Applying GAN-based data augmentation to improve transcriptome-based prognostication in breast cancer

Working title: Data augmentation for improved breast cancer prognostication

Cristiano Guttà[1, *], Christoph Morhard[2], Markus Rehm[1, 3, *]

1 Institute of Cell Biology and Immunology, University of Stuttgart, Allmandring 31, 70569 Stuttgart, Germany

2 ProKanDo GmbH, Elfriede-Breitenbach-Straße 38a, 71640 Ludwigsburg, Germany

3 Stuttgart Research Center Systems Biology, University of Stuttgart, Nobelstrasse 15, 70569 Stuttgart, Germany

***To whom correspondence should be addressed:**

Prof Dr Markus Rehm
Institute of Cell Biology and Immunology
University of Stuttgart
Allmandring 31, 70569 Stuttgart, Germany
E-mail: markus.morrison@izi.uni-stuttgart.de
ORCID ID 0000-0001-6149-9261

Cristiano Guttà
Institute of Cell Biology and Immunology
University of Stuttgart
Allmandring 31, 70569 Stuttgart, Germany
E-mail: cristiano.gutta@izi.uni-stuttgart.de

Abstract

Established prognostic tests based on limited numbers of transcripts can identify high-risk breast cancer patients yet are approved only for individuals presenting with specific clinical features or disease characteristics. Deep learning algorithms could hold potential for stratifying patient cohorts based on full transcriptome data, yet the development of robust classifiers is hampered by the number of variables in omics datasets typically far exceeding the number of patients. To overcome this hurdle, we propose a classifier based on a data

augmentation pipeline consisting of a Wasserstein generative adversarial network (GAN) with gradient penalty and an embedded auxiliary classifier to obtain a trained GAN discriminator (T-GAN-D). Applied to 1244 patients of the METABRIC breast cancer cohort, this classifier outperformed established breast cancer biomarkers in separating low- from high-risk patients (disease specific death, progression or relapse within 10 years from initial diagnosis). Importantly, the T-GAN-D also performed across independent, merged transcriptome datasets (METABRIC and TCGA-BRCA cohorts), and merging data improved overall patient stratification. In conclusion, GAN-based data augmentation therefore allowed generating a robust classifier capable of stratifying low- vs high-risk patients based on full transcriptome data and across independent and heterogeneous breast cancer cohorts.

Introduction

Breast cancer is the tumor with the highest incidence in women, accounting for 2.3 million new diagnoses and 685,000 deaths worldwide in 2020. According to the World Health Organization, nearly eight million patients were diagnosed with breast cancer in the five years before 2020, making it the most prevalent tumor disease worldwide [1]. In current clinical practice, the expression of estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) is determined by immunohistochemistry (IHC), with the expression patterns defining to which molecular subtype (luminal A, luminal B, HER2 positive or enriched and triple-negative breast cancer) individual tumors belong. Prognosis differs between these subtypes, and subtyping informs treatment plans in patients in which surgical resection of the tumor alone is insufficient [2]. However, substantial response heterogeneities to the current standard of care treatments can be observed in populations of breast cancer patients [3], highlighting the need for additional prognostic markers that could serve to identify high risk patients that could instead benefit from alternative treatments or for which the burden from inefficient standard of care treatments could be avoided [4].

Various multi-gene activity tests based on transcript abundance have been developed to assist in the clinical management of breast cancer (e.g. Oncotype DX [5], MammaPrint [6,7], Prosigna [8,9], OncoMasTR[10]) and received regulatory approval as prognostic tests [11]. Despite the prognostic value of these assays, their use is restricted to only subsets of patients with specific clinical characteristics (e.g. cancer stage, receptor or lymph node status, tumor size, menopause state, age group) [12–14]. It would therefore be desirable if more generally applicable prognostic tests based on transcriptome data could be developed.

The rapid advances in high-throughput sequencing technologies make tumor transcriptome data from larger patient cohorts increasingly available. The accessibility of -omics databases and companion clinical information now also encourages the application of deep learning (DL) methods to the oncology field, with the aim of learning and extracting features within

