Data generation with AI

GANs, GDPR and the Norwegian cancer registry

Kosovare Olluri, Phd

kol@nextbridge.no

+47 47 02 26 07

with Silje Nord, Daniil Shantsev, Jarle Strand & Kjetil Kalager

amesto NextBridge

Outline

- The problem
- GDPR
- The Norwegian cancer registry
- GANs
- Results
- Metrics
- Survival Analysis

The problem

Data in our society

- SoMe
- Entertainment
- Information
- Government
- Healthcare

Misuse of data

- Data theft
- Lack of awareness
- Personal gain
- Breach of confidentiality agreements

amesto NextBridge

Some large scale examples

Reverse engineering of de-identified data

IN AUSTRALIA

The government released an "**anonymized**" data set comprising the medical billing records, including every prescription and surgery, of 2.9 million people.

Scientist used 6 days to reidentify people for the dataset

#1500 downloads

IN GERMANY

A journalist and a data scientist secured data from **three million users** by creating a fake marketing company

"a canny broker can find an individual in the noise, just from a long list of URLs and timestamps"

IN THE USA

The Massachusetts Group Insurance Commission released "**anonymised**" data showing the hospital visits of state employees. A data scientist Sweeney where able to reidentify the governor who promised that the patients privacy where protected.

In later work, Sweeney showed that 87% of the population of the United States could be uniquely identified by their date of birth, gender and five-digit zip codes.

GDPR

EU general Data Protection Regulation

CONSENT

Users must explicitly consent to each type of marketing message

DATA PROTECTION

Personal data must be stored and processed with data protection at its core

DELETION AND CORRECTION

Users can request to have their data deleted, corrected or restricted in a timely manner

amesto
NextBridge

THE CASE

- Established in 1951, one of the oldest national cancer registries in the world
- All medical doctors in the country are instructed by law to notify new cancer cases
- 200 employees, among them 40 researchers (medicine, statistics, informatics and psychology ++)
- Administrative responsibility for the public screening programs in Norway (Breast, Cervical and from 2021 starting pilot program for colerectal cancer screening)
- Collects data
- Produce statistics of the cancer prevalence in Norway
- Extensive research activity
- Current Privacy disclosure methods
    - de-identify data
    - Random forest
    - Decision tree
    - Linear regression

Cancer
Registry of Norway

THE PROBLEM

- Biological markers
- Known/unknown attributes
- Re-engineering

ESTABLISHED METHODS ARE NO LONGER GOOD ENOUGH/
BE TRUSTED TO BE GDPR COMPLIENT

ONE SOLUTION

GANs

GENERATIVE ADVERSARIAL NETWORKS

2014 Goodfellow et. al. «Generative advererial networks»

2017 Choi et. al. «Generating Multi-label Discrete Patient Records using Generative Adversarial Networks»

2018 Camini et. al. «Generating Multi-Categorical Samples with Generative Adversarial Networks»

2020 Concalves et. al. «Generation and evaluation of synthetic patient data»

amesto NextBridge

Discriminator – a classifier, standard supervised learning

Generator – random noise, usually a convolutional network generate image from noise

Discriminator gets alternately real and fake image

The gradient of the discriminator is used to train the generator, gradient descent, adjust weights

Generator is being moved up the gradient for the discriminator error

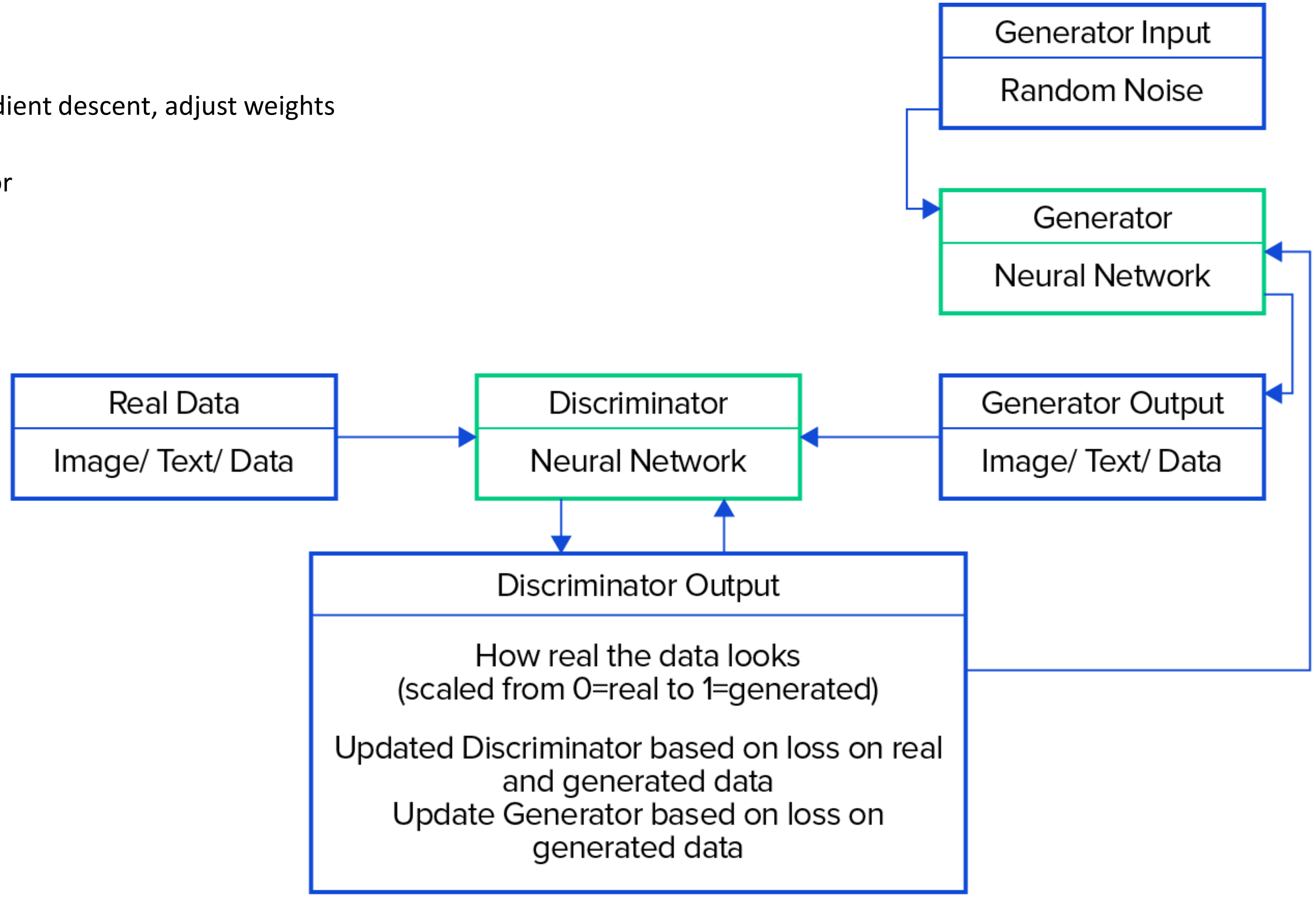
Tweak weights so that the discriminator is more wrong

