region västerbotten

Synthetic data validation

Results from Region Västerbotten

Petter Lindgren, Data Scientist, Region Västerbotten/Sogeti

Contents

Introduction

The aim of this project is to synthesize original electronic medical records, and test how well machine learning models based on synthesized data can perform compared to models based on original data.

The individual synthetic datasets created by AI Sweden and Syndata have been evaluated using several metrics and methods as described in Bilaga 1 and Bilaga 2. That evaluation is without knowing the AI model algorithm that is supposed to be used on the data. Here, we validate how well the different synthesized datasets perform when training AI models in the same way as for the original data.

The AI models of interest regard length of stay at an intensive care unit. One model predicts the probability of being discharged due to recovery, another model predicts both the probability of being discharged by death and the probability of being discharged by other reasons, respectively. The training dataset of the original models was used as bases for synthesized data creation. Both models input data contains 105 variables and 1689 observations.

Synthesized data

Four synthetized datasets by Syndata were evaluated, three for discharge hypothesis and two for the recovered hypothesis. In addition to the datasets generated by Syndata, AI Sweden generated synthetic datasets applying open source code from the Synthetic Data Vault, using the algorithms ctgan, dp-ctgan, pate ctgan and tvae on each hypothesis. Based on initial evaluation on the datasets, only ctgan and dp-ctgan are being validated for AI model performance.

The files used for validation from Syndata are in the folder \\vs995\c$\syndata\data_to synthesize\RVB_deliverables\synthetic_datasets_delivered_by_syndata and named:

- discharged_synthetic_model1__1x_dataset.csv
- discharged_synthetic_model2__1x_dataset.csv
- discharged_synthetic_model3__1x_dataset.csv
- recovered_synthetic_model1__1x_dataset.csv
- recovered_synthetic_model2__1x_dataset.csv

The files used for validation from AI Sweden are in the folder \\vs995\c$\syndata\data_to synthesize\ai.se_datasets and named:

- discharged_ctgan_synthetic_dataset_1.csv
- discharged_dp-ctgan_synthetic_dataset_1.csv
- recovered_ctgan_synthetic_dataset_1.csv
- recovered_dp-ctgan_synthetic_dataset_1.csv

The original data is in the folder \\vs995\c$\syndata\data_to synthesize and named:

- discharged_data_train.csv
- recovered_data_train.cs

Building models

For each of the synthetic data sets, a random forest survival analysis model (Ishwaran, o.a., 2014) was built with the same settings as for the pre-trained models. In that way, the only difference between the pre-trained model and the models based on synthetic data is the training data itself.

For each model during training, random forest uses out-of-bag error rates, which is a good estimate of the real error rate for unseen observations.

An error rate in random forest analysis models can here be interpreted how well the predictor correctly ranks (classifies) two random individuals in terms of survival. A value of 0.5 is no better than random guessing. A value of 0 is perfect.

Validating models

Each model was validated on unseen original validation data to find out the true error rate as well as the estimated discharge rate (survival function) from the model as compared to the true discharge rate. The survival function of the model was estimated using Kaplan Meier estimates (Kaplan & Meier, 1958)

We also examined the statistics of the distribution of the output variables in the synthetic dataset compared to the original dataset, to see the overall preconditions of creating good AI models from the synthetic data.

To examine the AI model robustness of the synthesized data set, we created synthesized data set according to each trained GAN algorithm and evaluated how the model performance fluctuated from dataset to dataset. This was presented as standard deviation of the error rate.

Results

Distribution of output variables

Statistics of the distribution of the response variables “facit_vardlangd_kvar” for the respective hypothesis are shown in Table 1 and Table 2.

For the discharge hypothesis, syndata datasets 1 and 2 generally had lower values than original datasets. After pointing that out to Syndata, they did a third model with more focus on getting the distribution of response variable as similar as possible to the original dataset distributions. Consequently, syndata 3 is the most similar synthetic dataset to the original dataset. For open source datasets, the ctgan seems to have a better distribution than dp-ctgan.

For the recovery hypothesis, open source datasets generally have too high values and datasets from Syndata generally have too low values. However, Syndata 1 has a good fit regarding 1st Quantile, median and mean.