large scale data that are not readily accessible by classical statistical and pattern recognition approaches. It is hoped that from DL-based methods tools can be developed that can aid in further advancing cancer diagnosis, prognosis or predicting treatment efficacy in the future [15]. DL algorithms such as convolutional neural networks (CNN) were originally applied for image analysis but could be successfully repurposed to take non-image objects as input, such as RNA-seq data [16]. One of the major pitfalls when applying DL models to transcriptome datasets is the typical imbalance between the number of quantified mRNAs (high) and the number of patients (low), which can lead to overfitting when solving classification tasks [17]. In addition, low numbers of samples or patients that represent one category (e.g. good prognosis) come at the risk of capturing patterns that are not robust when applied to larger populations [18]. Feature selection strategies [19], under- and over-sampling [20] are three strategies that may help mitigating effects arising from imbalanced source data. An alternative strategy lies in novel data augmentation approaches, such as generative adversarial networks (GANs), by which source datasets can be enriched with artificially generated additional data. GANs are typically applied to imaging data and are composed of two subnetworks, the generator and the discriminator. While the former produces synthetic images, the latter is challenged to discriminate fake vs. real images. Reiterating this process, the generator learns to produce images with features that can no longer be separated from the real images by the discriminator, with these generated images then enriching the source dataset [21]. In comparison to other generative models, GANs are currently preferred due to their computational speed and the quality of the generated images [22]. In addition, they exhibit a lower risk of overfitting classifiers and are less susceptible to the impact of non-pertinent image features (such as brightness) when enriching training data with synthetic images [23]. For example, GANs have been applied in the medical field to generate synthetic magnetic resonance, computed tomography or positron emission tomography images [24]. Aside from image-data, different

GAN implementations were also successfully applied to transcriptome data for cancer diagnosis [25,26], staging [27] and subtyping [28].

The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC, hereafter MB) [29] and The Cancer Genome Atlas – Breast Invasive Carcinoma (TCGA-BRCA, hereafter TCGA) [30] cohorts represent two of the largest and most exhaustively annotated breast cancer datasets, including, in addition to mRNA expression data, features such as patient demographics, cancer staging, receptor statuses, and follow-up information such as survival times. Despite not being directly interoperable due to different sequencing technologies, these datasets can serve as use cases to test new DL-based prognostication approaches.

In this study, we therefore set out to develop a prognostication framework that used the trained discriminator of a GAN architecture as a standalone classifier and compared its performance to classical breast cancer biomarkers and a classical CNN.

Materials and methods

Data integration

The METABRIC (MB) dataset was used to develop the prototype network implementation. Transcriptome data (median Z-scores), overall survival (OS), disease specific survival (DSS) and associated clinical records were downloaded from cbioportal.org [31,32]. The dataset was integrated with locoregional and distant recurrence information retrieved from Rueda et al. [33] and *Risk of Recurrence – Proliferation* (ROR-P) scores reported by Xia et al [9]. Clinical records, OS, DSS and progression free interval (PFI) of the validation TCGA-BRCA cohort (TCGA) were integrated from cbioportal.org [31,32] and Liu et al. 2018 [34], respectively. To merge the mRNA expression data of the two cohorts, normalized transcriptome datasets were downloaded using the R package MetaGxBreast [35]. The transcript amounts were rescaled as described by Gendoo et al. [35] so that the 2.5 percentile corresponds to -1 and the 97.5 percentile corresponds to +1. Subsequently, transcripts overlapping between the two cohorts

and with quantitative information missing in not more than five patients were retained, resulting in transcripts for m = 14042 genes. The R script used to download and rescale the datasets is available in the Zenodo repository [36].

Inclusion criteria and category definition

Both cohorts were filtered to exclude normal-like subtype samples [9,37,38] and patients for which less than 10 years of follow-up time from diagnosis were available. Low and high risk categories were defined according to published clinical records[8,9] as follows:

- high risk patients:
    - MB cohort: disease specific death, locoregional or distant recurrence event recorded before 10 years from initial diagnosis;
    - TCGA cohort: disease specific death, progression, local recurrence or distant metastases before 10 years from initial diagnosis.
- low risk patients: none of the above-mentioned events recorded before 10 years from initial diagnosis.

In total, 1248 patients of the MB cohort (n = 567 high risk, n = 681 low risk) and 165 patients of the TCGA cohort (n = 132 high risk, n = 33 low risk) satisfied the inclusion criteria. Four patients from each cohort were excluded after merging due to insufficient expression data.

Survival analysis and accuracy

Log-rank testing was used to compare predicted low vs high risk patients over a follow-up time of 10 years. Kaplan-Meier (KM) survival curves were computed using GraphPad Prism 8 (GraphPad Software, San Diego, CA). The area between the curves (ABC) displayed on the KM graphs for the pooled predictions was calculated as follows:

- Low risk AUC minus Predicted low risk AUC;
- Predicted low risk AUC minus Predicted high risk AUC;

- Predicted high risk AUC minus High risk AUC.

The ABCs values are shown on the graphs in the abovementioned order top to bottom. The AUC was computed using GraphPad Prism 8 (GraphPad Software, San Diego, CA). Univariate and multivariate hazard ratios were calculated using the function *coxph* from the R's library *survival* (v. 3.4.0, https://www.r-project.org/).