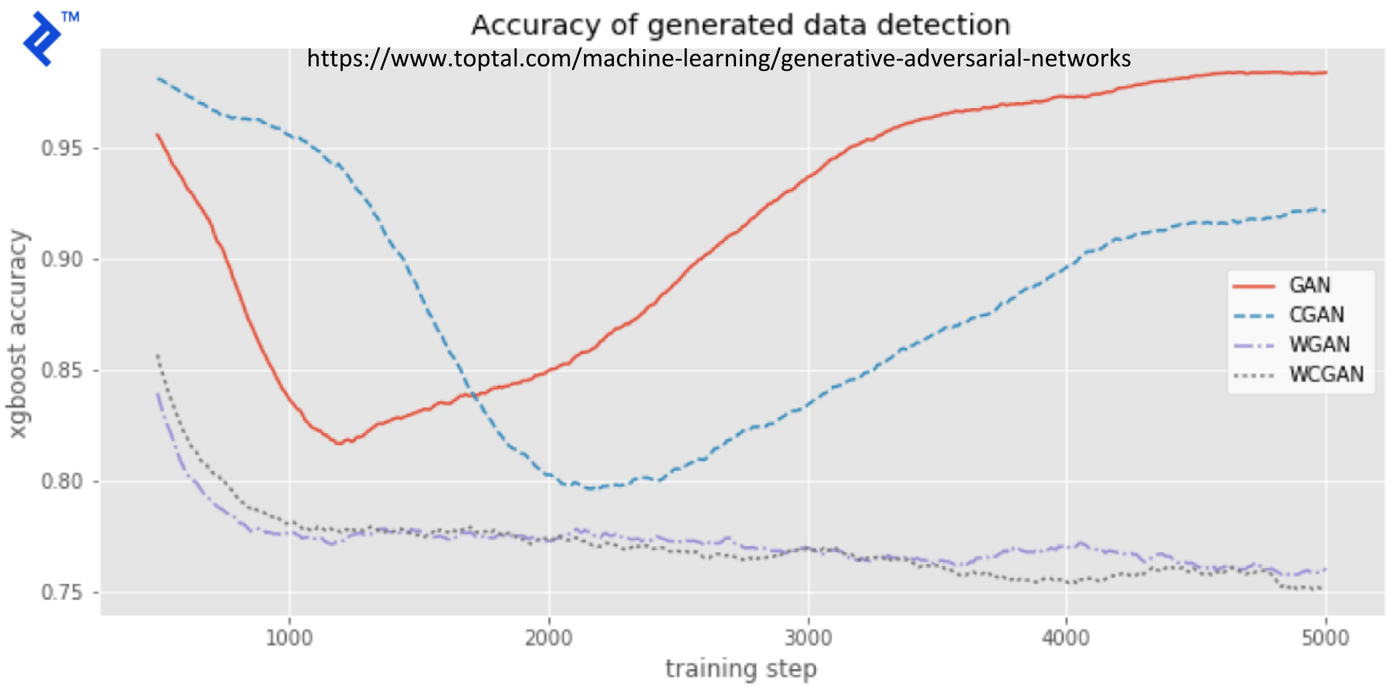
Generator output from random data randomly selected point in

Latent space

As the generator learns, the generator is making a mapping between the latent space and the desired results (cat images). As we move in latens space, the generator produce something that we consider real/meaningful about the object(cat).

The dimension of the latent space represent features of the original data, i.e. size, location in the image, color ++ ➔ the generator has structured its latent space in a way that it has some understanding of what the object is (cat) in general and in a meaningful way.





GAN challenges

Architecture and hyperparameter tuning of two networks

Generator/discriminator forget old tricks

Networks overpower each other

Mode collapse

Labeled data

Evaluations metrics for real/fake data: Cross-entropy loss vs Wasserstein distance

Xgboost fraud detection lossess
CGAN= Condissional GAN, WGAN= Wesserstein GAN

medGAN

From **one** continuous variable to **multiple** continuous **and** binary variables

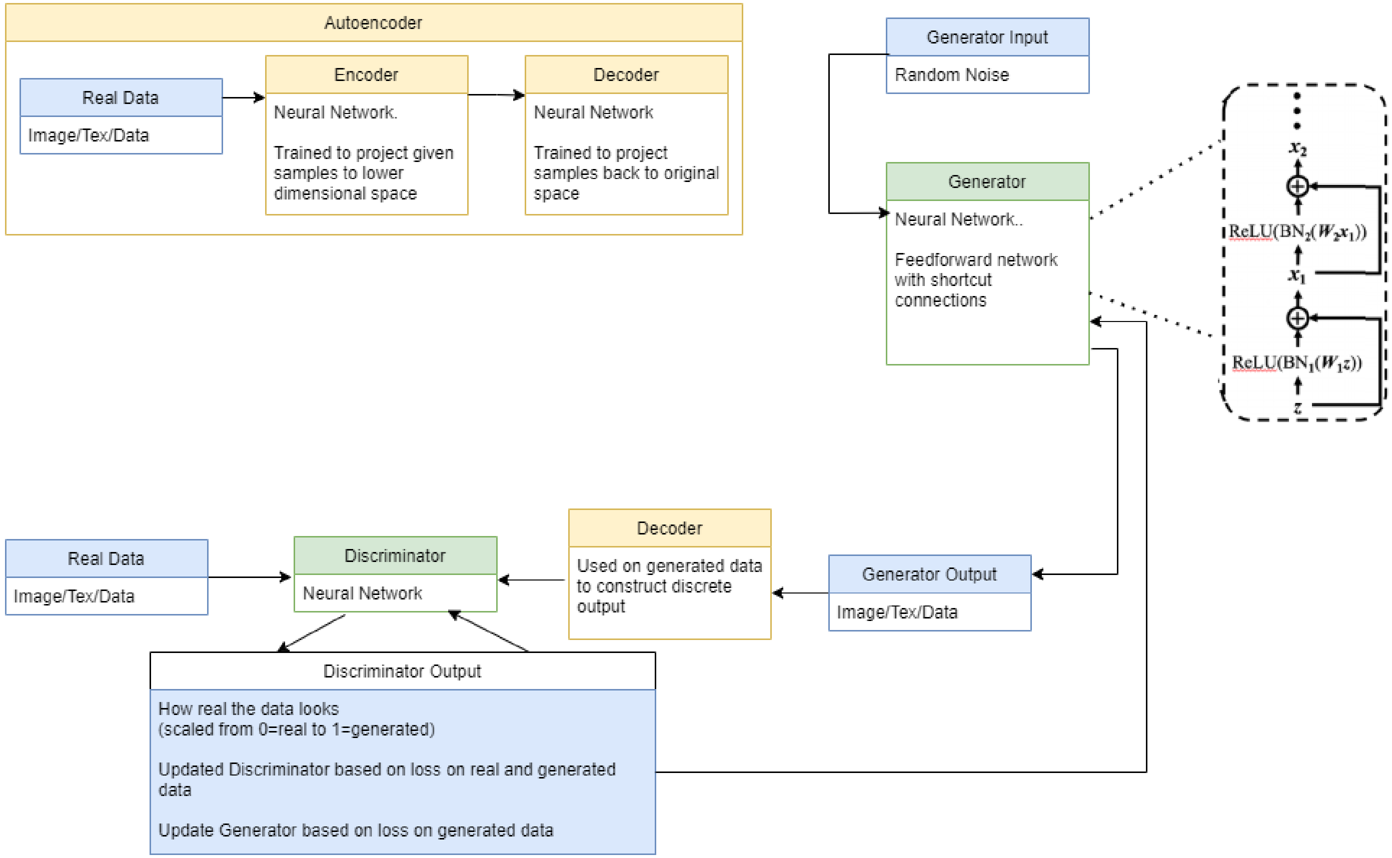
Need to generate realistic synthetic patient records, generate high-dimensional discrete variables (e.g. binary and count features)

Synthetic data need to achieve comparable performance to real data: distribution statistics, predictive modeling tasks, medical expert review

Result in limited identity and attribute disclosure

medGAN: combining an autoencoder with the original GAN to generate high-dimensional multi-label discrete samples

