Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)

doi:10.3233/FAIA230681

Assessing VTE Risk in Cancer Patients Using Deep Learning Synthetic Data Generation and Domain Adaptation Techniques

Sergi BECH [a], Bárbara LOBATO [b], Oriol PUJOL [a], José Manuel SORIA [b], Andrés MUÑOZ [c]

[a] *Dept. Matemàtiques i Informàtica, Universitat de Barcelona, Barcelona, Spain*
[b] *Unitat de Genòmica de Malalties Complexes, Institut de Recerca de l'Hospital de la Santa Creu i Sant Pau, Barcelona, Spain*
[c] *Medical Oncology, Hospital General Universitario Gregorio Marañón; and Cancer and Thrombosis Working Section, Spanish Society of Medical Oncology (SEOM), Madrid, Spain*

Abstract. This article focuses on the use of deep learning synthetic data generation methods to assess the risk of future treatments and medication for preventing venous thromboembolism (VTE) in cancer patients, based on a small dataset of genetic and clinical variables. The study employs CopulaGANs to generate synthetic tabular data, which is then used to train a Deep Learning-based classifier using domain adaptation techniques. The trained model is fine-tuned using real data and performs better than current state-of-the-art medical scores in assessing VTE risk. Additionally, the resulting Precision-Recall curve offers flexibility in selecting different and better operational points for VTE risk assessment.

Keywords. Synthetic Data Generation, Generative Adversarial Networks, Domain Adaptation, Application to Biomedical Data

1. Introduction

Cancer is one of the most common and fatal diseases nowadays. In 2020, over 19 million people were newly diagnosed and 10 million died worldwide [1]. Thanks to recent advances in oncologic treatments, many types of cancers are now regarded as chronic illnesses and, even in some cases, curable [2]. Cancer patients are in greater risk of suffering venous thromboembolism (VTE) when compared to the general population [3]. Although VTE can have dire effects, it can be easily prevented with anticoagulants. In fact, more effective anticoagulants are now accessible [4]. However, cancer patients also are in greater risk of bleeding [5]. Thus, the clinical management of such patients is difficult due to the delicate balance between the benefit of receiving anticoagulants to prevent VTE and the risk of severe bleeding. Besides, the longer the life expectancy in chronic cancer patients, the more likely it is that they will suffer VTE at some point [6].

The first attempt to create a tool able to classify cancer patients by VTE risk prior to anti-cancer treatment is the Khorana Risk Score (KRS) [7], which utilises five clinical features and assigns a score depending on their observation. Patients are then classified into low, intermediate or high risk. Despite its modest performance, the KRS has become the reference score and many of the current scores in clinical practice are modified versions of this.

With the increase in accessibility to genetic analysis and the evidences that support the role of genetics in VTE [8], [9], new scores based on combining genetic variables and phenotypic information have been developed. The first of such scores was TiC-Onco score [10]. This score has been recently replaced by ONCOTHROMB score [11], which establishes the current state-of-the-art and outperforms KRS when access to genetic information is available.

Due to the costs and potential impact of these studies, the creation and validation of these scores usually entails small datasets. For example, the validation of the ONCOTHROMB score only uses a few hundreds of patients. Although using the correct methodologies can lead to models with a certain degree of generalization, the small amount of samples combined with the large dimensionality (imposed by the combination of phenotypic and genetic features) certainly hinders predictive models from achieving the highest performance.

The machine learning community has recognized the problems of small data combined with the curse of dimensionality. Dimensionality reduction and feature selection techniques have been widely used to solve this issue. With the advent of deep learning techniques and the increased need of data for training complex models, there has been a resurgence of methodologies focused in the creation of new samples. Overcoming some of the problems of noise-based models and interpolation methods such as SMOTE, synthetic data generation constitutes many of the current efforts of the community. In this trend, one can find different approaches depending on the specific application, ranging from variational autoencoders [12], generative adversarial networks [13] or diffusion models [14], among others.

