Using Data Synthesis to Improve Length of Stay Predictions for Patients with Rare Diagnoses

Simon Schiff
German Research Center
for Artificial Intelligence
Ratzeburger Allee 160
23562 Lübeck
Germany

Sebastian Wolfrum
University Medical Center
Schleswig–Holstein
Campus Lübeck
Ratzeburger Allee 160
23538 Lübeck

Ralf Möller
German Research Center
for Artificial Intelligence
Warburgstraße 28
20354 Hamburg
Germany

Mattis Hartwig
German Research Center
for Artificial Intelligence,
singularIT GmbH
Lübeck, Leipzig
Germany

Abstract

In healthcare, managing small patient cohorts, particularly those with rare diseases, presents a unique challenge due to the scarcity of data required for effective machine learning applications. Addressing this issue, our paper investigates if a specific conditional data synthesis prior to learning the machine learning model using the CTGAN architecture improves the result. Data synthesis refers to the artificial generation of synthetic data with certain properties from the original data. We choose the specific learning task of predicting hospital length of stay (LoS) of patients leaving the emergency department. It can, e.g., be used to predict the bed occupancy in a hospital and thus enables better planning. The accuracy of the LoS-prediction is strongly dependent on rarity of the patients disease, ranging from an acceptable accuracy, e.g., for often occurring homogeneous cases to worse accuracy for, e.g., inhomogeneous and rare ones. To increase the accuracy for such cohorts, we enrich the dataset with new, synthesized patient admissions. Then, for each cohort, a model is trained to predict the LoS of a patient of this cohort. Our experiments show that adding synthetic data is able to increase the accuracy for the majority of cohorts. Indicators for a benefit of synthetic data seem to be cohorts that have a high LoS with high variance within in cohort.

Introduction

In machine learning, having data to learn from is critical. In the medical domain, data is often very scarce, especially, for patient cohorts with rare diseases or other special characteristics. The scarcity and specificity of data in these instances pose significant obstacles for traditional ML approaches, which typically thrive on large datasets. These smaller cohorts often lack the volume of data necessary to train robust and accurate models, leading to suboptimal performance for these groups. The issue of data scarcity is further compounded by stringent privacy regulations in healthcare, which restrict the use and sharing of patient data.

In this paper, we investigate the task of predicting hospital length of stay (LoS) for patients leaving the emergency department with rare diagnoses. Emergency units of hospitals require efficient and quick actions to cope with the amount of patients and time-critical tasks. Therefore, there is a need for optimizing and automatizing such processes as much as possible. In this study, we use the well-known freely accessible MIMIC-IV electronic health record (EHR) dataset (Johnson et al. 2023). The data consists, beside a few often occurring cases, of many small cohorts, as depicted in Figure 1.

The problem of unbalanced data occurs in many domains and is a widely researched topic. Different solutions, such as over-sampling (e. g. synthetic minority oversampling technique (SMOTE)), under-sampling, specific error functions, and many more can be found in the literature (see, e.g., the overview of (Wang et al. 2021)). SMOTE is a good example where individuals are not simply copied but where new synthetic individuals are created with a nearest neighbors based approach (Chawla et al. 2002). All these approaches assume that we have a large dataset, with patient's sharing a common data structure where the different classes just need to be balanced. In reality, practitioners often have selected small datasets that are not part of a big global database. Due to privacy regulations they cannot easily add more patients from other cohorts to their study. Therefore, we take a different approach in this paper and synthesize patients with the same condition as the patients in the cohort from a generative model based on the CTGAN (conditional generative adversarial network) architecture (Xu et al. 2019). In this paper, we loosely define data synthesis as the process to generate artificial data based on real data to be used in a specific task (Jordon et al. 2022b). According to the authors, CTGAN's synthesized data can replace real data for data science. Several studies have shown promising results for synthesizing data with CTGAN to learn machine learning models (Kuo and others 2020; Bourou et al. 2021) while others have also identified open issues to be solved (Mendikowski and Hartwig 2022).

To imitate the described use case, we split the dataset into cohorts per diagnosis and train and evaluate the individual models solely for the corresponding cohort. We use the conditional synthetic patient admissions to enrich the cohorts with artificial patients sharing the leading diagnoses of the corresponding cohort. We use two baselines for predicting the LoS of specific cohort: a model trained solely on the

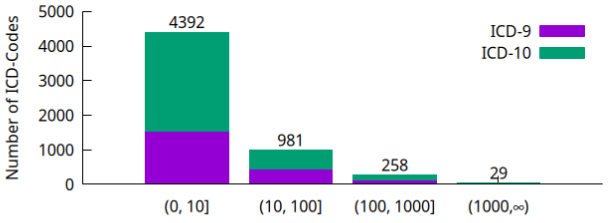

Figure 1: Size of all 5660 patient cohorts aggregated by their ICD-code (Primary diagnose)

training data from this cohort and a global model using all patients except the test set (a small sample of the specific cohort). That test set is used to evaluate the overall accuracy of that trained model.