Introduce minbatch averaging to avoid mode collapse, more efficient



amesto NextBridge

GANS FOR EPIDEMIOLOGICAL DATA

mc-medGAN

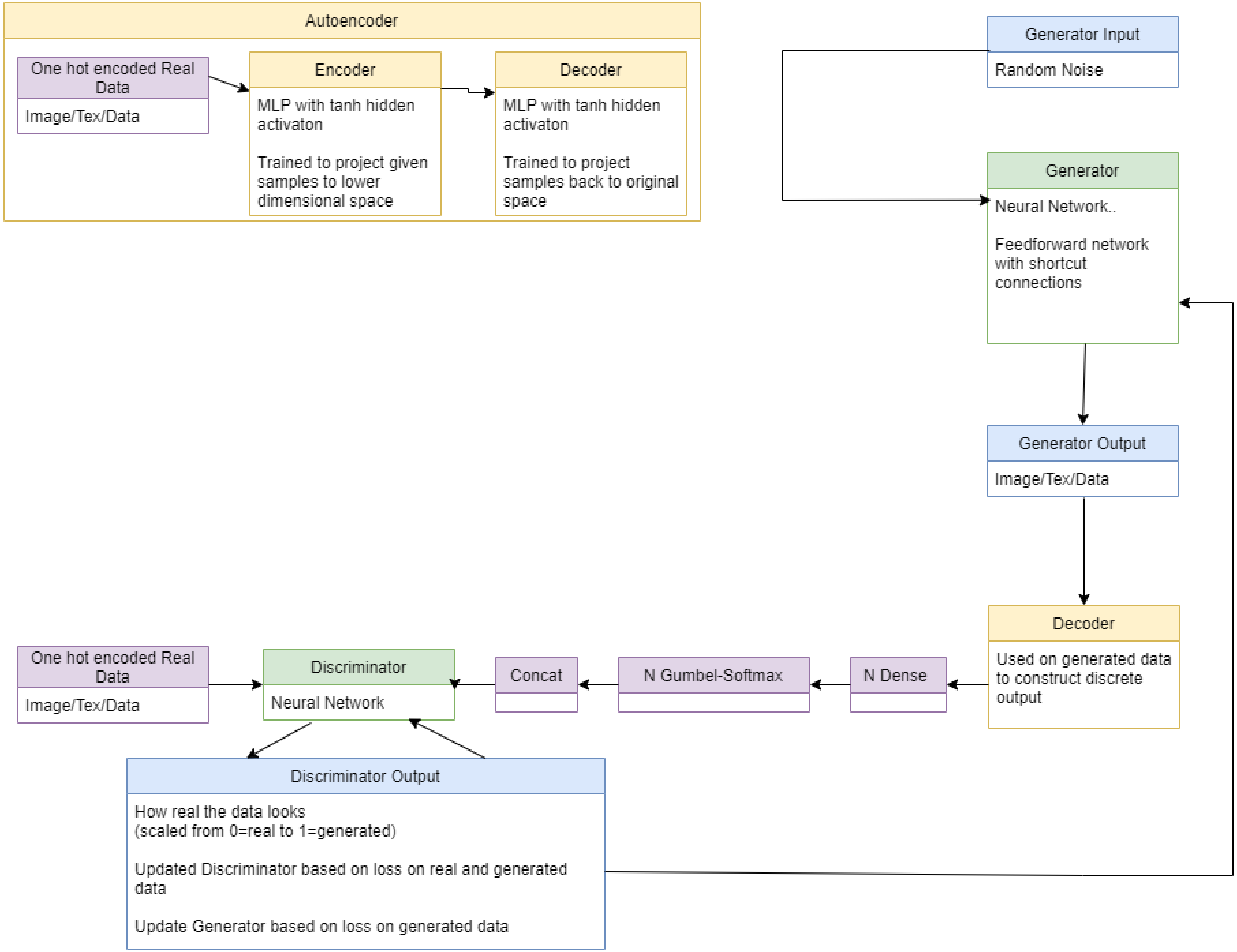
One-hot encoding of the multi categorical data

The decoder is modified by using a Gumbel-softmax activation after splitting the output with a dense layer per categorical variable

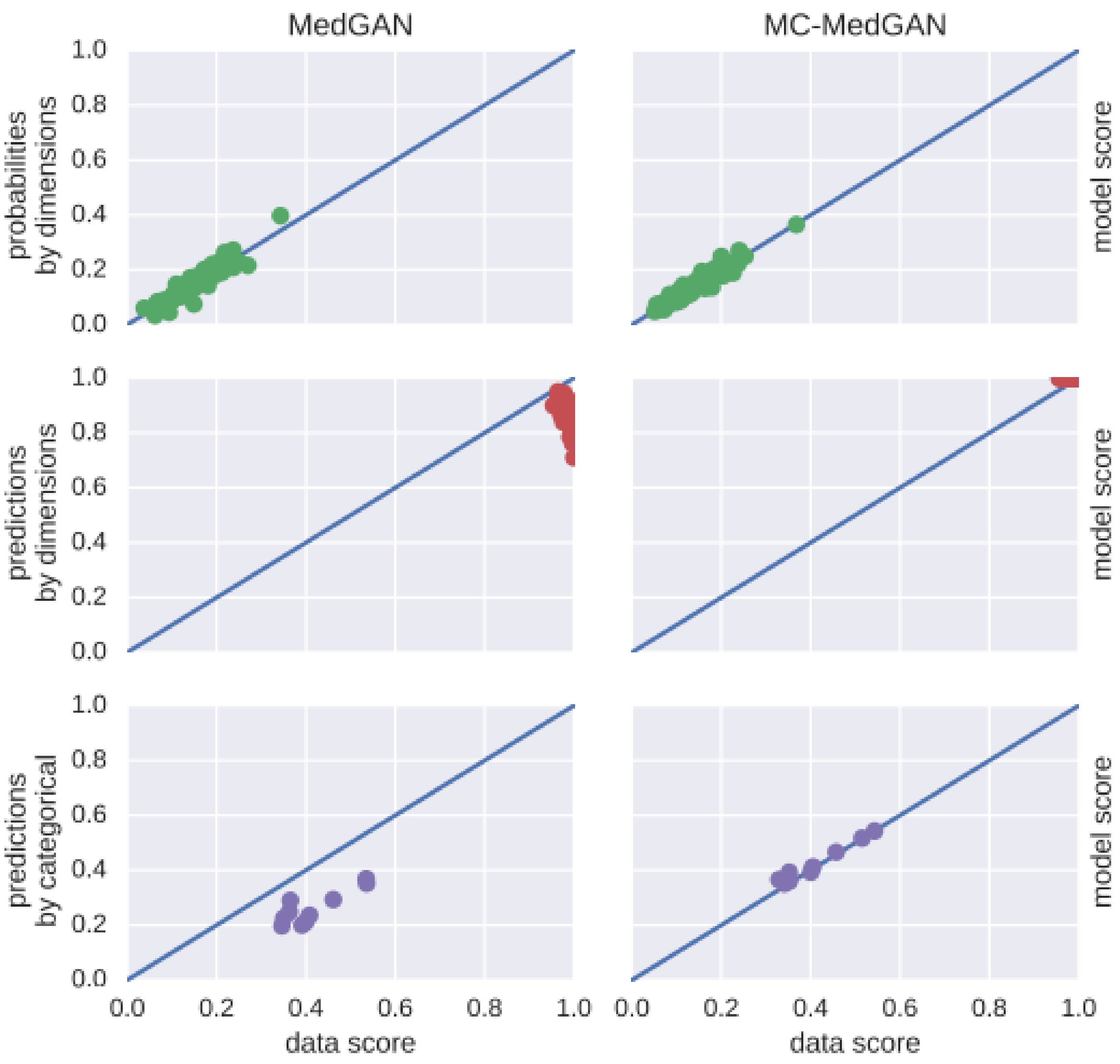
During training Gumbel-Softmax outputs are used separately to calculate the modified reconstruction loss

Color
Red
Red
Yellow
Green
Yellow

Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1



amesto NextBridge



2018 Camini et. al

Model evaluation

Information disclosure – How much of the real data can directly/indirectly be revealed

Data utility – gauge the extent of which the statistical properties of the real data are captured and transferred to the synthetic dataset

