AVATAR

Accelerating Innovation through data sharing data with no privacy concerns

Project: CONCEPT/0722/0019

D6, Synthetic data generator

Date: 29/09/2023

Classification of Dissemination: Confidential

Contents

1. Introduction

1.1 Scope of the document

The scope of this document is to define the workflow for generating synthetic data, to present the available real patient data and to implement the synthetic generator model using the collected data.
In particular, it describes the input data parameters needed for the successful implementation of the model and it creates the workflow (including the algorithm flowchart) for generating synthetic data and validating the results. The modules and functions of the generator model are also described along with the input arguments (and their requirements) and the files needed for the execution of the model. Finally, the report outlines common evaluation metrics that can be used for the evaluation procedure.

2. Definition of the workflow for generating synthetic data

2.1 Approach

A sequential procedure will be followed in this work for the development of the synthetic generator model. Initially, real data from patients with prostate cancer will be collected. The real patient data will be then cleansed to ensure data validity and sanity (i.e., high quality data). The cleansed data will be then fed into a synthetic data generator model. To this end, a synthetic data generator model will be developed. The filtered data will be used to train the model so that it learns the structure and the information contained. Once trained, the generator will be able to generate new synthetic data. Different models will be developed using machine learning principles. Ensemble learning will be then utilised for pooling the results of the multiple implemented models and averaging them using weights based on accuracy, to minimize modelling errors and bias. Once the best performing generator is derived, the quality of its generated synthetic data will be assessed using common metrics to define the extent to which the statistical properties of the real data are captured to the synthetic data sets and how much of the real data may be revealed (directly or indirectly) by the synthetic data.
Figure 1 summarises the workflow for generating the synthetic data.

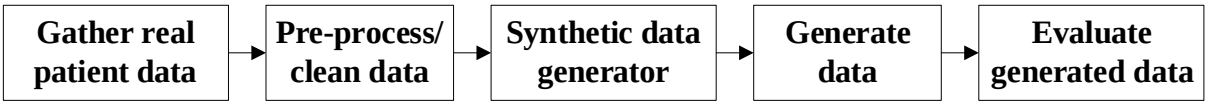


Figure 1. Flowchart of the proposed methodology.

3. Collection of patient data and available data set

3.1 Data collection

Records/Data from 1222 patients with prostate cancer were collected at 11 different centres [1]. These data will be used for the implementation of the synthetic data generator. The data were gathered by the German Oncology Center (GOC – PA1). To address patient confidentiality, the records were anonymised, and policies and confidentiality agreements were signed between the project partners.

3.2 Background information and description of collected data

The patients under study received salvage radiotherapy (sRT). The treatment decision process of sRT is directly affected by the prostate-specific membrane antigen positron-emission tomography (PSMA-PET), which is increasingly used for staging patients with biochemical relapse or prostate-specific antigen (PSA) persistence after radical prostatectomy.
The collected data include several variables such as the age of the patient, the disease and treatment (PSMA PET-guided salvage radiotherapy, sRT, for prostate cancer patients with biochemical relapse after prostatectomy), clinical data (such as PSA pre sRT, Gleason score, pathological [pT] stage, resection (R) status, local relapse in PET, nodal relapse in PET, time gap between surgery and relapse, PSA persistence), treatment variables (such as dose to the prostatic fossa, dose to elective pelvic lymphatics, dose to PET-positive pelvic lymph nodes [PLN], duration of androgen deprivation therapy) and the outcome variable - biochemical relapse (as time to event endpoint: biochemical-recurrence free survival).
Table 1 summarises the available input variables and provides a short explanation/description of the input parameters along with the collected values and coding.

Table 1: Description of the input variables - collected data.

ID	Input Variable	Units	Subgroups meaning	Code for subgroups / Collected values
1	age at sRT	years		36-81
2	initial PSA (code)	ng/ml		
3	pT stage at surgery (code)		2 3a 3b 4 unknown	1 2 3 4 empty
4	pT stage at surgery (binarized I)			0 1 empty
5	pT stage at surgery (binarized II)			0 1 empty
6	pN stage at surgery (code)		0 1 unknown	0 1 empty
7	R stage at surgery (code)		0 1 2 not sure unknown	0 1 2 3 empty
8	R stage at surgery (binarized)			0 1 empty
9	ISUP score in surgery specimen (code)		1+2 3 4 5 unknown	1 2 3 4 empty
10	ISUP score in surgery specimen (binarized)			0 1 empty
11	PSA persistence after surgery (defined as PSA ≥ 0.1 ng/ml) (code)	ng/ml	no yes unknown	0 1 empty
12	Time gap between surgery and recurrent disease (code)	years	0-1 >1 unknown	1 2 empty