The opinion on enriching medical datasets with synthesized data is ambivalent (see, e.g., (Hittmeir, Ekelhart, and Mayer 2019; Kuchin, Mukhamediev, and Yakunin 2020). Our experiments, however, show that in the context of the prediction of the LoS, it is a promising approach for creating machine learning models for small cohorts.

If a cohort is relatively large, training a model by adding synthetic data to the large cohort does almost nothing to improve the mean absolute error. As a rule of thumb, adding synthetic improves the trained model if the average LoS of the cohort is high and has a high variance in comparison to the whole dataset.

Our paper is structured as follows: After a discussion of the related work, we present the dataset MIMIC-IV in detail. Then, we describe the methodology of our experiments including the baseline model and the model training with the synthesis step. Afterwards, the results are evaluated and discussed. The paper ends with a short conclusion.

Related Work

Data synthesis is used in many contexts, e.g., for preserving privacy, de-biasing data or balancing imbalanced datasets (Jordon et al. 2022a). In this paper, the focus is to synthesize data from a model learned from a broader patient database to use for machine learning when the cohort of interest is relatively small. The related work can be structured around work on tabular data synthesis, using synthetic data for machine learning and our application of predicting hospital length of stay.

Approaches of synthesizing data are dependent on the properties of the data, here, as electronic health records (EHRs) are considered, tabular synthesis is of relevance. A widely acknowledged technique in this domain involves generative adversarial networks (GANs), especially CTGANs (Xu et al. 2019; Torfi, Fox, and Reddy 2022; Sauber-Cole and Khoshgoftaar 2022). These have been extensively used and further developed (Fang, Dhami, and Kersting 2022; Mendikowski and Hartwig 2022). There are also other approaches to synthesize data, e.g., a rule-based creation of new patient data, as in Synthea (Walonoski et al. 2017).

There are a lot of evaluation studies that investigate the usage of synthetic data instead of the real data for machine learning tasks (Hittmeir, Ekelhart, and Mayer 2019; Rashidian et al. 2020; Mendikowski et al. 2023). There are use cases where data synthesis is used to balance a dataset when e.g. the labels for a prediction task are unbalanced (Wang et al. 2021; Son et al. 2023). This turned out to be a useful data augmentation strategy, see, e.g., (Hoffmann et al. 2019) for a case study for crumpled sheets, (Perets and Rappoport 2023) for applications with ensemble models and (Sutojo et al. 2020) for general considerations. Other examples for the use of data synthesis are digital twins in sleep research (Kumi et al. 2023), simulation of chronic coronary disorders (Koloi et al. 2023). The focus of this paper is not to do a classical balance of labels in the dataset but to generate more training examples for a specific individual cohort. A related study has been performed by on CT-images (Prasanna Das et al. 2022), where the authors suggest to transfer the idea to other types datasets.

The promising results in related problem field show that it is a reasonable task to apply these techniques in the context of EHR data and LoS-prediction. The prediction of LoS for the MIMIC-dataset has been done, e.g., by (Hasan et al. 2023) for MIMIC-III and by (Winter, Hartwig, and Kirsten 2023) for MIMIC-IV, however, always on big datasets. A LoS-prediction which accounts for imbalanced data was presented by (Alsinglawi et al. 2022), however, does not use a CTGAN and is restricted to patients with lung cancer. As shown by (Hartwig et al. 2024), the predicted LoS can be further used, for example, to predict bed occupancy in a hospital.

Dataset

In our experiments, we extract data from the MIMIC-IV dataset (Johnson et al. 2023). MIMIC-IV is a freely available EHR dataset collected from the Beth Israel Deaconess Medical Center located in Boston. The collection of the data spans almost a decade between the years 2008 and 2019. More than $180k$ different patients were admitted to the hospital resulting in over $400k$ unique admissions. All patients where deidentified according to the HIPAA (Health Insurance Portability and Accountability Act) standard. For instance, dates are shifted for each patient consistently by a fixed offset such that the exact date of birth can not be reconstructed. However, by subtracting the admission time from the date of birth, one still gets the age of a patient. We extract the same features from the MIMIC-IV dataset as described by Winter et al. in (Winter, Hartwig, and Kirsten 2023). The total LoS of a patient in the hospital is predicted for patients leaving the emergency department (ED), therefore, for prediction only information given at that point in time is used.

LoS The length of stay (LoS) of a patient spans the time between the patient leaving the ED and leaving finally the hospital. The date of the admission is extracted from the ED module and the date of discharge from the *hosp* module of MIMIC-IV.

Admission location The location of the patient before being admitted to the hospital.