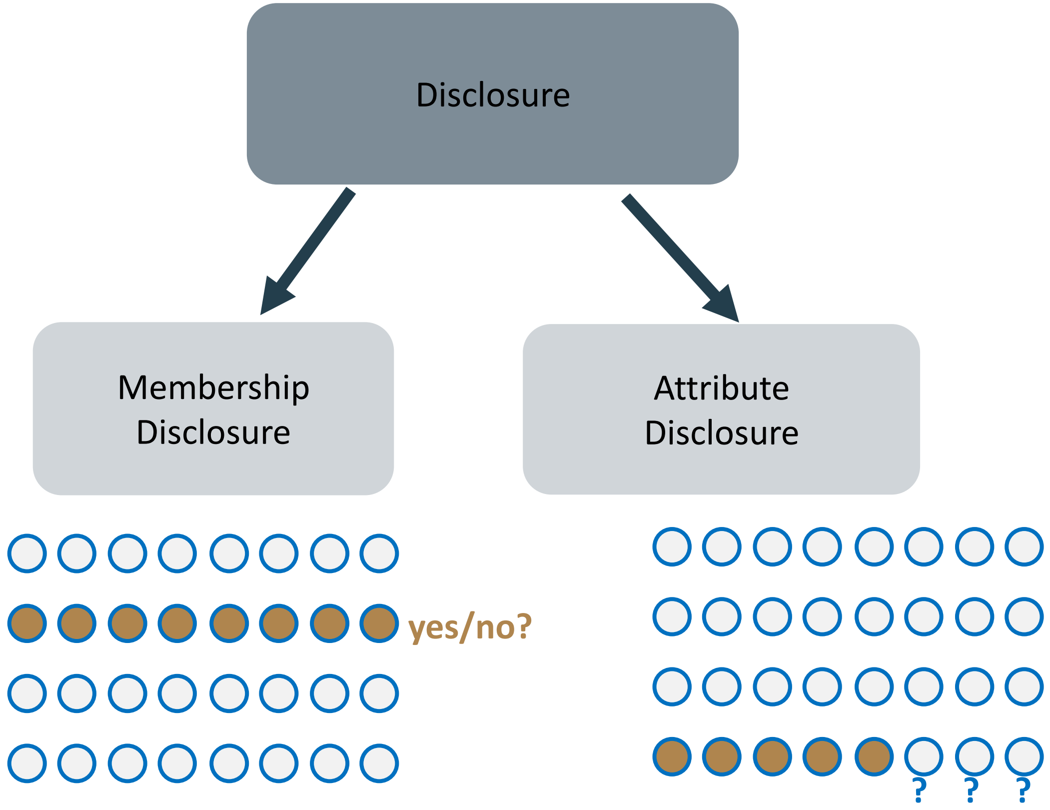
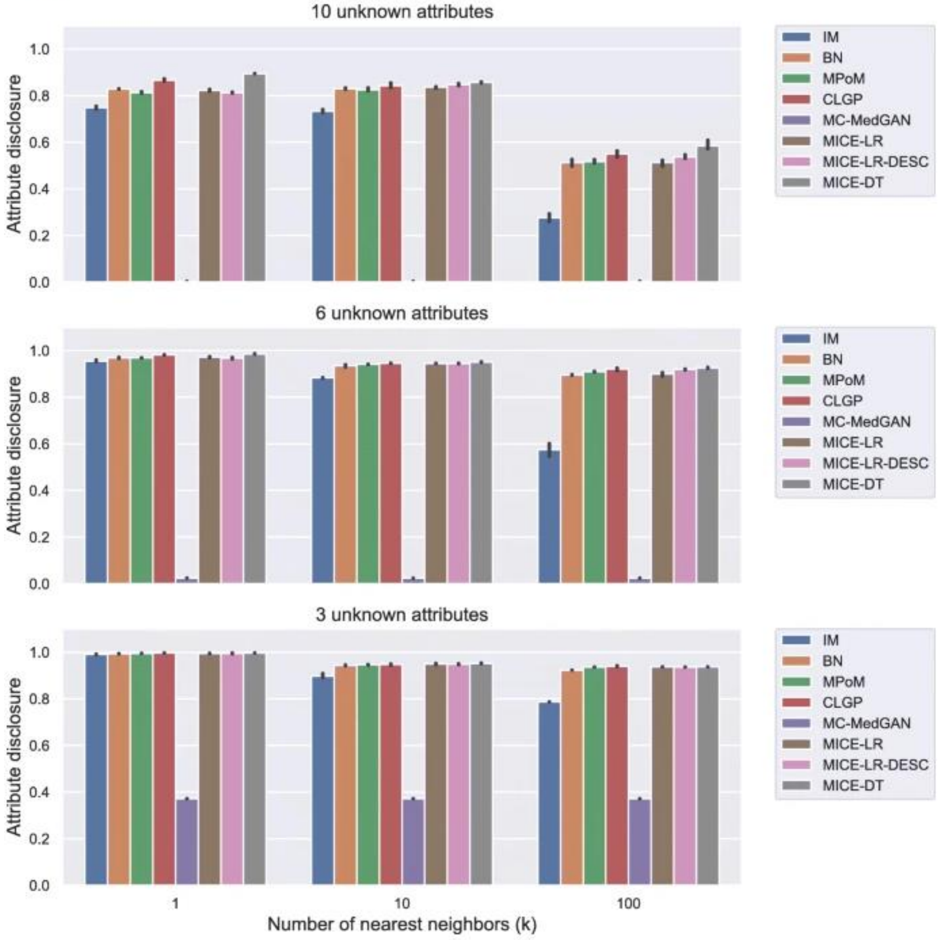


Fig. 17



Attribute disclosure for several values of nearest neighbors (k). BREAST `large-set`. Results show attribute disclosure for the case an attacker seeks to infer 10, 6, and 3 unknown attributes, assuming she/he has access to the remaining attributes in the dataset

Goncalves *et al. BMC Medical Research Methodology* (2020) 20:108
https://doi.org/10.1186/s12874-020-00977-1

