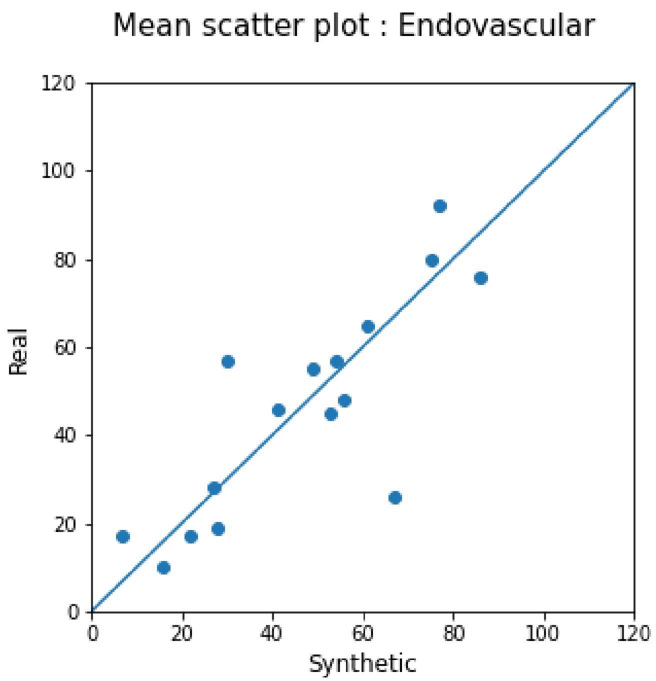
Data : 136 patients in 2 groups

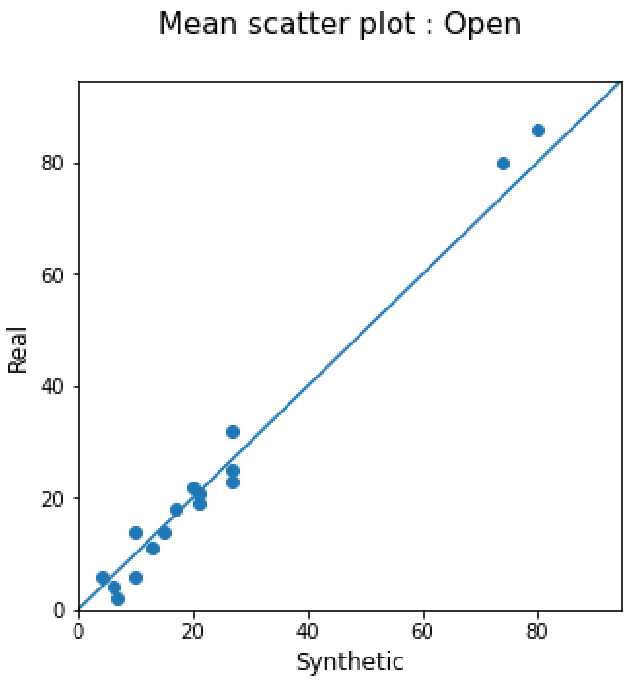


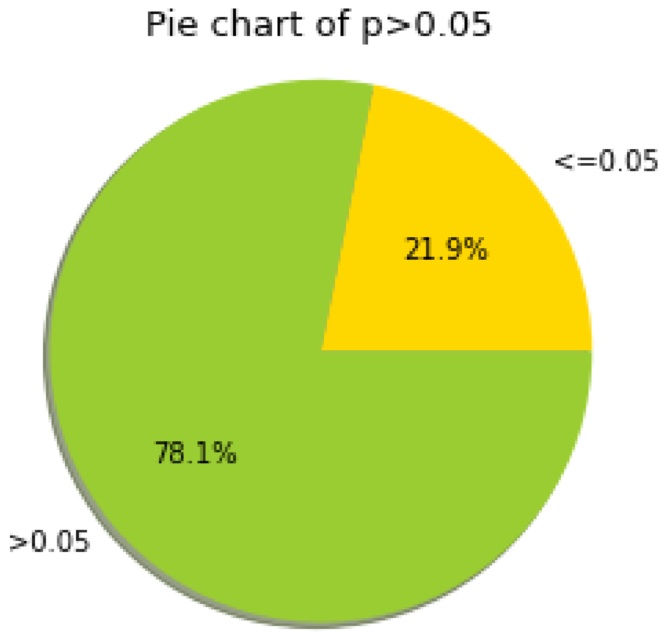
Study : Survival estimation against amputation / death / reintervention using Kaplan-Meier

Case Study 2 : Visceral Surgery - Results

Contingency Table







Case Study 2 : Visceral Surgery – Conclusion

- Similar mean values : The mean values of the real and synthetic datasets are relatively similar. For ~3/4 of the attributes, such synthetic mean values are **not** unlikely to be observed given the real mean values.

Case Study 3 : Pharmacokinetics

Pharmacokinetics Data

Cohort : 405 neonates with vancomycin monitoring
23 attributes

Study : The optimal vancomycin dosing for neonates
& Pharmacokinetic model of vancomycin

Input preprocessing :

SynTEG Model Adaptation : (cf. Case Study 1)

Table XIII: Pharmacokinetics Data

LABEL	TYPE	DESCRIPTION
ID	integer	Patient's Identifier
AMT	float	Administered drug amount (mg)
RATE	float	infusion rate
DV	binary	Dependent variable Concentration measurements in (mg/l)
EVID	binary	Event Identifier 1 for drug intake / 0 for plasma sampling
MDV	binary	Missing Dependent Variable 0 for concentration measurements / 1 otherwise
WT	integer	Body weight at drug administration (g)
BWT	integer	Body weight at birth (g)
GA	float	Gestational Age (weeks)
CA	float	Chronological Age (weeks)
PMA	float	Postmenstrual Age (weeks)
SEX	binary	0 for male / 1 for female
CRT	float	Serum creatinine (mmol/L)
URE	float	Urea (mmol/L)
ALB	float	Albumin (g/L)
SGA	binary	Small for gestational age: 1 for yes / 0 for no
PNA	integer	Postnatal age (days)
DOSE	float	Drug amount (mg)
DOSEKG	float	Drug amount / body weight (mg/kg)
TNDiamm	float	Size at birth (cm)
PCDiamm	float	Head circumference (cm)
DV_N	float	Normalized DV (mg/L)
DATETIME_CODE	timestamp	Datetime of drug intake or plasma sampling

Case Study 3 : Pharmacokinetics – Preliminary Results