ICD-code Each patients admission is associated with a primary diagnosis encoded either with an ICD-9 or ICD-10 code. Half of the diagnosis are encoded with ICD-9 and the other half with ICD-10.

Age Each patient has an `anchor_year` and an `anchor_age` attribute in the dataset. For instance, if the patient is admitted to the hospital in the year 2060, has the anchor year 2040, and the anchor age 20, then the patient is 20 years old in the year 2040 and 40 years old at time of hospitalisation.

Insurance The insurance a patient has at time of admission. In 9% of all cases, a patient has medicaid, in 38% has medicare and in all other 53% it is not further specified,

Ethnicity The ethnicity of a patient.

Gender A patient is either female or male in the dataset.

Resprate The respiratory rate of a patient during the triage right before the patient is admitted to the hospital. Values are normal distributed and have a mean of 17.77 breaths per minute and standard derivation of 2.52.

sbp The sbp (systolic blood pressure) is the measurement of the blood pressure.

Pain All patients are asked during the triage right before being admitted to the ED how much pain they have between 0 and 10.

Diagnosis count A count of the total number of diagnosis ICD-9 or ICD-10 codes associated with the patients admission.

Medication count The total count of all medications a patient enters during the triage.

ED LoS Time spend by a patient in the ED is extracted from the dataset as the ED LoS.

Average LoS of previous admissions The average LoS of a patient at previous admissions at the hospital. In the case of 52%, the patient was admitted to the hospital and for all other patients who are admitted to the hospital for the first time, we set the value to 0.

After extracting features from the MIMIC-IV dataset, approximately half of the admissions are filtered out, as not all patients are admitted to the ED at all. Additionally, we classify all admissions having a LoS longer than 50 days as outliers, resulting in a LoS distribution, as depicted in Figure 2. As one can see, the distribution has a high skew and only a few patients have a LoS higher than 20.

Methodology

Primary diagnoses, encoded as either an ICD-9 or ICD-10 code for each patients admission are used with varying frequency, as some diseases are more frequent than others, as presented in Figure 1. One can see in the figure that out of all 5660 codes, only 29 have a frequency higher than 1000, 258 a frequency between 10 and 100, and that most ICD-codes have a frequency lower than 10. Approximately half of the ICD-codes in our dataset are ICD-9-codes and the other half ICD-10-codes. The most and least common ICD-codes part of our dataset can be found in Table 1. The least common ICD-codes are chosen randomly, as 2080 patients have an ICD-code as the primary diagnose that occurs exactly once in our dataset.

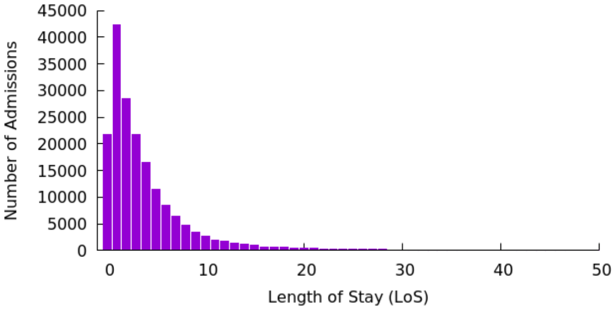

Figure 2: LoS distribution with a mean of 3.87 and standard derivation of 4.95 of patients admissions in our dataset after extracting features

ICD-code	#patients	Description
78650	6607	Unspecified chest pain
R079	4488	Other chest pain
R109	3886	Unspecified abdominal pain
F329	3273	Major Depressive Disorder
⋮	⋮	⋮
I519	1	Heart disease, unspecified
K3520	1	Acute appendicitis
A35	1	Other tetanus

Table 1: The most and least frequently occurring ICD-codes from our data set with their respective descriptions

In this paper, we examine how one could reduce the prediction error of the LoS of small cohorts of patients having the same ICD-code. As for synthesizing data a sufficient variety of real data is necessary, we consider only cohorts with more than 30 patient admissions. For each cohort considered, we split our data extracted from the MIMIC-IV database, as depicted in Figure 3. On the left hand side of the figure, one can see all cohorts $Cohort_1, \dots, Cohort_{n-1}$ sorted in an arbitrary order except for $Cohort_n$ (the cohort considered) which is depicted on the right hand side of the

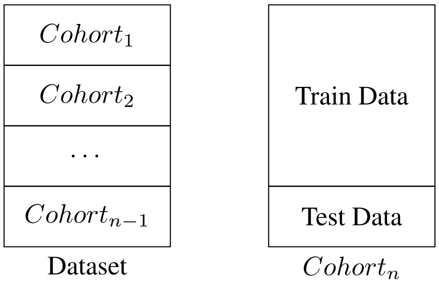

Figure 3: Dataset division for our models to be trained

figure. For $Cohort_n$, we aim to improve the prediction of the LoS of the patient admissions. It is split into a 70% train dataset for training and a 30% left out test dataset. In our evaluation, we create three different models, each trained with another training data collection. All models are trained using CatBoost (Prokhorenkova et al. 2018), as described by Winter et al. in (Winter, Hartwig, and Kirsten 2023), that we compare with each other. The test dataset is identical for each of the three models for an accurate and fair comparison. CatBoost is an open-source gradient boosting library for gradient boosting on decision trees that has the advantage of supporting, in addition to numerical data, categorical data as input. Categorical data such as the ICD-code of a patients admission do not need to be converted into a numerical feature prior training a model.