GAN architecture

The architecture was based on a Wasserstein [39] GAN [21] with gradient penalty [40] and an auxiliary classifier [41] as a variant of a conditional GAN implementation [42], yielding a AC-WGAN-GP architecture. The Wasserstein loss was implemented to reduce vanishing gradients and mode collapse [43] in the early phases of the training when the discriminator outperformed the generator. Stability was improved by exchanging the weights clipping approach described in Arjovsky et al. [39], with the gradient penalty described in Gulrajani et al. [40]. To create a conditional GAN, an auxiliary classifier network was implemented [41], resulting in a more stable training process and reduced mode collapse compared to the standard conditional GAN approach, supplying labels to both discriminator and generator [43]. A z-vector of size 250 was fed as input for the generator. Following good training practice [44], strided convolutions with step size 2, batch normalization and LeakyRELU as activation function were used. Since using batch normalization in the discriminator and/or the ADAM optimizer led to an unstable training process, batch normalization [45] was only used in the generator, and RMSprop was selected as the activation function. A shallow network consisting of two layers in both the discriminator and the generator led to the most stable training process, due to the smaller number of trainable parameters compared to deeper networks. Hyperparameters were tuned empirically, selecting 1000 epochs for the training process. Three "discriminator-only" training runs were performed before each full network training run, and the generated pictures were subsequently smoothed with a final convolution

layer with one filter and stride size of 1. The GAN architecture generated expression data of size 144x144 when using the entire transcriptome dataset of the MB cohort alone (m = 18543 genes) and 120x120 when merging the MB and TCGA cohorts (m = 14042 genes). In the latter setting, expression profiles with less than 14,440 transcripts were filled with random values, leading to better convergence. The resulting trained GAN Discriminator (T-GAN-D) was then used as an independent classifier to discriminate low and high risk patients. The Python code and the input files used to generate the predictions are available in the Zenodo repository [36].

CNN architecture

As the performance of the CNN implemented as the GAN's discriminator showed satisfactory performance, a similar architecture was used as a benchmark classifier. Batch normalization was employed to ensure shorter training periods and RELU was used as the activation function. A fixed training length of 1250 epochs was set due to the limited sample size and to generate comparable iterations.

The accuracy of both classifiers was calculated dividing the number of correct classifications by the total number of classifications performed.

Results

The METABRIC and BRCA-TCGA cohorts lend themselves as use cases for data augmentation and development of prognostication classifiers

One of the major challenges of machine learning applied to -omics data and companion medical records is the imbalance between the high amounts of variables compared to the limited number of patients available. Even in the case of breast cancer, one of the most frequent and widely studied malignant neoplasms, this limitation is apparent in the two major

public transcriptome datasets, namely the MB cohort (n = 1904 patients, m = 18543 transcripts) and the TCGA cohort (n = 1101 patients, m = 20532 transcripts). This imbalance is exacerbated for prognostic analyses that require long-term (10 years) follow-up information and the application of further exclusion criteria (see methods), reducing cohort sizes to n = 1248 and n = 165, respectively (**Fig. 1A, B**). Both cohorts behaved notably different, with patients in the MB cohort on average having an overall substantially better prognosis in overall survival and relapse-free, progression-free or disease specific survival (**Fig. 1C, D**). This is likely attributable to the MB dataset largely consisting of stage I and stage II patients (89.5% of patients with reported disease stage at diagnosis), whereas stage III and IV patients are more prominent in the TCGA dataset (40.4% of individuals with available disease stage at diagnosis). Despite these differences, the high risk subgroups of both cohorts showed comparable median survival times (MB = 31.9 months [**Fig. 1E**], TCGA = 26.3 months [**Fig. 1F**]). Due to the limited sizes of these cohorts, they lend themselves as suitably challenging use-cases for applying and testing data augmentation for improving prognostication. In particular, we set out to implement a classifier based on a data augmentation network for improved patient stratification in the MB cohort, to subsequently validate robustness and transferability by integrating the independent TCGA cohort.

A trained GAN discriminator robustly identifies low and high risk breast cancer patients

To tackle the problem of data scarcity, we implemented a GAN architecture to augment transcriptomic data of the MB cohort and tested the performance of a trained discriminator in stratifying breast cancer patients. First, individual patient transcriptome profiles were rescaled and converted into arrays of pixels (**Fig. 2A i**) in order to use these images as an input for the GAN. Independent of these true patient data, the generator created images representing the

transcript profiles of synthetic hypothetical patients together with their category (low or high risk) (**Fig. 2A ii**). After being exposed to a fraction of the real transcriptome images and associated categories, its adversary, the discriminator network then tried to distinguish fake from real transcriptome images for high or low risk patients (**Fig. 2A iii**). Reiterating this training process over 1000 epochs, the generator learned to create realistic synthetic transcriptome images for high and low risk categories, which then could be used to augment the original MB cohort data. Associated characteristics of this process (discriminator loss, discriminator class loss, generator loss) are shown in **Supplementary Fig. 1**. Using this approach, the discriminator learned to identify features relevant for the risk category definition, aided by the synthetic profiles that enriched the real training data at each epoch. The trained GAN discriminator (T-GAN-D) resulting from this process then was used as a standalone classifier to categorize images from the test fraction of the cohort into the high or low risk categories (**Fig. 2A iv**), thus prognosticating patient outcome.