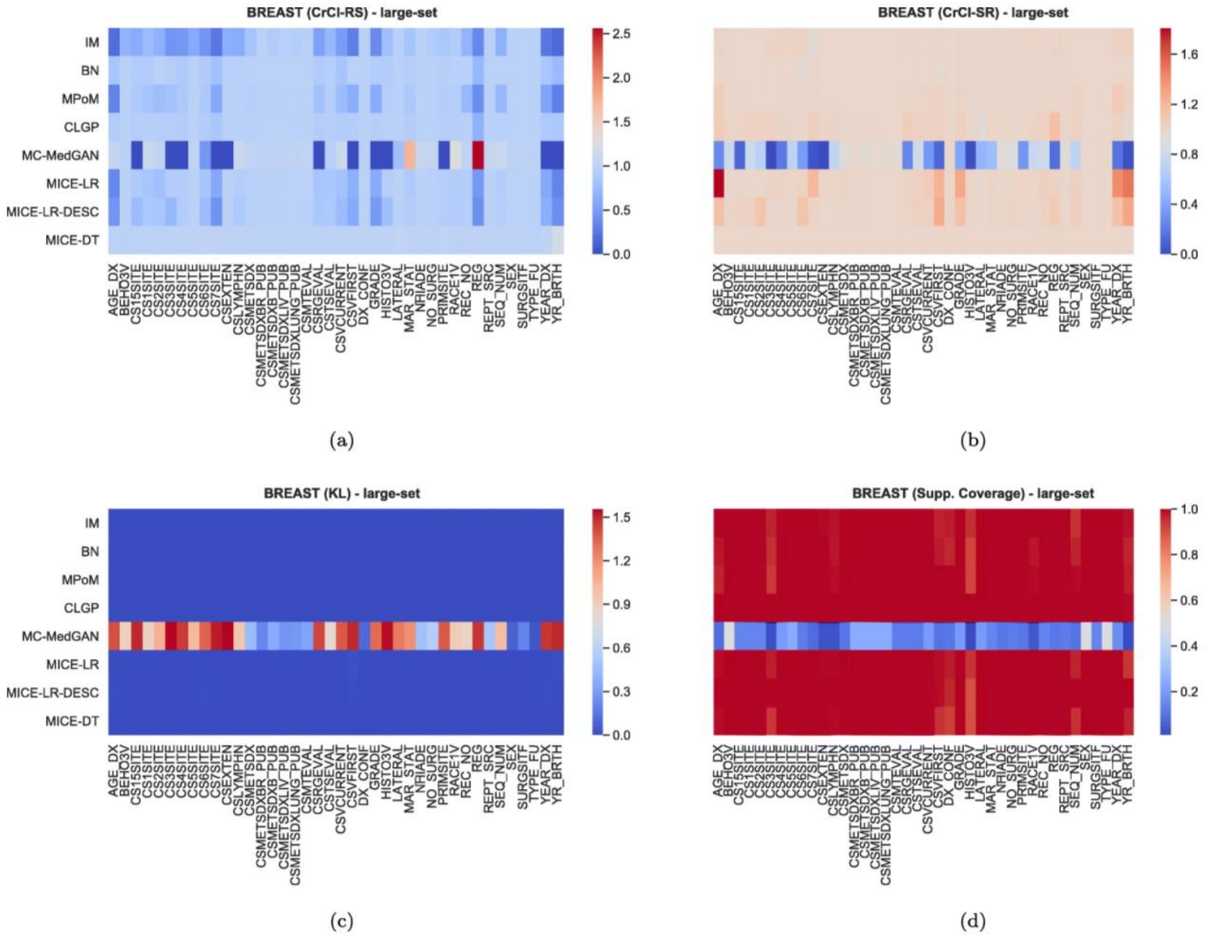
BMC Medical Research Methodology

RESEARCH ARTICLE **Open Access**

Generation and evaluation of synthetic patient data

Check for updates

Andre Goncalves[1*], Priyadip Ray[1], Braden Soper[1], Jennifer Stevens[2], Linda Coyle[2] and Ana Paula Sales[1]



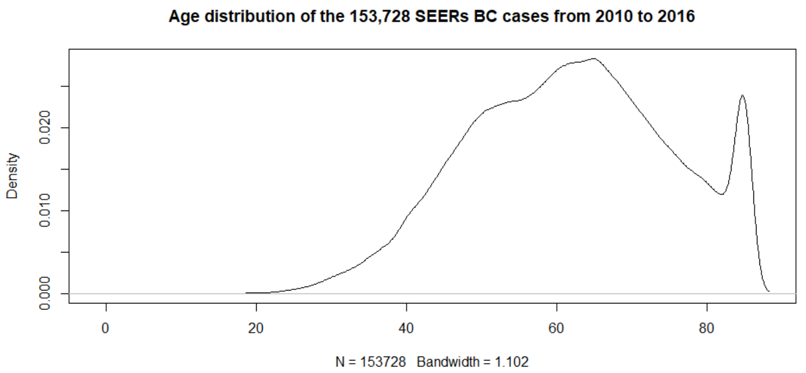
Heatmaps displaying the average over 10 independently generate synthetic datasets of (**a**) CrCl-RS, (**b**) CrCl-SR, (**c**) KL divergence, and (**d**) support coverage, at a variable level on BREAST `large-set`

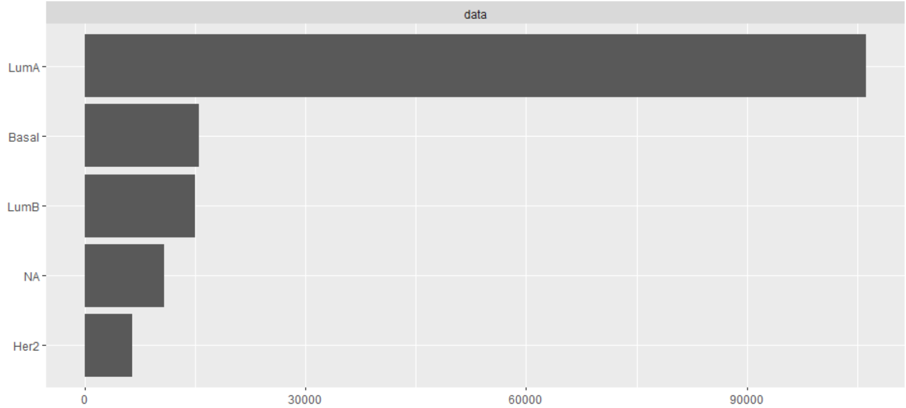
RESULTS FOR THE SAS HACKATHON

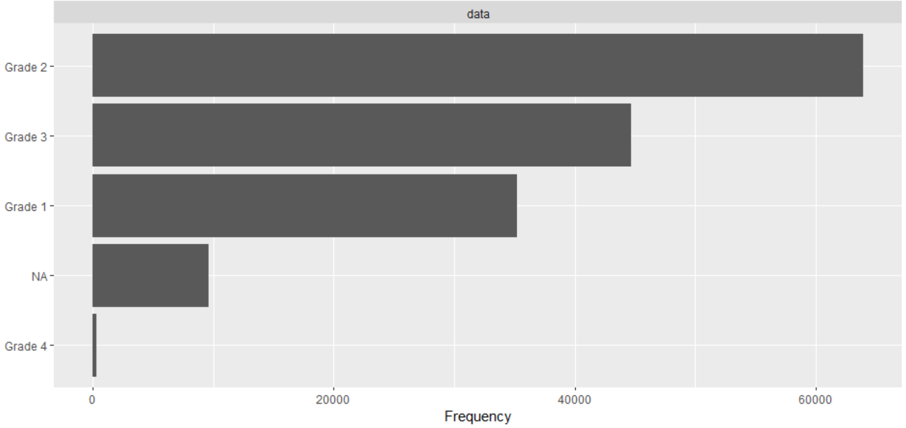
The SEERS data subset

The data set

- N = 153,728 cases
  - 152,490 females
  - 1,238
- Age distribution
- Can be divided into five subtypes:
  1. Luminal A, n=106,218
     - The most frequent BC subtype.
     - The tumor is estrogen positive with good prognosis (i.e., long term survival).
     - Patients with LumA tumors are given target therapy in the form of antiestrogen treatment such as tamoxifen.
  2. Luminal B, n=14,957
     - Estrogen and progesterone positive tumor with relatively good prognosis
     - The Lum B subtype is linked to a significantly worse prognosis than Lum A mainly due poorer response to antiestrogen treatment.
  3. Basal-like, n=15,408
     - Trippel negative tumor (Estrogen, progesterone and Her2 negative tumor)
     - The subtype with poorest outcome
  4. Her2, n=6,358
     - The 2nd worst subtype with respect to outcome
     - Her2 positive tumors
     - Receives anti-Her2 antibody treatment, e.g., trastuzumab
  5. Normal-like, n=10,787
     - The molecular profile of the tumor resembles normal breast tissue
     - God prognosis
- Tumor Grade
  - Tumor grade is based on how much the cancer cells look like normal cells
    - Higher grade results in poorer prognosis







amesto
NextBridge

Disclosure Probability

Attributes known to attacker

Attributes attacker tries to determine

MARITAL STATUS AT DX
RACE / ETHNICITY
SEX
AGE AT DIAGNOSIS
BIRTHDATE—YEAR
SEQUENCE NUMBER--CENTRAL
MONTH OF DIAGNOSIS
YEAR OF DIAGNOSIS
PRIMARY SITE
LATERALITY
HISTOLOGY (92-00) ICD-O-2
BEHAVIOR (92-00) ICD-O-2
HISTOLOGIC TYPE ICD-O-3
BEHAVIOR CODE ICD-O-3
GRADE
DIAGNOSTIC CONFIRMATION
TYPE OF REPORTING SOURCE
EOD—TUMOR SIZE
EOD—EXTENSION
EOD—EXTENSION PROST PATH
EOD—LYMPH NODE INVOLV
REGIONAL NODES POSITIVE
REGIONAL NODES EXAMINED
TUMOR MARKER 1
TUMOR MARKER 2
TUMOR MARKER 3
CS TUMOR SIZE

the number and sequence of all reportable malignant, in situ, benign, and borderline primary tumors, which occur over the lifetime of a patient.