In this work we explore the creation of tabular synthetic data compatible with features in ONCOTHROMB score with the goal of inducing a learning bias in such model. To that end, a variant of Conditional Tabular GAN — a CopulaGAN — is used to generate a large dataset of compatible data with the real ONCOTHROMB cohort. Then, using transfer learning techniques [15], a pretrained model with synthetic data is fine-tuned to the real data. The results show that the obtained model improves not only the generalization of the score but also the range of operational points in the Precision-Recall curve. This is of particular interest, as it allows to improve the evaluation of the two risks involved, namely complications due to VTE and those due to anticoagulant treatment.

The article is organized as follows: section 2 introduces the relevant concepts and techniques that will be referred to throughout this article. Section 3 briefly describes the proposed approach. Section 4 describes and report the obtained results and discusses them. Finally, conclusions are drawn and future lines are depicted.

2. Background and related works

2.1. Assessing VTE risk using ONCOTHROMB score

The ONCOTHROMB score was created using both genetic and clinical features with the aim to predict whether a patient diagnosed with cancer will develop VTE within the next six months since the diagnosis. The cohort of patients used to train this classifier is described in [10]. The predictive features used to train a logistic regression are: VTE risk according to tumor type as defined in the KRS, stage, whether the body mass index (BMI) is greater than 25 Kg/m2, and a Genetic Risk Score (GRS) that condenses the number of risk alleles of 9 Single Nucleotide Polymorphisms (SNPs) know to be associated with VTE. The genotyping and data collection procedures were performed using the protocols established in the Spanish ONCOTHROMB 12-01 study [11].

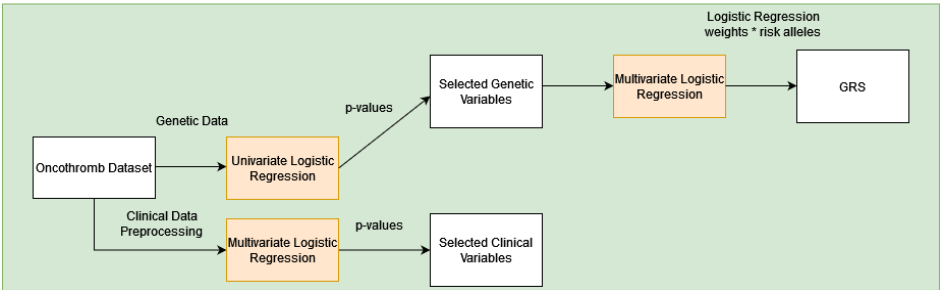


Figure 1. ONCOTHROMB score preprocessing pipeline

The steps in which the score was built are described in [11] and showed in Figure 1. The selection of the genetic variables is done using an univariate logistic regression, and variables with a p-value < 0.25 were chosen. Next, a multivariate logistic regression was performed using the number of risk alleles for the selected genetic variant. The weight of each genetic variant was multiplied by the number of risk alleles, thus generating a GRS for each patient. On the other hand, a multivariate logistic regression was also performed with the aim of selecting the clinical variables, and those with p-value ≤ 0.25 were kept. If any missing values were found for the selected variables for a patient, that patient was excluded from the analysis.

2.2. Deep Generative Models for tabular data

Deep generative models use deep learning techniques to generate synthetic data as similar as possible to real-world data. One of the most common types of deep generative models used for tabular data is the Generative Adversarial Network (GAN) [16]. These models were first developed to generate synthetic images, but several types of GANs were specifically developed for tabular data — some examples include the TabGAN, Conditional Tabular GAN (CTGAN) [17], Tabular Variational Autoencoder-GAN (TVAE-GAN), and CopulaGAN [13]. Firstly, TabGAN models generate synthetic tabular data by learning the underlying distribution of the original data using an adversarial loss. Secondly, CTGAN generates synthetic tabular data based on a given set of conditional vari-

ables. Thirdly, TVAE-GAN is a hybrid model that combines a variational autoencoder and a GAN; the VAE is used to encode the input data into a latent representation, and the GAN is used to generate new samples from the learned latent space. Finally, CopulaGAN is an extension of the CTGANs that utilizes copulas (a function that represents the correlation or dependence between variables with independence of their marginal distributions) to model the dependence structure of real-world data, and then generates synthetic data that preserves this same dependence structure using a GAN.