13	PSA doubling time (code)	months	0-6 6.1-12 >12 unknown	1 2 3 empty
14	PSA before PET (code)	ng/ml	0.01-0.2 0.21-0.5 0.51-1 >1 Unknown	1 2 3 4 empty
15	PSMA/PET tracer (code)		68Ga-PSMA-11 68Ga-PSMA-I&T 18F-PSMA-1007 18F-PSMA-DCFPyL 18F-rhPSMA-7 18F-rhPSMA-7.3 Other Unknown	1 2 3 4 5 6 7 empty
16	PET findings (code)	yes/no	No findings in PET any findings in PET Unknown	0 1 empty
17	Local failure - miTr			
18	Nodal failure – miN			
19	PSA before sRT (code)	ng/ml	0.01-0.2 0.21-0.5 0.51-1 >1 Unknown	1 2 3 4 empty
20	PSA before sRT (binarized)			0 1
21	Time between PET and beginning of sRT (code)	months	<3 months 3-6 months >6 months unknown	0 1 2 empty
22	All PET positive lesions located in the RT field (code)	yes/no	no yes unknown	0 1 empty
23	sRT to fossa (code)	yes/no	no yes unknown	0 1 empty
24	sRT dose to fossa or boost dose to PET positive lesion in the fossa (code)	Gy	<66 Gy 66-70 Gy >70 Gy Unknown	1 2 3 empty
25	sRT dose to fossa (binarized)			0 1 empty
26	sRT to pelvic LN (code)	yes/no	no yes unknown	0 1 empty
27	sRT dose to pelvic LN (code)	Gy	Unknown <50 Gy 50-60 Gy >60 Gy Unknown	0 1 2 3 empty
28	sRT to elective pelvic lymphatics (code)		no whole-pelvis half-pelvis other unknown	0 1 2 3 empty
29	Dose to elective pelvic lymphatics (code)	Gy	≤50 Gy >50 Gy	1 2

			Unknown	empty
30	ADT (code)	yes/no	no yes unknown	0 1 empty
31	Duration of ADT (code)	months	<6 months 6-12 months >12-24 months > 24 months Unknown	0 1 2 3 empty
32	Time gap between end of ADT and last follow (code)	months		0 1 2 3 11 12 17 22 48 empty
33	PSA relapse (code)	yes/no	no yes unknown	0 1 empty
34	Time to relapse	months		1-120 empty
35	Relapse under ongoing ADT (code)		no yes unknown	1 0 empty
36	Death of PCa (code)	yes/no	no yes unknown	0 1 empty
37	Time to death of PCa	months		6-82 empty
38	Time to last FU	months		2-86
39	Death at last FU	yes/no	no yes	0 1
40	Include FFBF			1 empty
41	Metastases	yes/no	no yes unknown	0 1 empty
42	Time to metastases	months		2-85 empty
43	PSMA re-staging			0 1 empty
44	Include DM (all imaging)			1 empty
45	Include DM (PSMA only)			1 empty

* Abbreviations: androgen deprivation therapy (ADT), follow up (FU), Freedom from biochemical failure (FFBF), International Society of Urological Pathology (ISUP), lymph node (LN), positron-emission tomography (PET), prostate cancer (PCa), prostate-specific antigen (PSA), prostate-specific membrane antigen positron-emission tomography (PSMA PET), salvage radiotherapy (sRT).

Finally, screenshots from the collected patient data and variables are shown in Figure 2.

Var1 Age Age at sRT (y) (y)	Var2 iPSA initial PSA in ng/ml (code)	Var3 pT pT stage at surgery (code)	Var4 pN pN stage at surgery (code)	Var5 R R stage (code)	Var6 ISUP ISUP score in surgery specimen (code)	Var7 PSAP PSA persistence (code)
69	2	1	0	0	1	0
71	1	1	1	0	2	0
64	1	1	0	0	1	0
71	2	1	0	1	1	1
73	2	1	0	1	2	0
70	1	2	1	0	2	0
61	2	1	0	1	3	0
68	2	1	1	0	1	0
72	3	3	0	1	1	0
55	1	1	0	3	2	0
83	2	2	0	0	2	1
78	1	1	0	1	1	0
62	1	1	0	0	1	1
70	1	1	0	0	1	0
78	2	1	0	1	3	0
76	1	3		0	2	1
57	1	1	0	0	2	0

Var13 LocarRec Local failure - miTr (code)	Var14 NodalRec Nodal failure - miN (code)	Var15 PSA sRT PSA before sRT (code)	V16 time PET sRT Time between PET and beginning of sRT (code)	V17 Lesions RT field All PET positive lesions in RT field (code)	V18 sRT Fossa sRT to fossa (code)	V19 dose Fossa RT dose to fossa (code)	End PSA rel PSA relapse (code)	End Time to relapse Time to relapse (in months)
0	0	1.00	1	1	1	2	0	69
0	0	1.00	1	1	1	2	1	44
1	0	3.00	2	1	1	1	0	77
0	1	4.00	2	1	1	2	1	41
0	0	1.00	1	1	1	2	0	44
0	1	1.00	1	1	1	2	1	26
0	0	2.00	2	1	1	2	0	66
0	0	2.00	2	1	1	2	1	41
1	0	4.00	2	1	1	2	0	50
0	0	2.00	2	1	1	2	0	57
1	0	4.00	2	1	1	3	1	13
1	0	3.00	2	1	1	2	0	47
0	0	3.00	2	1	1	2	0	49
1	0	1.00	1	1	1	2	0	47
0	0	3.00	2	1	1	2	0	46
1	0	4.00	2	1	1	2	1	25
0	0	2.00	2	1	1	2	0	50

Figure 2. Screenshots depicting the available data (provided by the GOC – PA1) and some of the recorded variables.

4. Cleansing of patient data and final data set

4.1 Data cleansing process

The available data will be first examined for missing values and then the cleaned (i.e., the final) data set will be created. For the data cleansing process, an analysis will be performed to test which of the variables have significant impact on the outcome variable. Only these input variables will be included in the final data set. It is also worth noting here that for missing values in the "significant" variables, the entire patient record will be excluded from the final data set. Finally, data statistics (e.g., min and max values, median, average, etc.) will be derived for each variable in the final data set. The same data set will be then used for the development of synthetic data generator model (MS6).