Using GridSearchCV from the scikit-learn library, we searched for the best parameters at where for each set of parameters the train dataset is split randomly into a train and evaluation dataset three times. Each of the three times, the model is trained with fixed parameters and evaluated with the evaluation dataset. The parameters we performed grid-search on are the learning rate, the maximum depth of a tree and L2 leaf regularization and all other parameters CatBoost offers are left by their default as can be found in the documentation[1]. The learning rate is used for reducing the gradient step. As CatBoost is based on trees, the parameter depth determines the maximum depth of the trees and the optimal depth ranges from 4 to 10. Any positive value for the $L2$ regularization coefficient of the cost function is allowed. The higher the value, the lower will be the value of the corresponding leaf. In the following, we present all three models. Specifically, we discuss which data we use to train them and what we expect from the results.

Model I: Training on All Cohorts

For training Model I, we use data from $Cohort_1$ to $Cohort_{n-1}$, and the training data part of $Cohort_n$. Only the test data of $Cohort_n$ is left out to measure the accuracy of the trained model. This model is only realistic if the performed study on a cohort can also leverage data from other studies. We include the model as a benchmark for further comparisons.

Model II: Training on Cohort Data

For training the second model, we use only the training data of $Cohort_n$ and evaluate it on the test data. This training regime often occurs in the real-world when a study is performed on a selected subset of patients without additional generated data. If $Cohort_n$ is small, then only a few training samples are given. If the model that is trained only on an individual cohort of patients would outperform the first model trained on all cohorts, then one can use it in combination with the other individual models trained for cohorts as ensemble method to improve the overall accuracy of predicting the LoS of the patients (This is of course also possible for model III where each model of the ensemble then contains additional synthesized data).

Model III: Training on Cohort and Synthetic Data

In addition to the real data of a patient cohort, the third model is trained on synthesized data. Data is synthesized from fitting a CTGAN on $Cohort_1$ to $Cohort_{n-1}$ and on the training data part of $Cohort_n$ to synthesize data belonging to $Cohort_n$.

We synthesize the data using conditional generative adversarial network (Xu et al. 2019) which is designed for fitting a model on a single table. The fitted model then is able to synthesize realistic data that is similar to the data the model was fitted on. Naive approaches which, e.g., duplicate rows of the data while preserving the data distribution do not protect the privacy of individuals in the data and do mostly not help to improve the training of any model. In contrast, CTGAN fitted on real data, synthesizes synthetic data not equal but similar to the real data. The SDV (synthetic data vault) library[2] is equipped with an API for checking the quality of synthesized data. Two metrics, namely the shapes of synthesized columns and column pair trends are summarized as an overall score, such that the library returns three scores. An overall score of 100% means that the synthesized data is identical to the real data and a score of 0% that synthesized data contains the opposite from the real data. Hence, ideally the score is never either close to 100% or to 0% as the data is then not useful for enriching real data with synthesized data for training a model. According to the documentation of SDV, column shapes is the marginal distribution of the pairwise comparison between a real and the corresponding synthetic column. The second measure, column pair trends, is the correlation between pairs of columns. If the overall score is around 80%, then the cohort can be enriched with synthetic data, that is close to the real data, but not equal. In that case, there is a chance that training the model with augmented data improves the result.

As already mentioned, the synthesizer is fitted on $Cohort_1$ to $Cohort_{n-1}$ and the training data of $Cohort_n$. However, we only want to enrich the training data of $Cohort_n$ only with synthetic data from the same cohort, i.e., having the same ICD-code. For that case, SDV has the ability to condition on a specific ICD-code during sampling synthetic data. Currently, only rejection sampling is supported when using the CTGAN synthesizer, as we do. Hence, synthesization of very small cohorts of approximately 30 patient admissions or less took too much time and was therefore omitted in our evaluation, as it is not promising due to missing variety in the data. For each cohort, we synthesized 4000 patient admissions and added them to the real cohort data.