Table 1. descriptive statistics of variable "facit_vardlangd_kvar" for discharge hypothesis

	Min.	1st Quantile	Median	Mean	3rd Quantile	Max.
Original	0,38	6,92	13,20	56,44	56,57	1673,70
Syndata 1	0,38	0,38	2,66	20,18	14,42	685,35
Syndata 2	0,38	4,35	15,08	28,28	39,28	282,60
Syndata 3	0,38	3,62	11,85	54,62	36,53	1673,70
Open source ctgan	0,46	10,49	18,10	58,98	61,52	728,39
Open source dp-ctgan	0,41	36,27	73,99	226,07	206,94	1672,66

Table 2. descriptive statistics of variable "facit_vardlangd_kvar" for recovered hypothesis

	Min.	1st Quantile	Median	Mean	3rd Quantile	Max.
Original	0,38	6,92	13,20	56,44	56,57	1673,70
Syndata 1	0,38	6,35	14,82	55,46	24,20	1355,96
Syndata 2	0,38	4,90	17,34	31,18	43,29	380,38
Open source ctgan	0,39	16,13	43,99	88,07	82,64	1104,15
Open source dp-ctgan	0,42	21,63	44,76	89,38	66,77	1656,83

Statistics of the distribution of the response variables "facit_avliden2" for the respective hypothesis are shown in Table 3 and Table 4.

For the discharge hypothesis Syndata 3 and Open source ctgan has the most similar distribution to original dataset. However, the number of missing is somewhat increased.

For the recovery hypothesis, open source ctgan has the most accurate distribution despite the increased number of missing data.

Table 3. descriptive statistics of variable "facit_avliden2" for discharge hypothesis

	Discharge due to other reasons	Discharge du to decease	NA's
Original	1537	150	2

Syndata 1	1612	77	1
Syndata 2	1685	5	0
Syndata 3	1549	128	13
Open source ctgan	1498	184	7
Open source dp-ctgan	1641	33	15

Table 4. descriptive statistics of variable "facit_avliden2" for recovered hypothesis

	Censored	**Discharged due to recovery**	**NA's**
Original	308	1379	2
Syndata 1	101	1589	0
Syndata 2	82	1608	0
Open source ctgan	364	1314	11
Open source dp-ctgan	18	1654	17

Error rate

The error rates of the different hypotheses are shown in see Table 5 and Table 6.

Generally, Syndata models 1 and 2 did perform almost as well as original model. However, and surprisingly, the error rate was considerably worse at training compared to validation. This is uncommon, but the reason could be that different populations are used at training (synthetic) as compared to validation (original). Open source models had higher validation error rate than all other models.

Table 5. AI model performance for discharged hypothesis

	Other reason		**Deceased**	
Model	**Train OOB Error rate**	**Validation Error rate**	**Train OOB Error rate**	**Validation Error rate**
Original	0.264	0.279	0.506	0.472
Syndata model 1	0.366	0.293	0.620	0.574
Syndata model 2	0.303	0.272	0.800	0.547
Syndata model 3	0.312	0.310	0.662	0.540
Open source ctgan model	0.492	0.486	0.525	0.565
Open source	0.518	0.348	0.511	0.548

dp-ctgan model				

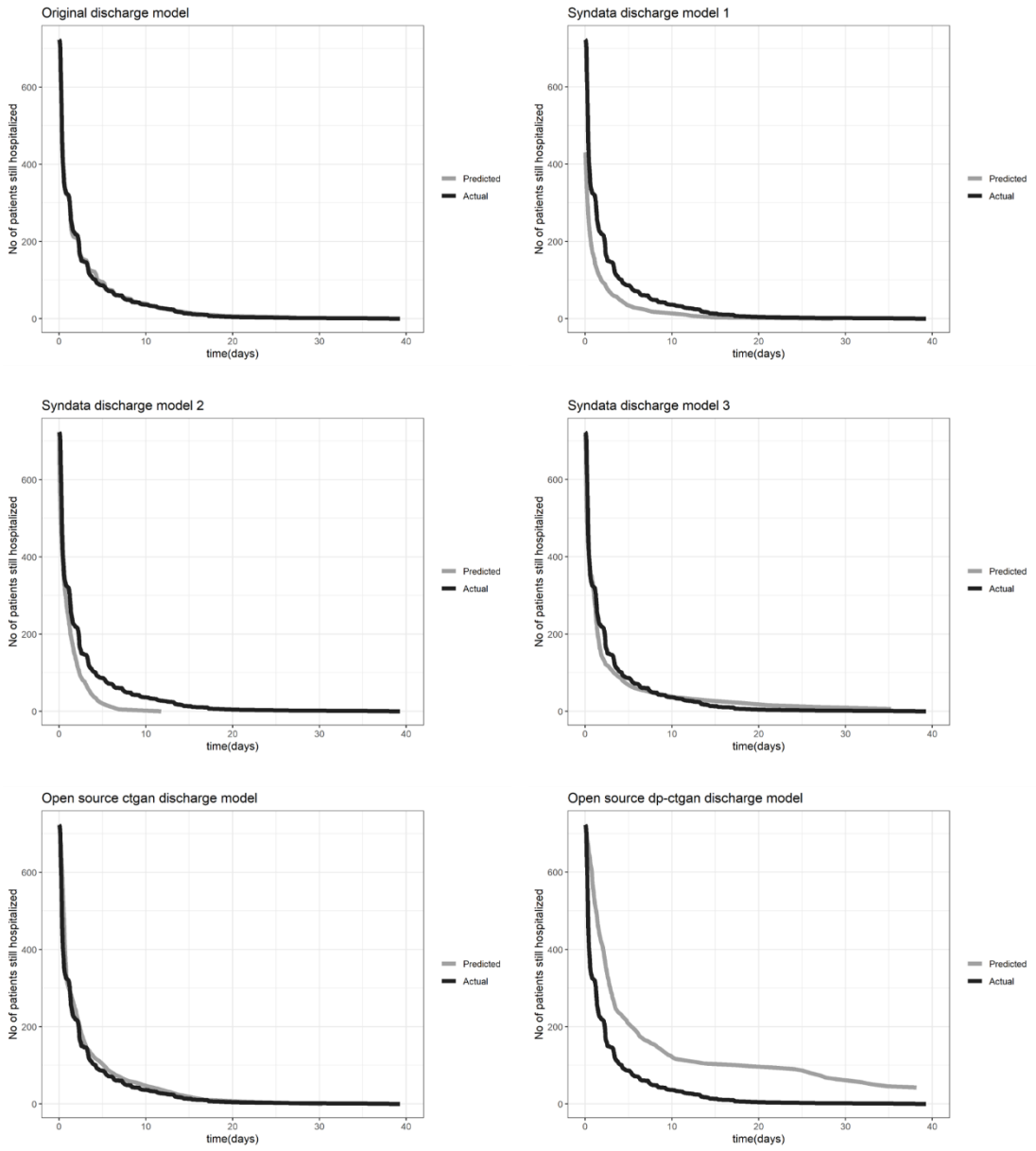
Table 6 AI model performance for recovered hypothesis