4.2 Input parameters selection, final data set and initial statistics

The statistical analysis for testing which of the available input variables have significant impact on the outcome variable has already been performed and reported by recent studies in the field [1]–[4]. Following their recommendations and the obtained results, the input variables of pT stage, R status, PSA serum values before sRT, ADT use, dose to the prostatic fossa, persistence of PSA after surgery and PET related variables (i.e., PLN or local recurrences prior to sRT) were selected as the significant ones. Other variables, such as the biochemical disease-free interval between surgery and sRT, the presence of PLN in the surgical specimen, PSA doubling time, and preoperative PSA, were not selected due to limited predictive values in previous studies [1]–[4].

To enable unbiased analysis and proper development of the generator, patients with missing clinical data for the significant variables (the ones with IDs 3-5, 7-11, 17-20, 22-25 and 30)

were excluded. Overall, 192 patients were excluded (more details are given in Table 2), thus the remaining 1030 patients were included in the final data set.

Table 2: Data cleansing results for the data set under study.

	Records
Total number of patients (all raw data)	1222
Removed due to missing medical records	183
Exclusion of Fossa from the sRT field	4 (55 in total with the incomplete records)
All PET lesions in RT field - PSMA-PET avid lesions not covered by the sRT field	5 (6 in total with the incomplete records)
Total Excluded	192
Remaining	1030

The statistics for the final data set (including the baseline patient and treatment characteristics) are summarised in Table 3. The analysis revealed that 632 patients (61.36%) had PSA serum values before sRT of 0.5 ng/mL or less, 438 patients (42.52%) had PET scan-detected locally recurrent disease, and 317 patients (30.727%) had at least 1 PLN-PET. ADT was prescribed for 325 of 1030 patients (31.55%). None of the patients in this study received an escalation of systemic therapy beyond ADT.
In addition, the most frequently applied equivalent dose in 2 Gy per fraction (EQD2, $\alpha/\beta = 1.6$ Gy) to the prostatic fossa or to local recurrent disease was 66 to 70 Gy (551 of 1030 [53.49%]). On results of PSMA-PET scan prior to sRT, 438 of 1030 patients (42.52%) had local recurrences, while 313 of 1030 patients (30.38%) had nodal recurrences. Salvage radiotherapy to elective pelvic lymphatics was delivered to 395 of 1030 patients (38.34%). All PLN-PETs received dose-escalated sRT; the most frequent dose (129 of 264 [48.86%]) was 50 to 60 Gy (EQD2, $\alpha/\beta = 1.6$ Gy). Finally, 338 patients (32.81%) had biochemical relapse after a median follow-up time (the median time to relapse was 26 months).

Table 3: Data statistics for the final data set (including patient's and treatment characteristics).

Characteristic	**Patients**					
Age at sRT, years	**Median**	**IQR**	**Average**	**Min**	**Max**	**Total**
	70	64-74	69	42	89	
pT stage						
2						461
3a						327
3b						235
4						7
Resection status in surgery						
R0						674
R1						327
R2						3
Rx						26
ISUP grade in surgery						
1+2						372
3						324
4						156
5						178
PSA persistence after surgery						
No						751
Yes						279
PSA before sRT, ng/mL						
0.01-0.2						246
0.2-0.5						386

0.5-1						172
>1						226
Local recurrence after PSMA-PET						
No						592
Yes						438
Pelvic lymph nodes after PSMA-PET						
No						713
Yes						317
Dose to the prostatic fossa, Gy						
<66						103
66-70						551
>70						376
sRT to elective pelvic lymphatics						
No						634
Yes						395
Dose to elective pelvic lymphatics, Gya						
≤50						267
>50						47
Unknown						716
Irradiation to positive pelvic LNs *						
No						713
Yes						317
Dose to positive pelvic LNs, Gy						
≤50						21
50-60						129
>60						114
Unknown						766
ADT						
No						705
Yes						325
Duration of ADT admission, months						
<6						65
6-12						110
12-24						57
>24						49
Unknown						749

* PLN defined by PSMA-PET imaging or based on localization of pN+ status in surgery.

5. Implementation of the synthetic data generator model

5.1 Literature review

A literature survey revealed that the healthcare synthetic data are generated by utilising process-driven or data-driven methods [5]. The first-class of synthetic data are derived from computational or mathematical models of an underlying physical process. Process-driven methods include numerical and Monte Carlo simulations, agent-based modelling, and discrete-event simulations. On the other hand, data-driven methods operate on observed data to derive the synthetic data using generative models. In this work, focus is shed on data-driven models.
As indicated by Goncalves et al. [5], there are 3 main types of data-driven methods: imputation based methods, full joint probability distribution methods and function approximation methods. Imputation based methods are fully probabilistic and include multiple imputation techniques in the context of Statistical Disclosure Control (SDC) and Statistical Disclosure Limitation (SDL) methodologies. In this domain, generalized linear regression models and non-linear methods (e.g., Random Forest and neural networks) have been utilised. Full joint probability distribution methods include statistical algorithms (e.g., parametric and non-parametric Bayesian networks)

for generating fully synthetic data by estimating, learning or approximating a joint probability distribution.