Evaluation

In the evaluation, we selected different patient cohorts depending on different statistics in order to cover as many different cohorts as possible as testing all cohorts was not possible due to its computational requirements. First of all, we

[1] https://catboost.ai/en/docs/references/training-parameters/

[2] https://docs.sdv.dev/sdv/single-table-data/evaluation/data-quality

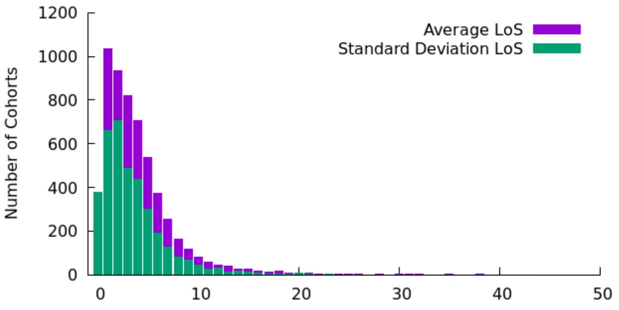

Figure 4: Average and standard derivation distribution of all cohorts in our dataset

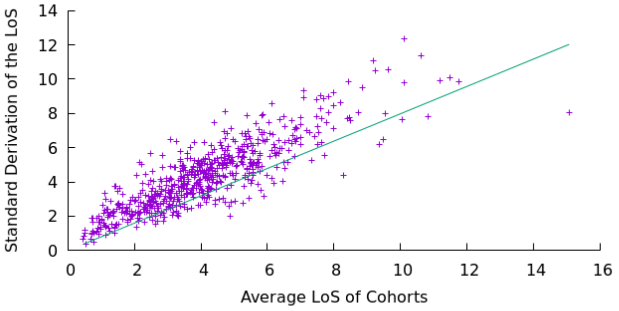

Figure 5: Correlation of 0.798 between the average LoS and its standard derivation for cohorts with more than 30 admissions

chose cohorts based on their average and standard derivation of their LoS. The distribution is depicted in Figure 4. Both the average and standard derivation of the patient cohorts LoS have a high positive skew. If one takes all cohorts with more than 30 patients admissions into account, then the average LoS and its standard derivation are correlated with each other as depicted in Figure 5. Hence we chose large, medium, and small cohorts each with patient admissions having a long and short LoS mean and standard derivation respectively. Another statistic we take into account is the skew of the LoS distribution of each individual cohort. The distribution of the skew is depicted in Figure 6 with an average of 0.89 and standard derivation of 0.24. Hence, we also chose cohorts with different skews.

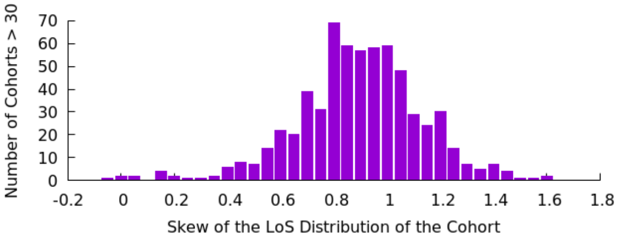

Figure 6: Distribution of the skew of all cohorts with more than 30 patients admissions

The final results of our evaluation can be found in Table 2. The first column of Table 2 is the cohort we are analysing, i.e., all patient admissions having the same *ICD-code* as the primary diagnose at when the patient enters the ED of the hospital. The columns *All*, *Cohort* and *Synthetic* correspond the accuracy in tearms of mean absolute error of Models I, II and III respectively. *A* - *C* is the difference between *All* and *Cohort* and *A* - *S* the difference between *All* and *Synthetic*. The table is sorted in descending order by *A* - *S*, at where the most upper rows represent the best results we got from adding synthesized data to real cohort data for training. The other columns are statistics about the cohort under consideration. *Quality* is the overall quality of the synthesized data returned by the SDV library. *Avg(L)* is the average, *Stddev(L)* the standard derivation, and *Skew(L)* the skew of the LoS of the corresponding patient admissions of the cohort. Finally, *#* denotes the number of patient admissions of the cohort.

Discussion

We tested 26 patient cohorts in our evaluation, as listed in Table 2. In roughly two thirds of the cases, adding synthetic data to patient cohorts prior training has led to an improvement in the accuracy of the trained model in contrast to a model trained only on the cohort's data. The same ratio holds true when we compare the accuracy to a model trained on all data. Adding synthetic data is particularly effective when the average LoS and the standard derivation are large while the cohort itself is rather small. All others cases, separated by a horizontal line at the table, the accuracy of the model is worse than the one of the model trained on all cohorts.

As a rule of thumb, if a cohort is very large, then the models accuracy differs only moderately. If a cohort is small, the average of the LoS and its standard derivation is high, then one could consider adding synthetic data to the cohort before training a model.