the site in which the primary tumor originated

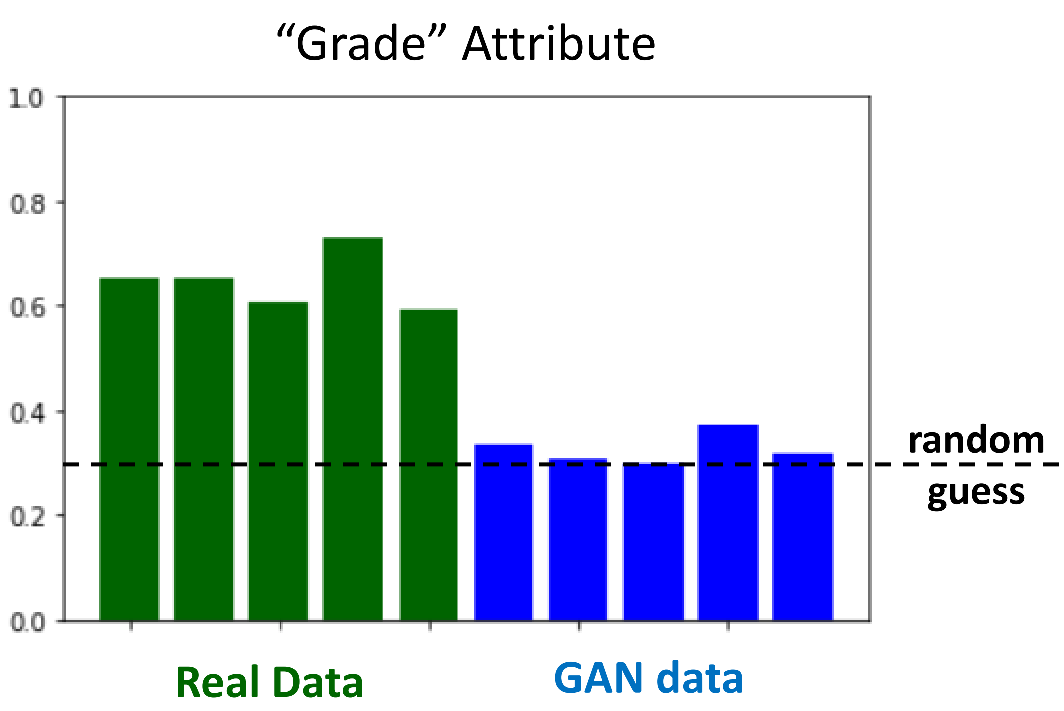
the side of a paired organ or side of the body on which the reportable tumor originated

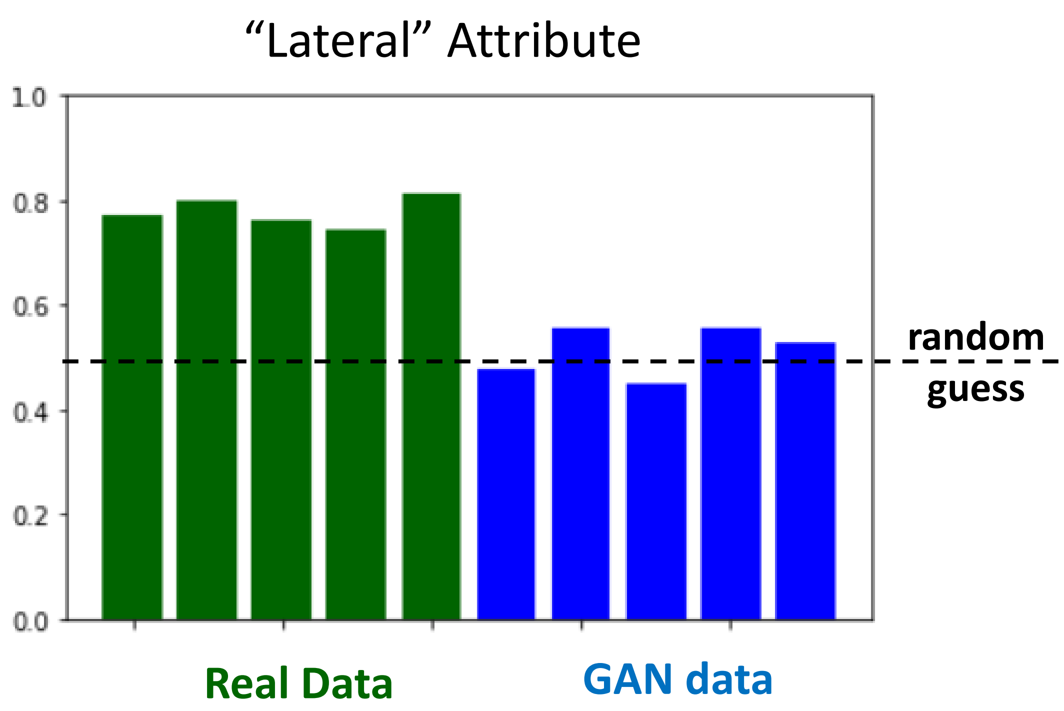
Code	Description
1	Grade I; grade i; grade 1; well differentiated; differentiated, NOS
2	Grade II; grade ii; grade 2; moderately differentiated; moderately differentiated; intermediate differentiation
3	Grade III; grade iii; grade 3; poorly differentiated; differentiated
4	Grade IV; grade iv; grade 4; undifferentiated; anaplastic