With the current advances in the field of artificial intelligence (AI) and the development of computational power, the generation of synthetic data from machine learning (ML) techniques has attracted a lot of attention lately. State-of-the-art ML and deep learning algorithms have seen an unprecedented growth in popularity, and they have been incorporated into probabilistic models, creating a new generation of models (i.e., the Deep Generative Models, DGMs) that exploit deep learning for creating synthetic data. Such models are based on function approximation methods using the conventional train and test set approach. Regarding DGMs, Generative Adversarial Networks (GANs), Markov Chain and Variational AutoEncoders (VAE) models were used to generate new data instances with or without an explicit formulation of the data probability density function. GANs have shown remarkable results in many fields [6] and are considered as "the preferable method" to generate synthetic data due to the construction of robust models with less labelled data. In the health sector, the GANs method has been used successfully for unsupervised learning tasks and for generating health data. A general limitation of GANs is the inability for generating categorical synthetic data sets. To alleviate the GANs drawback, improved generative GANs methods (or extended GANs-based models such as the medGAN, HealthGAN, etc.) have been lately introduced by researchers [7]. Such methods are based on the Wasserstein GAN and use a novel variant of the categorical encoding method to handle mixed categorical and discrete data.

From a literature search, open-source software packages also exist for synthetic data generation (e.g., the R packages synthpop [8] and SimPop [9], the Python package DataSynthesizer [10], the Python library Synthetic Data Vault (SDV) [11] and the Java-based simulator Synthea [12]). Though, such generators were not deemed suitable to simulate healthcare synthetic data as part of the research work to be undertaken in this proposal. In particular, the R package synthpop for generating synthetic does not include appropriate procedures for synthesizing multiple event data and the choice for replacing only selected cases from selected variables is not available (currently all values of variables chosen for synthesis are replaced). Likewise, the R package simPop implements model-based methods to simulate synthetic populations and it is based on household survey data and auxiliary information. The Python package DataSynthesizer increases the computational burden (compared to the other methods) due to the constructed differentially private Bayesian network and enforced consistency among the noisy marginals. Similarly, the Synthea simulator is based on publicly available summary statistic data, and therefore it does not provide the flexibility of creating faithfully generative models that resemble real data. Finally, one of the most popular open-source synthetic data generators is the SDV library, that builds a ML model using synthetic data. The SDV uses a variety of ML algorithms to learn patterns from the real data and emulate them in synthetic data. Though, it was recently shown that SDV's performance was measured at 40% accuracy, highlighting a significant disparity in the results [13].

Apart from the open-source software packages, dedicated websites (such as MOSTLY AI [14]) for generating synthetic data were recently developed. Such websites combine the most recent advances in generative AI with a thorough grasp of data protection and compliance. Currently, MOSTLY AI is claimed to be one of the best AI-powered synthetic data generator, using different types of data sets [13].

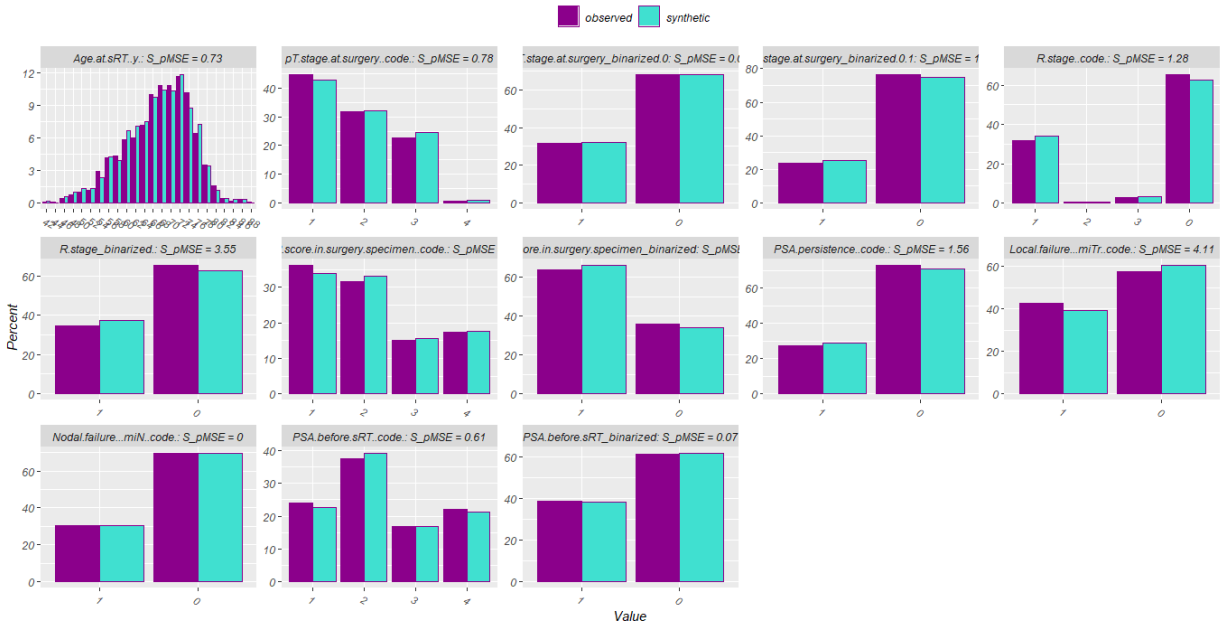
To sum up, different techniques and tools have been proposed for data synthesis and all of them have their merits/advantages and drawbacks. The selection of the model is case dependent, and

the quality of the synthetic generated data is highly dependent on the quality of the model (and hence data) that created it. There is always a trade-off between the features, the computational burden, and the data availability, as well as a trade-off between data utility and privacy.
Since cancer patient data are available for the AVATAR project, the team members decided to proceed with the development of an AI-powered generator based on DGMs for high-quality (in terms of data utility and information disclosure) synthetic health data.