The LoS distribution in the synthesized data has a high effect on whether the model that is trained on both real and synthetic data performs better than a model trained on real cohort data alone, as depicted in Figure 7. On the bottom right of Figure 7, the LoS distribution of all cohorts is depicted. The distribution has a high positive skew and most patient admissions have a length of roughly one day. On the bottom left one can see the LoS distribution of both the real and synthetic data of the patients cohort with the primary diagnose 78097. Both distributions are not identical, however mostly similar in contrast to the others. That is reflected by the results from Table 2. Adding synthetic data to the cohort 78097 has almost no effect, as model *All* has an mean absolute error of 2.66, model *Cohort* of 2.55, and model *Synthetic* of 2.63. At the top right of Figure 7, the distribution of the synthetic data of cohort with the ICD-code 0389 is almost identical to the distribution of the real data at the bottom right of the image. This is also reflected in the performance of the models. Adding synthetic data to the patient cohort improves our baseline model *All* by 0.44. As a counter example, the LoS distribution of the synthetic data on the top left of cohort with the ICD-code 95901 is dissimilar to the real LoS distribution of all cohorts. Especially, the

ICD-Code	Cohort	A - C	Synthetic	A - S	All	Quality	Avg(L)	Stddev(L)	#	Skew(L)
K652	7.66	0.08	7.07	0.67	7.74	66.32%	11.19	9.91	49	1.58
0389	4.54	−0.10	4.00	0.44	4.44	67.19%	7.62	6.30	451	0.80
28800	4.27	−0.01	3.87	0.39	4.26	71.02%	6.64	6.45	225	0.91
28419	9.58	−1.48	7.82	0.28	8.10	66.74%	8.45	9.84	35	1.12
D708	6.81	−1.30	5.28	0.23	5.51	68.14%	8.87	9.52	106	1.06
4589	4.00	0.23	4.00	0.23	4.23	68.48%	5.14	5.74	286	1.08
J9601	4.73	0.63	5.14	0.22	5.36	69.81%	7.99	7.10	52	1.00
8080	1.70	−0.37	1.20	0.13	1.33	67.92%	4.44	2.67	59	0.01
R188	3.51	−0.09	3.31	0.11	3.42	69.19%	6.43	5.94	178	1.17
R410	4.51	−0.62	3.80	0.09	3.89	68.73%	5.87	6.41	252	0.92
A419	5.40	−0.31	5.02	0.07	5.09	65.01%	8.42	7.70	912	0.97
3489	3.69	0.02	3.65	0.06	3.71	62.97%	7.54	6.90	117	0.45
N179	4.37	0.03	4.34	0.06	4.40	67.15%	6.21	6.17	1483	1.02
78097	2.55	0.11	2.63	0.03	2.66	65.22%	3.47	4.99	2329	1.03
J90	4.34	−0.26	4.08	0.00	4.08	68.19%	7.50	7.23	332	1.02
I6201	4.33	0.22	4.56	−0.01	4.55	66.26%	6.90	7.10	123	0.97
I509	3.93	−0.11	3.83	−0.01	3.82	60.32%	7.53	6.19	1946	0.79
R110	1.83	0.13	2.04	−0.08	1.96	70.23%	3.06	3.09	151	0.42
79902	3.94	0.13	4.17	−0.10	4.07	69.60%	4.98	6.45	126	0.93
R17	5.41	−0.38	5.15	−0.12	5.03	69.32%	6.73	7.58	226	1.02
M25562	2.25	0.03	2.59	−0.31	2.28	64.85%	3.25	4.86	115	1.28
430	6.58	−0.21	6.71	−0.34	6.37	65.89%	9.55	7.99	110	0.60
G939	3.18	−0.44	3.16	−0.42	2.74	67.33%	7.73	5.57	96	0.56
8082	1.47	0.08	2.06	−0.51	1.55	69.07%	2.90	1.92	80	0.22
71946	1.52	0.07	2.39	−0.80	1.59	65.82%	2.42	3.14	162	1.38
95901	1.12	0.26	2.82	−1.44	1.38	66.61%	1.73	2.30	341	1.26

Table 2: Overview of the results of our evaluation

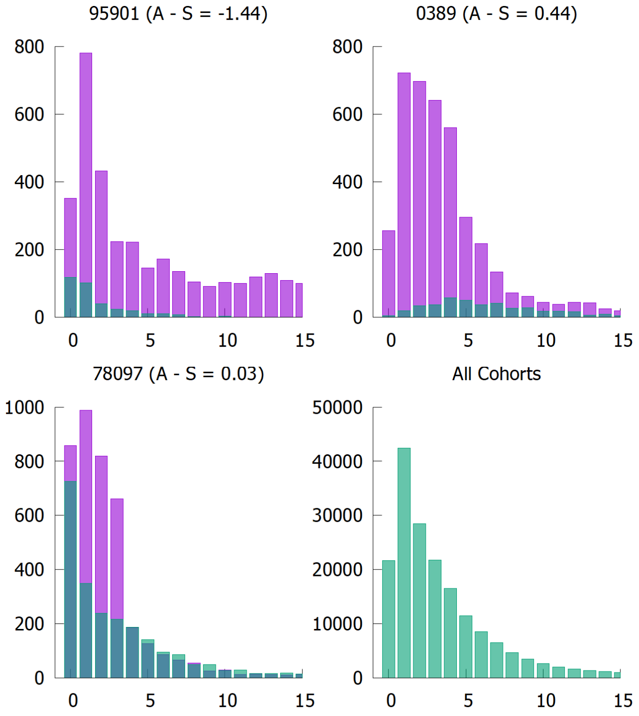

Figure 7: LoS distribution of different patient cohorts with purple representing the real and green synthetic data.

synthetic data contains too many patient admissions with a LoS longer than 5 days.