	Recovered	
Model	**Train OOB Error rate**	**Validation Error rate**
Original	0.206	0.209
Syndata model1	0.356	0.236
Syndata model 2	0.293	0.226
Open source ctgan model	0.496	0.411
Open source dp-ctgan model	0.508	0.436

Discharge rate

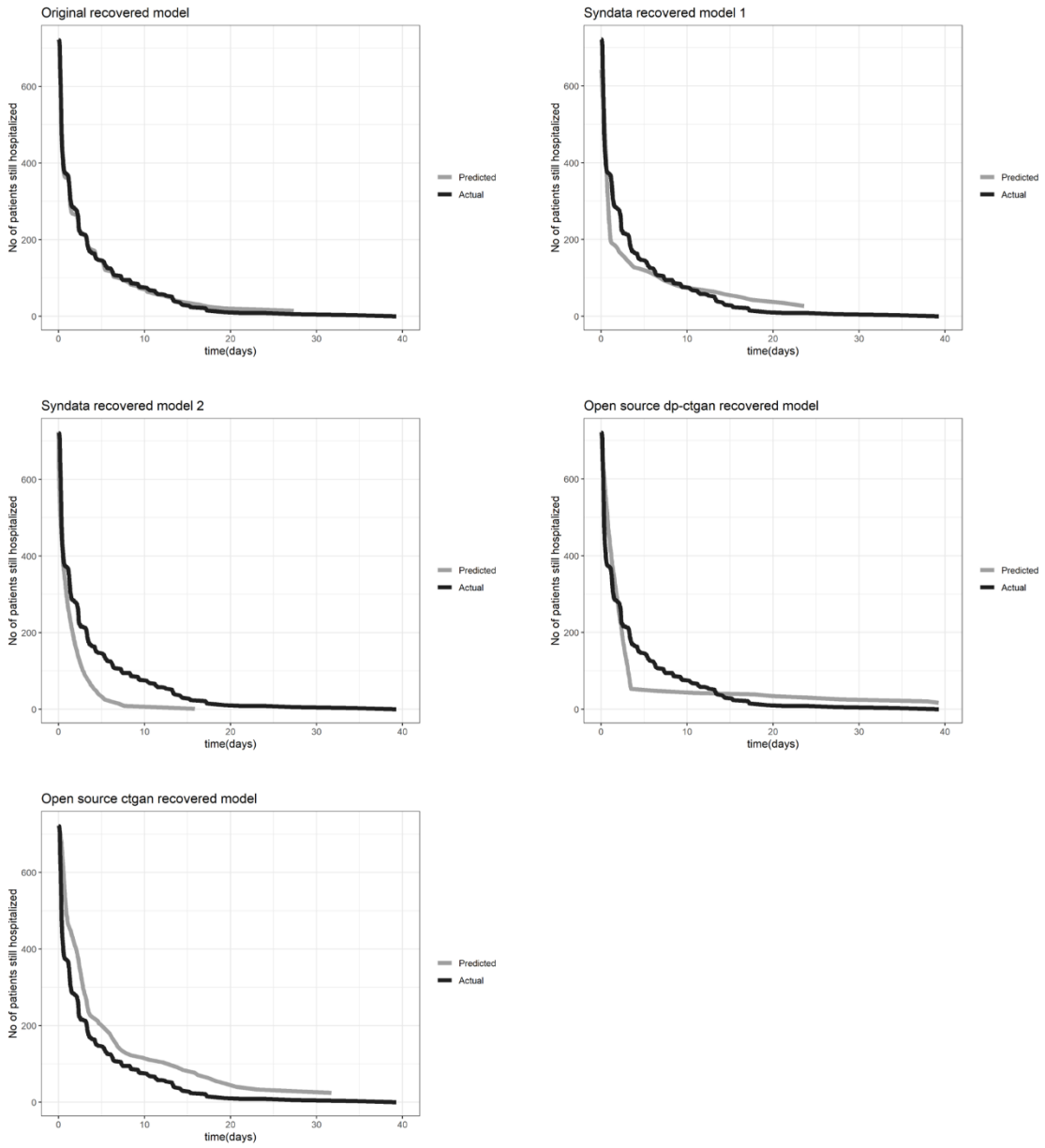
The overall discharge rate for discharge hypothesis for the different models is presented in Figure 1. Syndata models 1 and 2 predicted a faster rate than reality, while Syndata 3 and open source data ctgan model almost had as good fit as the original model. The reason for the fast discharge rate on Syndata model 1 and model 2, is probably due to the difference in distributions for output variables as compared to the original data set, see Table 1 and Table 3.