More specifically, the fitting process of a CopulaGAN begins by standardizing non-categorical variables using Gaussian normalization. In order to generate synthetic data, the CopulaGAN samples from the learned joint distribution using CTGAN, and the resulting data is transformed back to its original distribution by applying the inverse cumulative distribution function, ensuring that the generated synthetic data has the same marginal distributions as the original input data. The code for CopulaGAN and GaussianNormalizer can be found in the sdv library [18].

3. Transferring knowledge from synthetic data to the ONCOTHROMB cohort

In the ONCOTHROMB cohort, each patient contains clinical and genetic information that can be distilled into a small set of aggregated variables. The main hypothesis of this work is that, by means of deep learning generative methods, a large set of compatible data can be synthesized and used to induce a learning bias that can be transferred to the real setting, thus improving the generalization and operational points of the original ONCOTHROMB score.

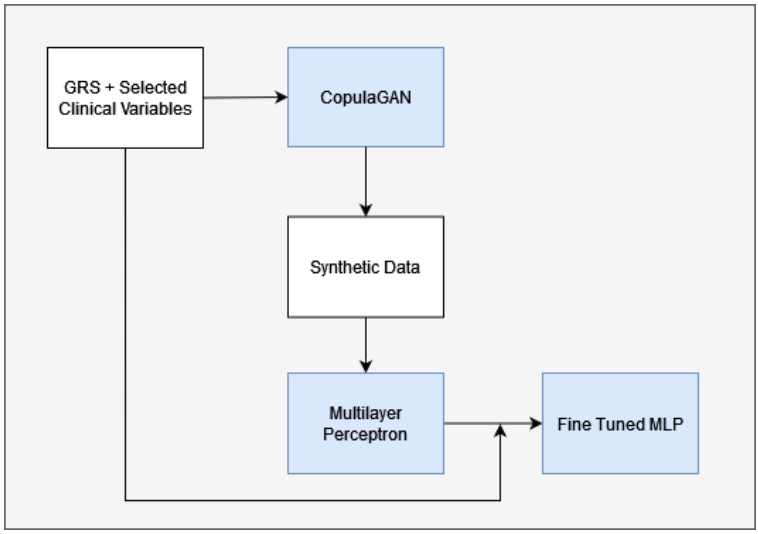


Figure 2. Methodology proposal

For that purpose, we first generate synthetic tabular data utilizing generative models. It is crucial to ensure that the synthetic data closely resembles the distribution of real data and, as a result, performs well in a classification task. Therefore, we propose to

train a multilayer perceptron (MLP) only with synthetic data and assess its performance. Lastly, a fine-tuning process of the MLP is carried out using real-world data. Figure 2 summarizes the proposed methodology.

Synthetic Data Generation: The first step of our proposal is to generate synthetic data samples using the CopulaGAN. As shown in Figure 2, we generate synthetic samples containing the GRS and the clinical variables that are selected on [11]. Then, the CopulaGAN is trained and 150,000 observations of synthetic data are sampled. The quality of the synthetic data is evaluated both visually and by training classifiers for the prediction of VTE.

Fine-tuning with Real Data: The synthetic data generated by the CopulaGAN is used to train a MLP. Then, we proceed to fine-tune the model with real data in order to shift the biases of the classifier. We experiment with various fine-tuning strategies, such as adding more layers to the pretrained model and freezing all layers except for these new ones, as well as retraining all layers for a small number of epochs and several learning rates. Both strategies produced indistinguishable results. For the sake of simplicity, in this article we report the results using fine-tuning with reduced learning rates.

4. Experiments and Results

In this section we explain the experimental setting, describing the details of the data, methods and models used, parameter set-up, and performance measures as well as the results obtained. The results are divided in three different subsections: first, the plausibility of the generated synthetic data is assessed; then, the results of the model trained on synthetic data are reported and discussed; and, finally, the transferred model is evaluated.

Data description: As already mentioned, the dataset used in this study comes from the ONCOTHROMB12-01 cohort, and it comprises clinical and genetic information on 390 patients with 19 clinical features plus 54 genetic features representing the number of risk alleles present in a set of genetic variants known to be linked to VTE. There are some missing values in the dataset, primarily in the genetic data. No imputation is used to deal with the missing values; instead, patients with missing values in key variables are dropped. Thus, 29 instances are dropped and the information of 361 patients is used in subsequent experiments. The same four features used in the original ONCOTHROMB (mentioned in Subsection 2.1) are the ones used in further analyses.