Conclusion

The considerations in this paper show that enriching patient data with snythesized data for LoS-prediction is promising. However, it turned out that it is not a miracle cure for solving the problem of data scarcity for this topic and thus opens up a wide range of topics based on synthesizing data for LoS-prediciton that can be considered in detail. Although, the discussion shows that there are some rules determinable to discriminate between advantageous and non-advantageous synthetizations, it is necessary to examine that in detail, e.g., by enhancing the study to a greater variety of ICD-codes, to different cohorts than the ones based on ICD-codes or to other datasets. Additionally, as mentioned in the discussion, CT-GAN suffers from some drawbacks regarding EHRs. Thus, it would be worth examining whether it is possible to circumvent some of the limitations of the models by using a different, case-specific approach for data synthesis.

Acknowledgements

The research for the paper was funded by the state of Schleswig-Holstein as part of the APONA project, project no. 220 23 020.

References

[Alsinglawi et al. 2022] Alsinglawi, B.; Alshari, O.; Alorjani, M.; Mubin, O.; Alnajjar, F.; Novoa, M.; and Darwish, O. 2022. An explainable machine learning framework for lung cancer hospital length of stay prediction. *Scientific Reports* 12(1).

[Bourou et al. 2021] Bourou, S.; El Saer, A.; Velivassaki, T.-H.; Voulkidis, A.; and Zahariadis, T. 2021. A review of tabular data synthesis using GANs on an IDS dataset. *Information* 12(09):375.

[Chawla et al. 2002] Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16:321–357.

[Fang, Dhami, and Kersting 2022] Fang, M. L.; Dhami, D. S.; and Kersting, K. 2022. *DP-CTGAN: Differentially Private Medical Data Generation Using CTGANs*. Springer International Publishing. 178–188.

[Hartwig et al. 2024] Hartwig, M.; Schiff, S.; Wolfrum, S.; and Möller, R. 2024. Aggregating predicted individual hospital length of stay to predict bed occupancy for hospitals. In *Proceedings of the 17th International Joint Conference on Biomedical Engineering Systems and Technologies - HEALTHINF*, 175–184. INSTICC.

[Hasan et al. 2023] Hasan, M. N.; Hamdan, S.; Poudel, S.; Vargas, J.; and Poudel, K. 2023. Prediction of length-of-stay at intensive care unit (ICU) using machine learning based on MIMIC-III database. In *2023 IEEE Conference on Artificial Intelligence (CAI)*. IEEE.

[Hittmeir, Ekelhart, and Mayer 2019] Hittmeir, M.; Ekelhart, A.; and Mayer, R. 2019. On the utility of synthetic data: An empirical evaluation on machine learning tasks. In *Proceedings of the 14th International Conference on Availability, Reliability and Security*, ARES '19, 1–6. ACM.

[Hoffmann et al. 2019] Hoffmann, J.; Bar-Sinai, Y.; Lee, L. M.; Andrejevic, J.; Mishra, S.; Rubinstein, S. M.; and Rycroft, C. H. 2019. Machine learning in a data-limited regime: Augmenting experiments with synthetic data uncovers order in crumpled sheets. *Science Advances* 5(4).

[Johnson et al. 2023] Johnson, A. E. W.; Bulgarelli, L.; Shen, L.; Gayles, A.; Shammout, A.; Horng, S.; Pollard, T. J.; Hao, S.; Moody, B.; Gow, B.; Lehman, L.-w. H.; Celi, L. A.; and Mark, R. G. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data* 10(1).

[Jordon et al. 2022a] Jordon, J.; Szpruch, L.; Houssiau, F.; Bottarelli, M.; Cherubin, G.; Maple, C.; Cohen, S. N.; and Weller, A. 2022a. Synthetic data – what, why and how? *arXiv*.

[Jordon et al. 2022b] Jordon, J.; Szpruch, L.; Houssiau, F.; Bottarelli, M.; Cherubin, G.; Maple, C.; Cohen, S. N.; and Weller, A. 2022b. Synthetic data–what, why and how? *arXiv preprint arXiv:2205.03257*.