5.2 Synthetic data generator models

5.2.1 Implementation of existing solutions for synthetic data generation

Initially, the open-source software package synthpop [8] was used for generating synthetic data. The results are summarized in Figure 3. To measure utility, the propensity mean squared error (pMSE) metric, that predicts whether the synthetic data can be distinguished from the original, was used. The pMSE ranged from 0.04 to 4.11.




```
Selected utility measures:

                                          pMSE   S_pMSE df
Age.at.sRT..y.                        0.000178 0.732680  4
pT.stage.at.surgery..code.            0.000141 0.775500  3
pT.stage.at.surgery_binarized.0       0.000002 0.040227  1
pT.stage.at.surgery_binarized.0.1     0.000104 1.706754  1
R.stage..code.                        0.000234 1.283839  3
R.stage_binarized.                    0.000215 3.545114  1
ISUP.score.in.surgery.specimen..code. 0.000145 0.796632  3
ISUP.score.in.surgery.specimen_binarized 0.000137 2.257548  1
PSA.persistence..code.                0.000095 1.561658  1
Local.failure...miTr..code.           0.000249 4.110751  1
Nodal.failure...miN..code.            0.000000 0.004585  1
PSA.before.sRT..code.                 0.000111 0.608298  3
PSA.before.sRT_binarized              0.000004 0.065641  1
```

Figure 3. Screenshots of the results generated by synthpop package.

Apart from the open-source R package, the MOSTLY AI (available at https://mostly.ai/) solution was tested. The results are summarized in Figure 4. The trained model by MOSTLY AI showed high accuracy when representing the statistics of the real patient data. Identical matches between the synthetic and real samples were found for 4.27%.

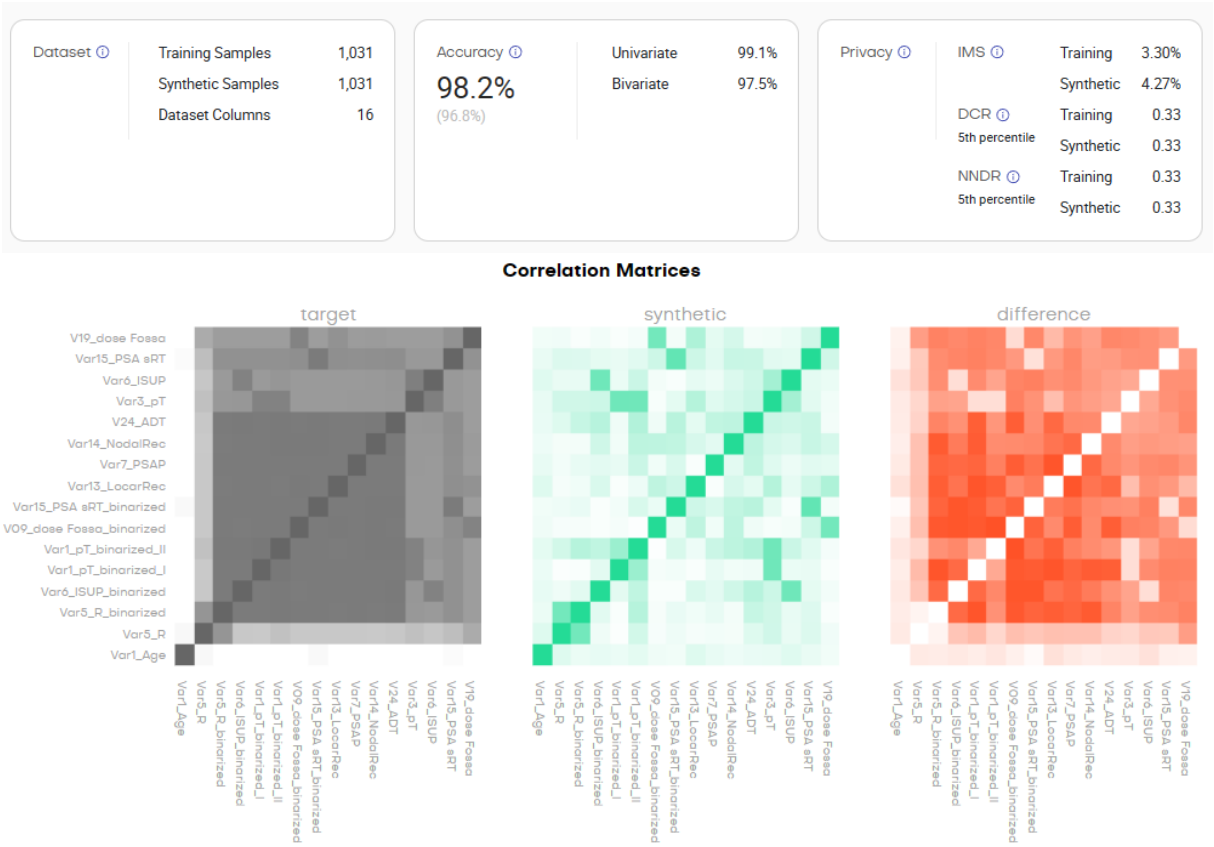


Figure 4. Screenshots of the results generated by MOSTLY AI [14].

5.2.2 Proposed synthetic data generator model

An innovative AI-powered generation system that leverages machine learning principles (e.g., DGMs) and ensemble modelling was developed to create privacy preserving synthetic heath data. AI-driven generative models were selected for the implementation of the generator since they can capture complex functions, they can compute non-linear input–output mappings and their effectiveness depend on the provided data. Given the available data (that are balanced and of high quality – ensured by the data pre-processing step) from PA1, more robust ML models can be built. The available data were used for the training procedure of the ML models. The final synthetic generator model (derived from ensemble modelling) was used to generate the new data (i.e., the synthetic data). Comparison/analysis between the real and synthetic generated data was then performed to evaluate the performance of the generator. The innovative character of this generator lies in the usage of the ensemble modelling technique for pooling the results of multiple models (in this work, different DGMs with different number of nodes and hidden layers will be implemented) and averaging them using weights based on accuracy, in order to minimize modelling errors and bias. After the successful implementation of the model (MS6), synthetic data files will be generated.