Figure 1. Survival function for models on discharge hypothesis vs truth on validation data



The overall discharge rate for recovered hypothesis for the different models is presented in Figure 2. Here, the interpretation is not as clear as for discharge hypothesis. The original model has the best fit and Syndata model 1 is second best while the other models are deviating from the actual survival curve in different ways.

Figure 2. Survival function for models on recovered hypothesis vs truth on validation data



Robustness

To evaluate the robustness of the model we generated ten datasets of each GAN model, ran separate survival analysis on each dataset and evaluated the variance of the error rate for the ten runs. The goal is that AI models based on synthetic data should behave stable, and not depend on random fluctuation of the synthetic data generated by the GAN model.

Table 3 describes the mean and variance of error rate when training survival analysis models on synthetic data for the discharge hypothesis. In all Syndata models, the variance of the error rate at validation is small, meaning that all GAN models generate datasets of similar behaviour from run to run. For open source models, the variation is bigger than for Syndata models but still acceptable.

Table 3. Robustness of error rate for discharge due to other reason than deceases

	Train		Validation	
Model	**Error rate Mean**	**Error rate sd**	**Error rate Mean**	**Error rate sd**
Syndata model 1	0.372	0.0043	0.291	0.0045
Syndata model 2	0.298	0.0031	0.270	0.0025
Syndata model 3	0.306	0.0056	0.305	0.0055
Open source ctgan model	0.493	0.0069	0.461	0.0480
Open source dp-ctgan model	0.509	0.0139	0.375	0.0394

To evaluate the robustness of variable importance, we investigated how the variable importance ranking list match between the ten different survival analysis runs on dataset produced by Syndata model 1 for the discharge hypothesis. All ten runs did have the same ranking of the ten most important variables indicating a very strong robustness of variable importance (data not shown). Due to clear result and the fact that measuring variable importance is very computer intensive, we didn't evaluate this robustness for the other synthetic data models.

Discussion

Generally, the result of using synthetic data for creating machine learning models on medical records looks promising. Especially impressive is how the Syndata models have low error rate, i.e., they are capable of ranking patients according to the care length.

The predicted discharge rate of Syndata's models were at first faster than reality but after correcting distribution of output variables, the rate turned out to be much more accurate. Unfortunately, as the discharge rate became better, the error rate increased. It might be that the correlation between the output variables and explanatory variables were diminished when optimizing the GAN training to produce as accurate output variable distribution as possible. More work needs to be done to investigate if we can keep the good ranking of patients from Syndata model 1 and 2 and get as good discharge rate as Syndata model 3.

The open source dataset has similar performance regarding individual variable distribution as Syndata datasets but has higher error rate throughout. The reason for the high error rate is probably due to that open source scripts are by default optimized to get as similar distribution as possible and not to keep the correlation structure in place.

Litterature

Ishwaran, H., Gerds, T. A., Kogalur, U. B., Moore, R. D., Gange, S. J., & Lau, B. M. (2014). Random survival forests for competing risks. *Biostatistics, 15*(4), 757-773. Hämtat från https://academic.oup.com/biostatistics/article/15/4/757/266340 den 14 12 2021

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc., 53*(282), 457–481. Hämtat den 14 12 2021