[Koloi et al. 2023] Koloi, A.; Loukas, V. S.; Sakellarios, A.; Bosch, J. A.; Quax, R.; Nowakowska, K.; Tachos, N.; Kaźmierski, J.; Papaloukas, C.; and Fotiadis, D. 2023. A comparison study on creating simulated patient data for individuals suffering from chronic coronary disorders. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE.

[Kuchin, Mukhamediev, and Yakunin 2020] Kuchin, Y. I.; Mukhamediev, R. I.; and Yakunin, K. O. 2020. One method of generating synthetic data to assess the upper limit of machine learning algorithms performance. *Cogent Engineering* 7(1):1718821.

[Kumi et al. 2023] Kumi, S.; Hilton, M.; Snow, C.; Lomotey, R. K.; and Deters, R. 2023. SleepSynth: Evaluating the use of synthetic data in health digital twins. In *2023 IEEE International Conference on Digital Health (ICDH)*. IEEE.

[Kuo and others 2020] Kuo, K., et al. 2020. Generative synthesis of insurance datasets. Technical report.

[Mendikowski and Hartwig 2022] Mendikowski, M., and Hartwig, M. 2022. Creating customers that never existed: Synthesis of e-commerce data using CTGAN. In *18th International Conference on Machine Learning and Data Mining (MLDM-22). New York, US: IBAI Publishing*, 91–105.

[Mendikowski et al. 2023] Mendikowski, M.; Schindler, B.; Schmid, T.; Möller, R.; and Hartwig, M. 2023. Improved techniques for training tabular GANs using cramer's v statistics. *Proceedings of the Canadian Conference on Artificial Intelligence*.

[Perets and Rappoport 2023] Perets, O., and Rappoport, N. 2023. Ensemble synthetic EHR generation for increasing subpopulation model's performance. *arXiv*.

[Prasanna Das et al. 2022] Prasanna Das, H.; Tran, R.; Singh, J.; Yue, X.; Tison, G.; Sangiovanni-Vincentelli, A.; and Spanos, C. J. 2022. Conditional synthetic data generation for robust machine learning applications with limited pandemic data. *Proceedings of the AAAI Conference on Artificial Intelligence* 36(11):11792–11800.

[Prokhorenkova et al. 2018] Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A. V.; and Gulin, A. 2018. CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems* 31.

[Rashidian et al. 2020] Rashidian, S.; Wang, F.; Moffitt, R.; Garcia, V.; Dutt, A.; Chang, W.; Pandya, V.; Hajagos, J.; Saltz, M.; and Saltz, J. 2020. *SMOOTH-GAN: Towards Sharp and Smooth Synthetic EHR Data Generation*. Springer International Publishing. 37–48.

[Sauber-Cole and Khoshgoftaar 2022] Sauber-Cole, R., and Khoshgoftaar, T. M. 2022. The use of generative adversarial networks to alleviate class imbalance in tabular data: a survey. *Journal of Big Data* 9(1).

[Son et al. 2023] Son, B.; Myung, J.; Shin, Y.; Kim, S.; Kim, S. H.; Chung, J.-M.; Noh, J.; Cho, J.; and Chung, H. S. 2023. Improved patient mortality predictions in emergency departments with deep learning data-synthesis and ensemble models. *Scientific Reports* 13(1).

[Sutojo et al. 2020] Sutojo, T.; Syukur, A.; Rustad, S.; Fajar Shidik, G.; Agus Santoso, H.; Purwanto, P.; and Muljono, M. 2020. Investigating the impact of synthetic data distribution on the performance of regression models to overcome small dataset problems. In *2020 International Seminar on Application for Technology of Information and Communication (iSemantic)*. IEEE.

[Torfi, Fox, and Reddy 2022] Torfi, A.; Fox, E. A.; and Reddy, C. K. 2022. Differentially private synthetic medical data generation using convolutional GANs. *Information Sciences* 586:485–500.

[Walonoski et al. 2017] Walonoski, J.; Kramer, M.; Nichols, J.; Quina, A.; Moesel, C.; Hall, D.; Duffett, C.; Dube, K.; Gallagher, T.; and McLachlan, S. 2017. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association* 25(3):230–238.

[Wang et al. 2021] Wang, L.; Han, M.; Li, X.; Zhang, N.; and Cheng, H. 2021. Review of classification methods on unbalanced data sets. *IEEE Access* 9:64606–64628.

[Winter, Hartwig, and Kirsten 2023] Winter, A.; Hartwig, M.; and Kirsten, T. 2023. Predicting hospital length of stay of patients leaving the emergency. In *International Conference on Health Informatics*.

[Xu et al. 2019] Xu, L.; Skoularidou, M.; Cuesta-Infante, A.; and Veeramachaneni, K. 2019. Modeling tabular data using conditional GAN. In *Advances in Neural Information Processing Systems*.