Methods and models used: The first step in this study is to generate synthetic tabular data from the preprocessed dataset. A CopulaGAN model is used to generate 150,000 samples of synthetic data. The CopulaGAN is trained with 5 discriminator updates for each generator update as in [19]; we also set the number of epochs to 1,500 and the batch size to 60, and the rest of parameters are the default ones from [18].

A MLP is trained with synthetic data only with the aim of learning the inductive biases of our problem. The MLP has 3 hidden layers with 128, 64, and 16 neurons, using the Rectified Linear Unit (ReLU) activation function with dropout rates of 0.2 and 0.1 after the first and second layers, respectively. The model was trained using binary cross-entropy as the loss function with Adam optimizer, a learning rate of 1×10^{-3}, and a sigmoid activation function on the output layer. Early stopping was applied by monitoring the validation precision-recall. Then, domain adaption techniques are applied to this pretrained deep learning model. Specifically, the model is fine-tuned using real

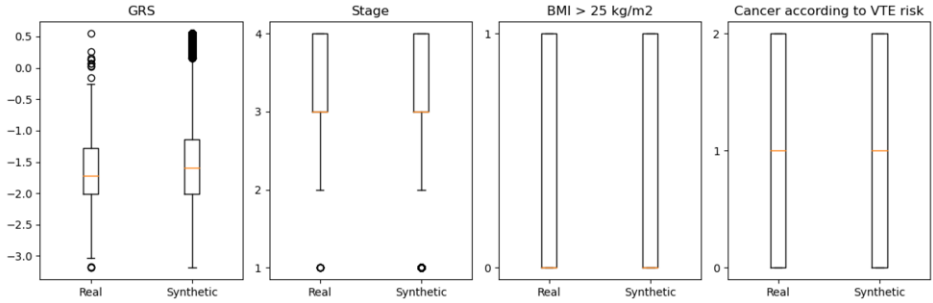


Figure 3. Comparison of real and synthetic data feature's distribution

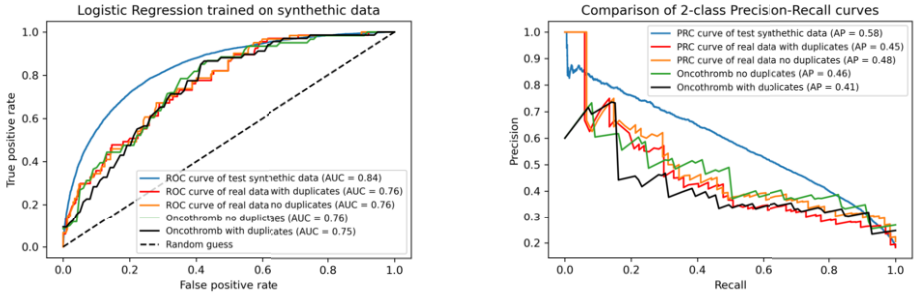
data to adapt and learn its specific patterns. In this step, the architecture of the model remains exactly the same and all layers are trained again, but this time with real data and for a smaller number of epochs and smaller learning rate (1×10^{-4}). To balance the class distribution of the real data, we computed class weights by dividing the total number of examples by twice the number of negative (or positive) examples. These class weights help to ensure that the model does not become biased towards the majority class during fine-tuning step. The early stopping criteria used during the previous training step is still used for this training as well.

Performance metrics: The evaluation of our method is done by comparing the ROC (Receiver Operating Characteristic) curve and PRC (Precision-Recall Curve) with the ONCOTHROMB score's curves. Additionally, we report the area under the curves, AUC-ROC and AUC-PRC, respectively.

We observed that, due to the variable selection process conducted in the ONCOTHROMB score, the original data set contains duplicated observations. This may undesirably bias the results. In order to provide a fair comparison, we evaluate all metrics in the dataset with duplicates (as done in the original paper) and also without duplicates.