With the current advances in the field of AI, generation of synthetic data from ML techniques has attracted a lot of attention lately. State-of-the-art ML algorithms include Bayesian networks (BNs) and neural networks using data augmentation methods. In this domain, Generative Adversarial Networks (GANs) have become particularly popular as a method to generate synthetic data and to build robust models with less biased data. In the health sector, GANs method has been used successfully for unsupervised learning tasks and to generate health data.

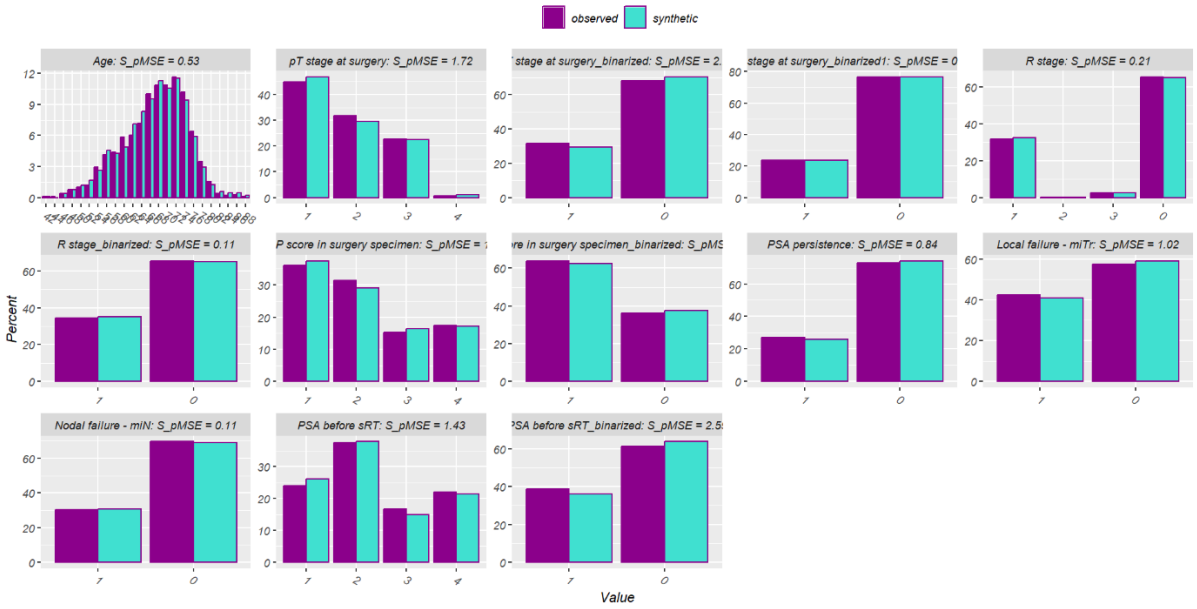
A general limitation of GANs is the inability for generating categorical synthetic data sets. To alleviate the GANs drawback, recent works introduced improved generative GANs based methods (e.g., medGAN, HealthGAN, etc.). Such methods can handle mixed categorical and discrete data and have already been used to generate synthetic data related to cancer. The HealthGAN method was developed as an open-source Python package available in GitHub (GitHub - TheRensselaerIDEA/synthetic_data: Repository for the UHF synthetic data project). From a literature search, other open-source software packages do exist for synthetic data generation (e.g., the R packages synthpop [8] and SimPop [9], the Python package DataSynthesizer [10], and the Java-based simulator Synthea [12]). Though, such generators (e.g., Synthea [5]) are based on publicly available summary statistic data, and therefore they do not provide the flexibility of creating generative models faithfully resembling real data.
In this proposal, improved generative GANs will be utilised for creating privacy preserving synthetic health data. Initially, the real-world data are gathered, pre-processed, and then used to train the generator model inside the secure environment. Then the model (no need for real data to generate the synthetic ones) is exported outside the secure environment. Finally, data is generated using the model, which can be then used for different applications.
Advantages of this methodology include the effectiveness to capture resemblance, privacy, utility and footprint using novel and existing metrics. In addition, this method generates high quality synthetic data by maintaining the relationships that exist in real patient data. It deviates from the classical methods by focusing on methods that create new data points that approximately mimic the real data rather than altering the real data points.

5.3 Development of the data generator model and results

5.3.1 Ensemble data generator

Initially, an ensemble approach was utilised to create synthetic data by combining two different methods: k-means clustering and resampling (bootstrap).
For the k-means clustering method, the procedure is as follows:
The generator first applies k-means clustering to the numeric columns of the original data set. It clusters similar data points into groups (clusters) based on their attributes. Then, the numeric columns are isolated from the data set, and k-means clustering is performed on these columns to identify clusters. The centroid (mean) of each cluster is used to create synthetic data by representing the characteristics of each cluster.
For the resampling (bootstrap) method, the procedure is as follows:
The generator uses resampling, specifically the bootstrap method, to create synthetic data. Random samples are drawn with replacement from the original data set. This resampling technique generates synthetic data by replicating observations from the original data set.
Combining the two methods for synthetic data generation:
The generator then combines the synthetic data sets created from the k-means clustering and resampling methods. In particular, the synthetic data sets generated by k-means clustering (based on cluster centroids) and resampling are merged to form an ensemble or combined synthetic data set.
Output:
The final output of the generator is an ensemble synthetic data set that incorporates characteristics from both the k-means clustering-based centroids and the resampled data. The aim of this process is to create a diverse synthetic data set that captures various aspects of the

original data through clustering-based representation and resampling, providing a more comprehensive synthetic representation for further analysis or modelling purposes.
The results obtained from the ensemble synthetic data set are summarized in Figure 5, with pMSE ranging from 0.005 to 2.59. It can be observed that the ensemble technique outperformed the generator implemented by the open-source software package synthpop.