amesto
NextBridge

Disclosure Probability

for Real & MC-GAN synthetic data

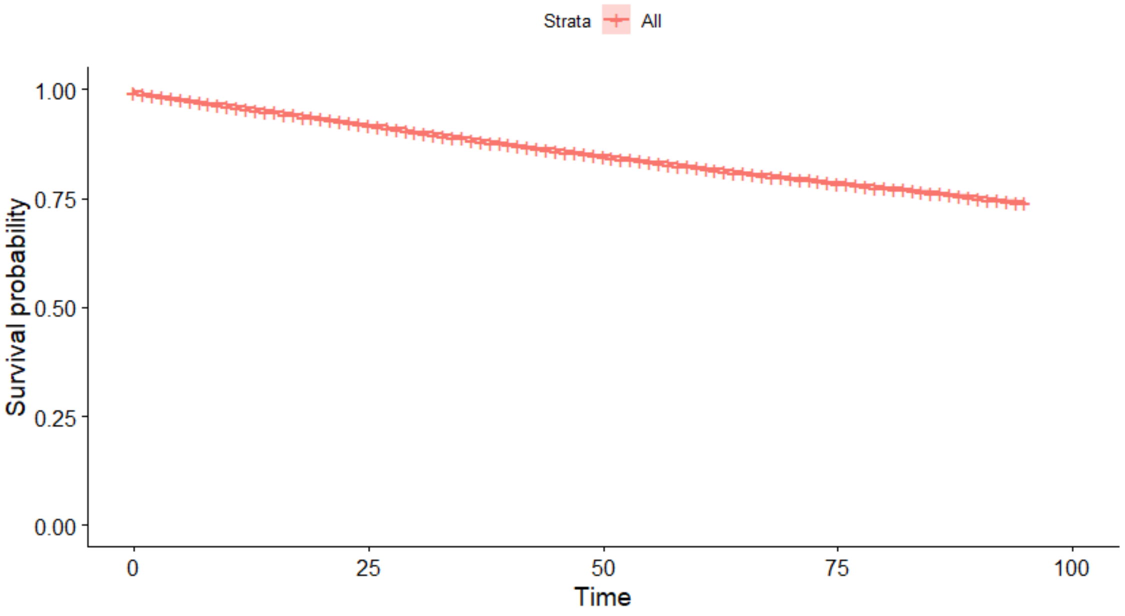




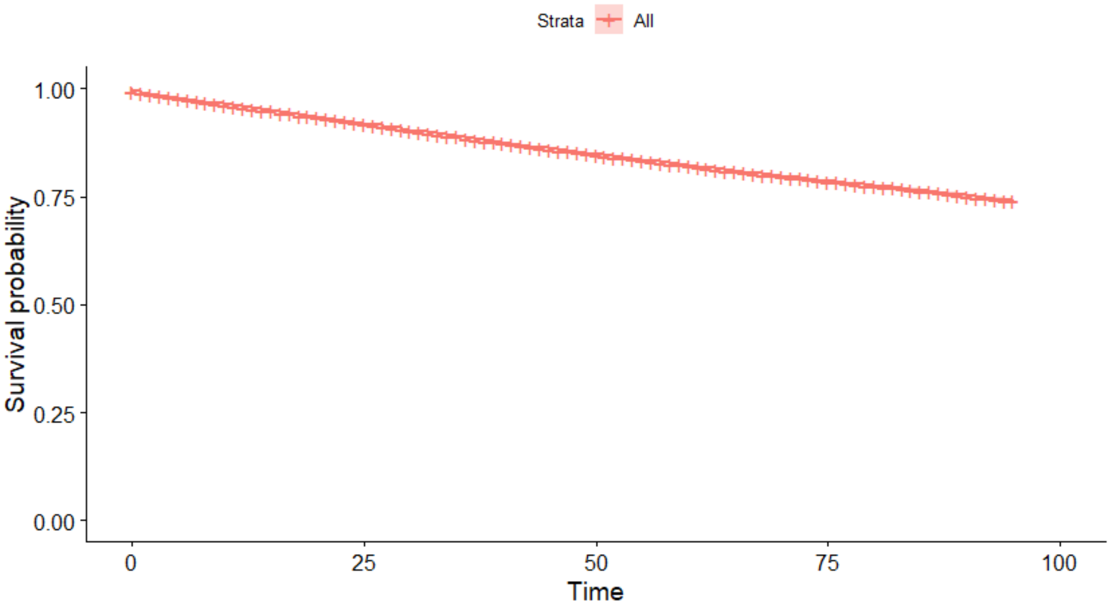
amesto NextBridge

Overall survival

Real data

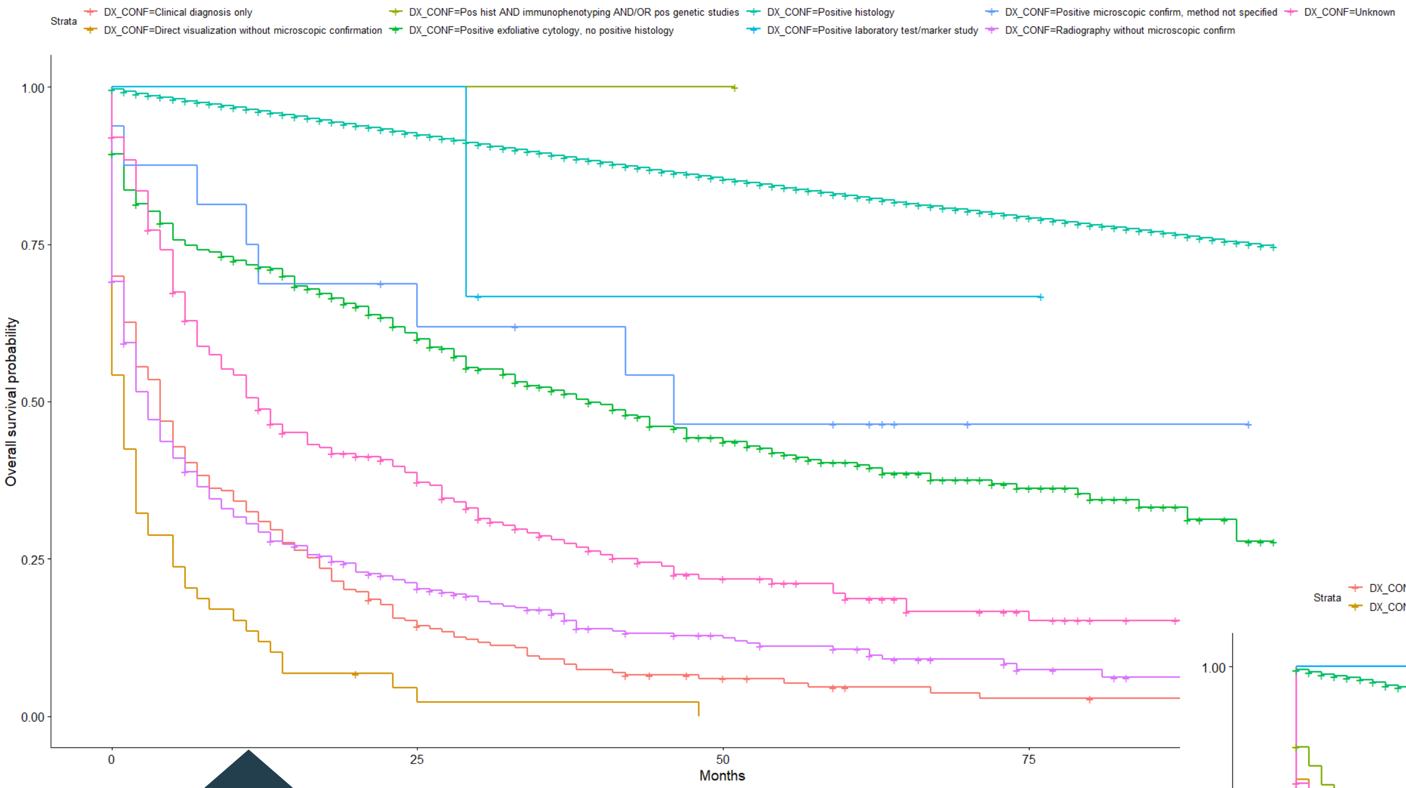


Synthetic data



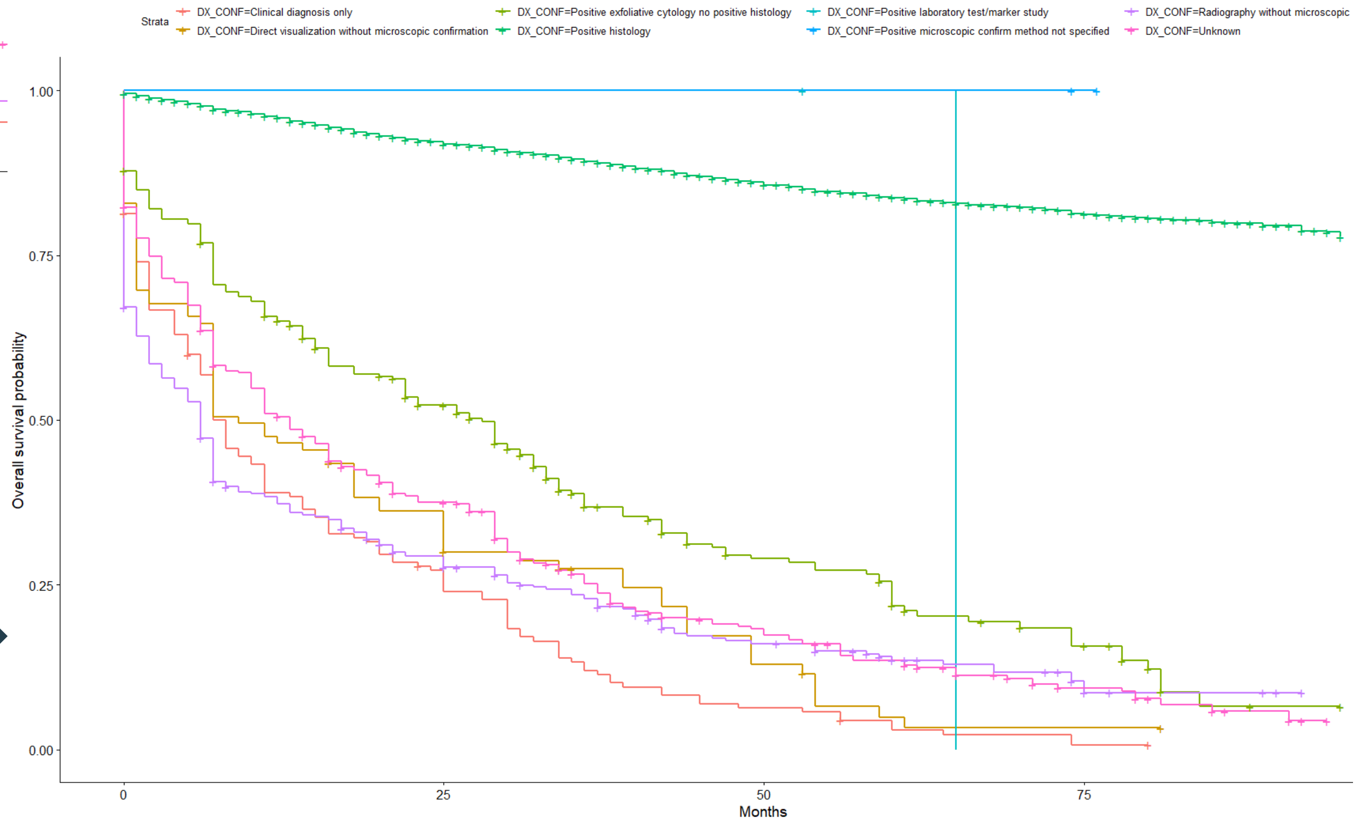
amesto NextBridge

Survival versus clinical characteristics



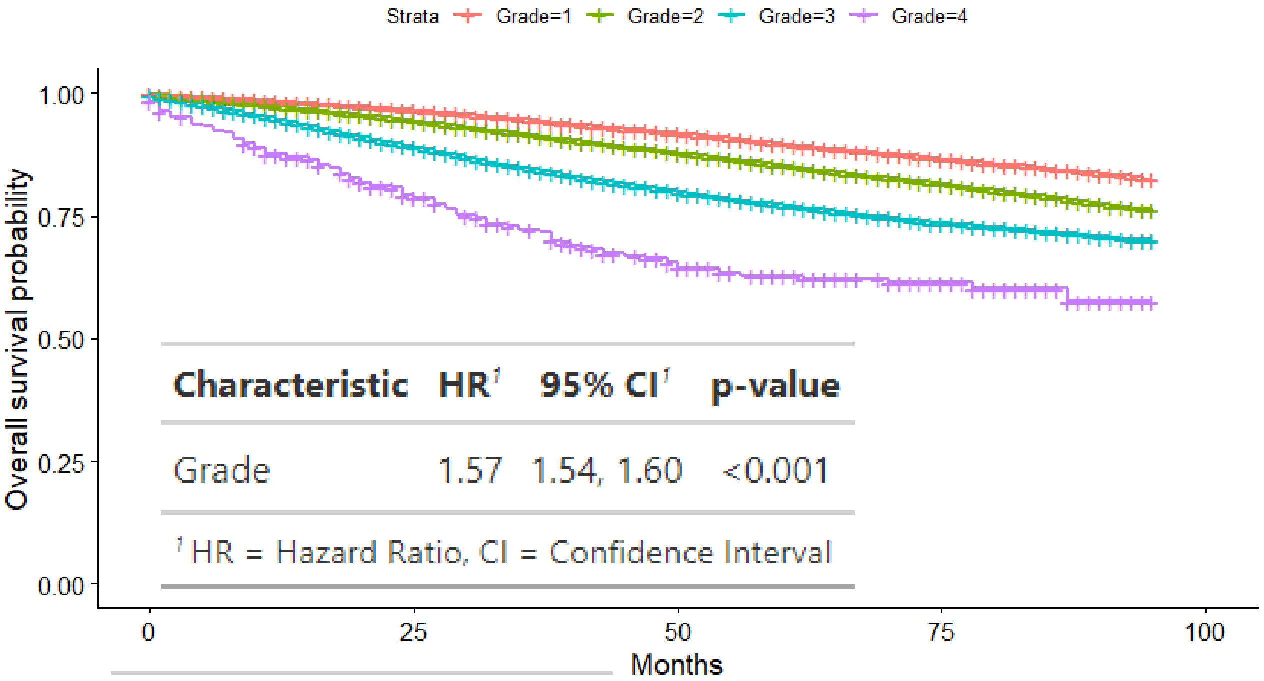
Real data

Synthetic data



Survival by breast cancer subtype

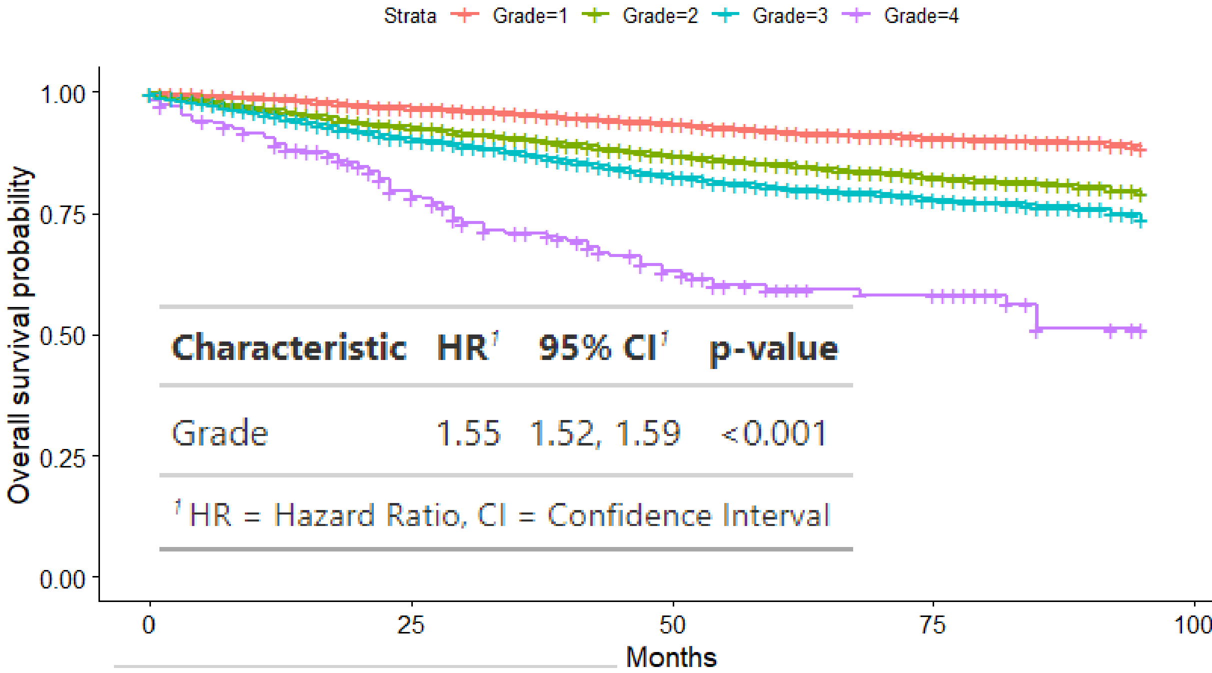
Real data



Characteristic	HR[1]	95% CI[1]	p-value
as.factor(Grade)			
1	—	—	
2	1.48	1.42, 1.54	<0.001
3	2.40	2.31, 2.51	<0.001
4	4.24	3.47, 5.17	<0.001

[1] HR = Hazard Ratio, CI = Confidence Interval

Synthetic data



Characteristic	HR[1]	95% CI[1]	p-value
as.factor(Grade)			
1	—	—	
2	2.01	1.92, 2.11	<0.001
3	2.66	2.54, 2.79	<0.001
4	6.28	5.13, 7.68	<0.001

[1] HR = Hazard Ratio, CI = Confidence Interval

Summary

MC-MedGAN

- has the best attribute
- produces synthetic data with poor data utility performance, indicating that the synthetically generated data does not carry the statistical properties of the real dataset
- relies on continuous embeddings of categorical data obtained via an autoencoder
- generated data show less then 1% failure when run through the SEER datachecks

Propose to make new medGAN variation, with alternative to one-hot-encoders and autoencoders

- Target encoding
- Leave-one-out encoding
- Bayesian Target
- Weight of evidence

Check out blog: https://www.toptal.com/machine-learning/generative-adversarial-networks

amesto NextBridge