Experimental set-up: The experiments are divided in three blocks:

- **Experiment 1: Synthetic data quality assessment.** The CopulaGAN is trained over all the preprocessed dataset. To assess the quality of the synthetic data, a graphical comparison is performed between the distributions of both synthetic and real data.
- **Experiment 2: Assessment of synthetic data trained classifiers.** Subsequently, the synthetic data was divided into training and test sets, and a logistic regression classifier was trained on this data. This logistic regression is evaluated using the synthetic fake data and also the real data in order to compare it to the baseline. The purpose of this evaluation was to verify that the logistic regression model trained with synthetic data produced similar results to that trained with real data. Afterwards, a MLP consisting of three layers was trained using the synthetic training data, and precision-recall was utilized as the stopping criteria. This model is evaluated with the test data and also with the real data to be compared with the baseline.
- **Experiment 3: Assessment of the transfer learning approach.** As a last step, the real data is divided into 5 stratified folds and is used to adapt the model to our domain by fine-tunning the MLP.



(a) (b)

Figure 4. Synthetic and real data evaluation on classification task. The blue curves represents the logistic regression model tested on test synthetic data. The green curves are the ONCOTHROMB model's and the yellow/red curves are from the logistic regression tested on the real data.

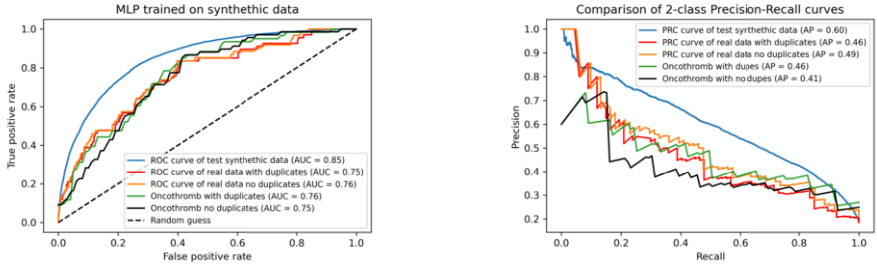
Experiment 1: Synthetic Data Qualitative Assessment

Synthethic data generated with CopulaGAN is first evaluated by comparing the cumulative sum distribution per feature (not shown) as well as the feature's data distribution represented in boxplots.

Figure 3 suggests that the synthetic data preserves the distribution and patterns of the original data despite the fact that synthetic data is generated in a non-constrained parameter space that does not necessarily resemble that of the real data. Moreover, we proceed to evaluate how good is this synthetic data compared to the real data when classifying patients by VTE risk.

Experiment 2: Assessment of synthetic data trained classifiers

The information captured in synthetic data is demonstrated to be useful for classifying real data through the performance of logistic regression trained on synthetic data (Figure 4). We observe that the differences between the model trained exclusively using synthetic data is able to achieve comparable results to the model trained using the original data. The ROC curve in Figure 4 (a) indicates that our synthetic data still contains valuable information. Similar conclusions can be drawn from Figure 4 (b), where syn-



(a) ROC curve (b) PRC curve

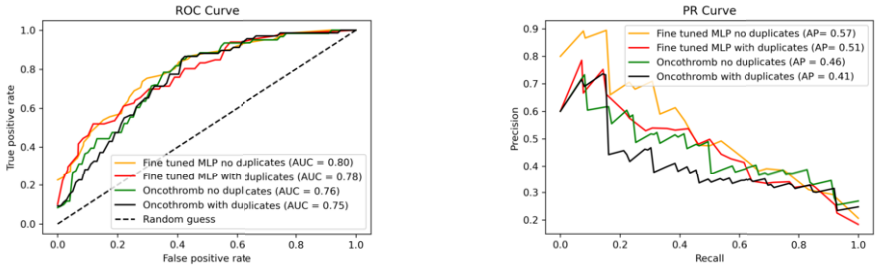
Figure 5. Synthetic data and real data evaluated on classification task. The blue curves represents the MLP model tested on test synthetic data. The green curves are the ones from ONCOTHROMB replications and the yellow/red curve are from the MLP tested on the real data.

thetic models are on pair with the models trained with real data in terms of PR curve. These results support our hypothesis that the generated synthetic data contains valuable information for making predictions.