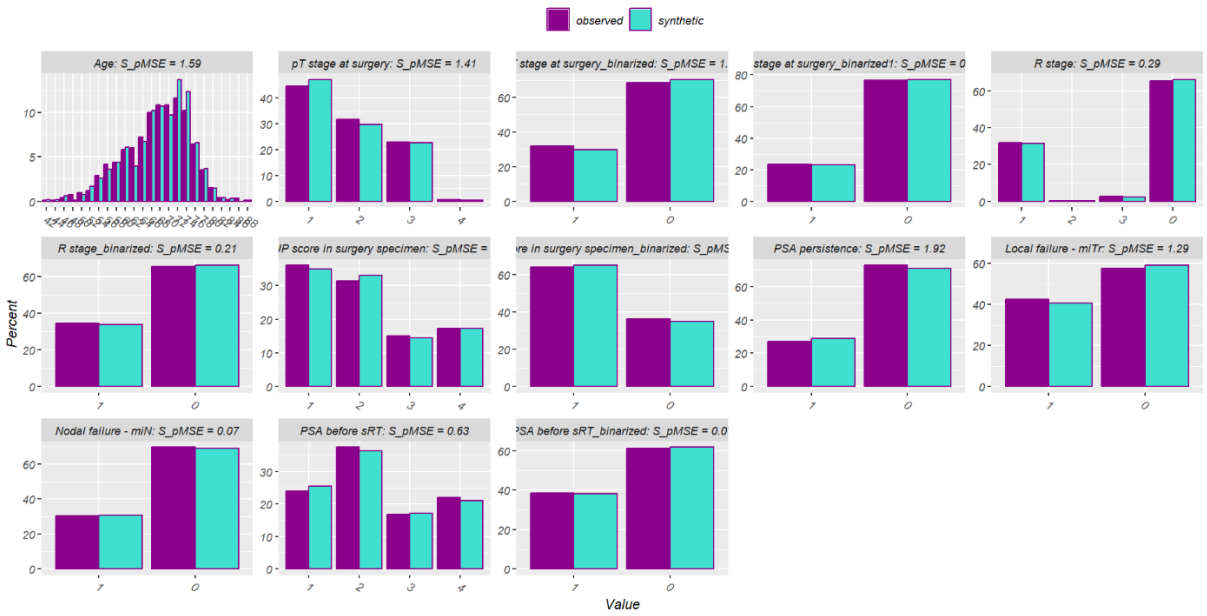



```
Selected utility measures:
                                          pMSE   S_pMSE df
Age                                   0.000128 0.526093  4
pT stage at surgery                   0.000314 1.722402  3
pT stage at surgery_binarized         0.000134 2.209517  1
pT stage at surgery_binarized1        0.000000 0.005394  1
R stage                               0.000038 0.208235  3
R stage_binarized                     0.000006 0.106965  1
ISUP score in surgery specimen        0.000207 1.139365  3
ISUP score in surgery specimen_binarized 0.000050 0.818226  1
PSA persistence                       0.000051 0.843286  1
Local failure - miTr                  0.000062 1.022016  1
Nodal failure - miN                   0.000007 0.114229  1
PSA before sRT                        0.000260 1.429823  3
PSA before sRT_binarized              0.000157 2.591015  1
```

Figure 5. Screenshots of the results generated by an ensemble generator.

5.3.2 AI-driven generator with ensemble modelling

To further improve the performance of the generator, an innovative AI-powered generation system that leverages machine learning principles (i.e., DGMs) and ensemble modelling was developed to create privacy preserving synthetic heath data. The innovative character of this generator lies in the usage of the ensemble modelling technique for pooling the results of multiple models (different DGMs with different number of nodes and hidden layers were implemented) and averaging them using weights based on accuracy, to minimize modelling errors and bias.
The results obtained from the AI-powered generator are summarized in Figure 6, with pMSE ranging from 0.06 to 1.92.



Selected utility measures:

	pMSE	S_pMSE	df
Age	0.000387	1.594595	4
pT stage at surgery	0.000257	1.411748	3
pT stage at surgery_binarized	0.000111	1.822839	1
pT stage at surgery_binarized1	0.000008	0.136010	1
R stage	0.000053	0.289107	3
R stage_binarized	0.000013	0.211332	1
ISUP score in surgery specimen	0.000091	0.501171	3
ISUP score in surgery specimen_binarized	0.000037	0.610310	1
PSA persistence	0.000117	1.923894	1
Local failure - miTr	0.000079	1.294346	1
Nodal failure - miN	0.000004	0.073171	1
PSA before sRT	0.000115	0.631925	3
PSA before sRT_binarized	0.000004	0.065641	1

Figure 6. Screenshots of the results generated by the AI-powered generator.