Figure 5 shows the same kind of evaluation when training a MLP. Again, we observe in both the ROC curve, Figure 5(a), and PR curve in Figure 5(b), that the MLP classifier replicates the findings shown in the former experiment.

Experiment 3: Assessment of the transfer learning approach

As described in the experimental set-up, we report the results after using the transfer learning approach. Figure 6 shows the curves obtained after the fine tuning.



(a) ROC curve (b) PRC curve

Figure 6. Real data evaluated on classification task. The blue curves are the one from our ONCOTHROMB replications and the yellow/red curves are from the fine tunned MLP tested on the real data.

When looking at the ROC curves in Figure 6 (a) we observe a consistent improvement of the AUC-ROC values compared to the baseline in both settings; i.e. considering duplicates and non-duplicates. This is shown by the clear improvement in the TPR for the first half of the FPR in the ROC curves. However, for large values of FPR all methods seems to converge to the same behavior. A more drastic improvement is found in the Precision-Recall curves. Figure 6 (b) shows the obtained curves. A detailed examination of these segments according to the two different set-ups, namely 'duplicates' and 'non-duplicates', provides evidence that the proposed approach improves the AUC-PRC by 10%. All the relevant figures and its performance metrics are summarized in Table 1.

	ONCOTHROMB score	Trained with synthetic data	Fine-tuned with real-world data
Non-duplicates	0.46/0.76	0.49/0.76	**0.57/0.80**
Duplicates	0.41/0.75	0.46/0.75	**0.51/0.78**

Table 1. Results of AUC-PRC/AUC-ROC for the baseline model, the model trained exclusively with synthetic data, and the transfer learning approach.

The large improvement in AUC-PRC might have important consequences in the clinical practice. If we examine the plots in Figure 6 (b) we observe that the fine-tuned approaches clearly improve the operational point range of the score. In the case of considering the original dataset (curves in black and red), we observe that for recalls between 0.2 and 0.8, the proposed methodology achieves important gains. If we focus on the modified dataset (curves green and yellow), we observe the same effect in the range of recalls between 0.0 and 0.65. Again, the proposed method clearly outperforms the

baseline. As said earlier, this has might have a huge impact in clinical practice as it allows the physician to leverage the different risks involved in the process.

For the sake of discussion, the recall value stands for the rate of VTE detected from all the VTE population. Thus, the larger the recall, the best one can prevent VTE complications by administering anticoagulant treatment. However, as expected, this value trades-off with the precision. In this case, precision is related to the risk created by the administration of the anticoagulant treatment. The operational point of the original ONCOTHROMB article is $(0.8, 0.3)$. This is justified by the flat area in the black precision curve for the range of values between 0.4 and 0.8 recall. Precision of 0.3 means that the original score would recommend to overmedicate 70% of detected patients. The new score proposed in this article allows for the exploration and selection of different operational points, with a clear reduction of the risk of overmedication while keeping a large sensitivity value.

5. Conclusions

In this paper, we explored the use of synthetic data generation and transfer learning methodologies for improving the operational points in the PRC and ROC curves of a classifier that predicts the risk of VTE in cancer patients. Results show clear improvements in the AUC-ROC and AUC-PRC. These findings have impact in the clinical practice as it may allow to reduce the risk of complications due to the treatment of VTE.

The next steps in this line of work follow two main avenues. First, the obtained results are very promising but further validation is recommended. Thus, new data is being collected that may further confirm these improvements. Second, we plan on focusing on the influence of the genetic variants in VTE. In this line of thought, we plan on exploring the use of causal machine learning algorithms to better understand the underlying mechanisms of cancer in VTE.

6. Acknowledgements

This work is supported by the Sociedad Española de Trombosis y Hemostasia (SETH), Sociedad Española de Oncología Médica (SEOM), Asociación ActivaTT por la Salud and PI/20/00325 FIS project, and partially supported by MCIN/AEI/10.13039/501100011033 under project PID2019-105093GB-I00 and PID2022-136436NB-I00.

References

[1] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: A Cancer Journal for Clinicians. 2021;71(3):209-49. doi:10.3322/caac.21660.

[2] Pituskin E. Cancer as a new chronic disease: Oncology nursing in the 21st Century. Canadian OnCOlOgy nursing JOurnal. 2022;32(1):87.

[3] Blom JW, Doggen CJM, Osanto S, Rosendaal FR. Malignancies, Prothrombotic Mutations, and the Risk of Venous Thrombosis. JAMA. 2005 02;293(6):715-22. doi:10.1001/jama.293.6.715.

[4] Desai R, Koipallil GK, Thomas N, Mhaskar R, Visweshwar N, Laber D, et al. Efficacy and safety of direct oral anticoagulants for secondary prevention of cancer associated thrombosis: a meta-analysis of randomized controlled trials. Scientific Reports. 2020;10(1):18945. doi:10.1038/s41598-020-75863-3.

[5] Al-Samkari H, Connors JM. Managing the competing risks of thrombosis, bleeding, and anticoagulation in patients with malignancy. Hematology 2014, the American Society of Hematology Education Program Book. 2019;2019(1):71-9. doi:10.1182/bloodadvances.2019000369.
[6] Mahajan A, Brunson A, Adesina O, Keegan TH, Wun T. The incidence of cancer-associated thrombosis is increasing over time. Blood advances. 2022;6(1):307-20. doi:10.1182/bloodadvances.2021005590.
[7] Khorana AA, Kuderer NM, Culakova E, Lyman GH, Francis CW. Development and validation of a predictive model for chemotherapy-associated thrombosis. Blood, The Journal of the American Society of Hematology. 2008;111(10):4902-7. doi:10.1182/blood-2007-10-116327.
[8] Zöller B, Li X, Ohlsson H, Ji J, Sundquist J, Sundquist K. Family history of venous thromboembolism as a risk factor and genetic research tool. Thrombosis and haemostasis. 2015;114(11):890-900. doi:10.1160/TH15-04-0306.
[9] Zöller B. Genetics of venous thromboembolism revised. Blood. 2019 11;134(19):1568-70. doi:10.1182/blood.2019002597.
[10] Muñoz Martin AJ, Ortega I, Font C, Pachón V, Castellón V, Martínez-Marín V, et al. Multivariable clinical-genetic risk model for predicting venous thromboembolic events in patients with cancer. British journal of cancer. 2018;118(8):1056-61. doi:10.1038/s41416-018-0027-8.
[11] Muñoz A, Ay C, Grilz E, López S, Font C, Pachón V, et al. A Clinical-Genetic Risk Score for Predicting Cancer-Associated Venous Thromboembolism: A Development and Validation Study Involving Two Independent Prospective Cohorts. J Clin Oncol. 2023. doi:10.1200/JCO.22.00255.
[12] Wan Z, Zhang Y, He H. Variational autoencoder based synthetic data generation for imbalanced learning. In: 2017 IEEE Symposium Series on Computational Intelligence (SSCI); 2017. p. 1-7. doi:10.1109/SSCI.2017.8285168.
[13] Abedi M, Hempel L, Sadeghi S, Kirsten T. GAN-Based Approaches for Generating Structured Data in the Medical Domain. Applied Sciences. 2022;12(14):7075. doi:10.3390/app12147075.
[14] Azizi S, Kornblith S, Saharia C, Norouzi M, Fleet DJ. Synthetic Data from Diffusion Models Improves ImageNet Classification; 2023. doi:10.48550/arXiv.2304.08466.
[15] Iman M, Arabnia HR, Rasheed K. A Review of Deep Transfer Learning and Recent Advancements. Technologies. 2023 mar;11(2):40. doi:10.3390/technologies11020040.
[16] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. Advances in neural information processing systems. 2014;27:2672-80. doi:10.1145/3422622.
[17] Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. Modeling Tabular data using Conditional GAN; 2019. doi:10.48550/arXiv.1907.00503.
[18] Team SD. SDV: Synthetic Data Vault. GitHub; 2022. Version 0.11.0. https://github.com/sdv-dev/SDV.
[19] Arjovsky M, Chintala S, Bottou L. Wasserstein GAN; 2017.