6. Performance evaluation metrics

The generated synthetic data will be first tested for privacy, resemblance, and utility. During the evaluation, different metrics (e.g., pairwise correlation difference, log-cluster, support coverage, nearest neighbours' adversarial accuracy and cross-classification) will be used to define the extent to which the statistical properties of the real data are captured to the synthetic data set and how much of the real data may be revealed (directly or indirectly) by the synthetic data. To this end, different utility assessment methods for synthetic data and privacy metrics will be used.

6.1 Common metrics used in the literature

Metrics commonly used in the literature include the Membership Inference Attack, ground truth tables, receiver operating characteristic (ROC) curve, Hellinger distance, correlations, statistical distances (such as the total variation distance and exact matches between synthetic and original data), precision, accuracy and recall [5], [15]. Data statistics (such as the variable distribution, mean and median), correlation differences and the Cox-regression analysis can also be used for the evaluation process. Other performance evaluation metrics include the nomogram and the k-fold cross-validation [1].

A detailed literature review and a final list of the evaluation metrics (that will be used in this project) will be derived in Task 4.1 (Deliverable 6 of WP4). The final list will include only the privacy and utility metrics applicable for this specific project application and the ones that best capture the quality of the artificially generated data.

7. Conclusions

This document provided the detailed methodology for generating synthetic data and the procedure to adapt the methodology for cancer data. It also described the collected data and created final data set. Finally, a detailed description of the data generator model was also provided along with common evaluation metrics.

8. Acknowledgement

Co-funded by the European Union

RESEARCH & INNOVATION FOUNDATION

Republic of Cyprus

EUNDO we make it work

German Oncology Center

The project is implemented under the programme of social cohesion "THALIA 2021-2027" co-funded by the European Union, through Research and Innovation Foundation.

References

[1] C. Zamboglou *et al.*, "Development and Validation of a Multi-institutional Nomogram of Outcomes for PSMA-PET–Based Salvage Radiotherapy for Recurrent Prostate Cancer," *JAMA Netw. Open*, vol. 6, no. 5, p. e2314748, 2023, doi: 10.1001/jamanetworkopen.2023.14748.

[2] Zamboglou, C *et al.*, "Metastasis-free survival and patterns of distant metastatic disease after prostate-specific membrane antigen positron emission tomography (PSMA-PET)–guided salvage radiation therapy in recurrent or persistent prostate cancer after prostatectomy.," *Int J Radiat Oncol Biol Phys*, vol. 113, no. 5, 2022, doi: https://doi.org/10.1016/j.ijrobp.2022.04.048.

[3] D. Bartkowiak, A. Siegmann, D. Böhmer, V. Budach, and T. Wiegel, "The impact of prostate-specific antigen persistence after radical prostatectomy on the efficacy of salvage radiotherapy in patients with primary N0 prostate cancer," *BJU Int*, vol. 124, no. 5, pp. 785–791, 2019, doi: https://doi.org/10.1111/bju.14851.

[4] R. D. Tendulkar *et al.*, "Contemporary update of a multi-institutional predictive nomogram for salvage radiotherapy after radical prostatectomy," *J. Clin. Oncol.*, vol. 34, no. 30, pp. 3648–3654, 2016, doi: 10.1200/JCO.2016.67.9647.

[5] A. Goncalves, P. Ray, B. Soper, J. Stevens, L. Coyle, and A. P. Sales, "Generation and evaluation of synthetic patient data," *BMC Med. Res. Methodol.*, vol. 20, no. 1, pp. 1–41, 2020, doi: 10.1186/s12874-020-00977-1.

[6] H. Jiri, C. Massimo, D. I. L. E. O. Margherita, D. E. N. Sarah, O. Nicole, and N. Nicholas, "Multipurpose synthetic population for policy applications," Joint Research Centre

(Seville site), 2022.

[7] R. D. Camino, C. A. Hammerschmidt, and R. State, "Generating Multi-Categorical Samples with Generative Adversarial Networks," *ICML 2018 Work. Theor. Found. Appl. Deep Gener. Model.*, pp. 1–7, 2018.

[8] B. Nowok, G. M. Raab, and C. Dibben, "Synthpop: Bespoke creation of synthetic data in R," *J. Stat. Softw.*, vol. 74, no. January 2017, 2016, doi: 10.18637/jss.v074.i11.

[9] M. Templ, B. Meindl, A. Kowarik, and O. Dupriez, "Simulation of synthetic complex data: The R package simPop," *J. Stat. Softw.*, vol. 79, no. 10, 2017, doi: 10.18637/jss.v079.i10.

[10] B. Howe, J. Stoyanovich, H. Ping, B. Herman, and M. Gee, "Synthetic Data for Social Good," in *Bloomberg Data for Good Exchange Conference*, 2020, vol. 2657, pp. 1–9, doi: 10.1145/nnnnnnn.nnnnnnn.

[11] "SDV: The synthetic Data Vault (SDV)." [Online]. Available: https://github.com/sdv-dev/SDV.

[12] J. Walonoski *et al.*, "Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record," *J. Am. Med. Informatics Assoc.*, vol. 25, no. 3, pp. 230–238, 2018, doi: 10.1093/jamia/ocx079.

[13] G. Loizou, "MOSTLY AI: The most accurate synthetic data generator," 2023. [Online]. Available: https://machinelearningmastery.com/mostly-ai-the-most-accurate-synthetic-data-generator/.

[14] "MOSTLY AI." [Online]. Available: https://mostly.ai/.

[15] K. El Emam, L. Mosquera, and R. Hoptroff, *Practical synthetic data generation: balancing privacy and the broad availability of data*. O'Reilly Media, 2020.