We first implemented and tested the T-GAN-D for its prognostic capability using follow-up and mRNA expression data of the prototyping MB cohort, consisting of n = 1248 individuals and m = 18543 genes. Within this cohort, we independently cross-validated (CV) five-fold with randomly composed training data. Kaplan-Meier curves and log rank testing for each run yielded significant class separations in 4 out of 5 iterations (**Fig. 2B, Supplementary Fig. 2A**). Pooling the results so that each patient of the MB dataset was present once in the survival analysis, the T-GAN-D separated high and low risk patients with high statistical significance (p-value = 2.71E-12) (**Fig. 2C**). To obtain a reference performance baseline, a classical CNN was challenged with the same task, using the same training and test sets for each iteration. The CNN yielded class separations with a $p<0.05$ in only two out of five iterations (**Fig. 2D, Supplementary Fig. 2B)**. In the pooled comparison, the CNN performed well yet failed to outperform the T-GAN-D in separating low vs. high risk patients (**Fig.2E, Supplementary Table 1**). These results therefore demonstrate that the reiterative learning

process of a GAN to train its discriminator and use it as an independent classifier provides a more robust and slightly improved patient stratification than a classical DL approach.

Introducing and independent cohort improves MB patient classification

A common limitation of predictors and classifiers is their limited robustness and transferability to independent datasets. This might arise from overfitting or overtraining within the initial cohort but also from heterogeneity and batch effects between source datasets. For validating our approach further, we therefore merged the mRNA expression data of the MB and TCGA cohorts, which originally were quantified with bead-based microarray technology (Illumina Human V3) or RNA-Seq (Illumina HiSeq) platforms respectively [46], by rescaling the expression of transcripts overlapping between the two cohorts (m = 14042). We then retrained the discriminator using the entire TCGA data plus a fraction of the MB data from the merged dataset and generated predictions on an independent subset of MB patients (**Fig. 3A**), using five-fold cross-validation. The T-GAN-D again separated patients into low and high-risk categories with high statistical significance (**Fig. 3B, Supplementary Fig. 3A**). The CNN trained and tested with the same data performed similarly well (**Fig. 3C, Supplementary Fig. 3B**). The T-GAN-D trained on the merged and reduced dataset also showed improved accuracy when compared to all settings where both a CNN or the GAN were trained on the full or reduced MB dataset alone (**Supplementary Table 1, 2**). Therefore, in our setting, rescaling and converting transcriptome profiles into images was sufficient to successfully merge the two cohorts without the need for further preprocessing steps and allowed to stratify patients into high and low risk classes.

The T-GAN-D outperforms classical outcome predictors and accurately stratifies early stage patients into risk categories

We next compared the performance of CNN and GAN based classifications to other established clinical markers in breast cancer. These included a scoring system based on a multi-transcript signature (Risk-of-recurrence - proliferation, [ROR-P]), estrogen receptor status (ER), human epidermal growth factor receptor 2 status (HER2), and progesterone receptor status (PR). Likewise, tumor staging was included, yet was available for only 911 out of 1248 patients of the MB cohort. The hazard ratios (HR) obtained from a univariate analysis were comparable for ROR-P, HER2 or tumor staging as classifiers, and similar HRs were also obtained for the CNN and T-GAN-D classifiers developed from only the MB transcriptome dataset (**Fig. 4A**). Interestingly, the T-GAN-D classifier resulting from the merged cohort data returned a mean HR>2.0 (+/- 0.4), thereby surpassing all other markers. This feature was even more pronounced in a multivariate analysis including ER, HER2 and PR biomarkers (**Fig. 4B**). When reducing the MB cohort to those patients for which staging information was available, HRs based on staging and T-GAN-D were comparable (**Fig. 4C**). To test whether both classifiers might be redundant, we performed a T-GAN-D based survival analysis within the tumor stage I and stage II subcohorts, which dominate the MB dataset. T-GAN-D based classification allowed separating high and low risk patients within both tumor stages (**Fig. 4D, E**), indicating non-redundancy of the T-GAN-D classification to tumor staging information. Taken together, these results show that training through data augmentation can enhance the prognostic performance of DL classifiers, and in this case surpasses individual classical biomarkers. In addition, the T-GAN-D performed well in prognostication of early stage breast cancer cases.

The T-GAN-D stratifies TCGA patients despite these being scarcely represented

After observing that introducing TCGA patients into the training set of the T-GAN-D did not degrade, but improved the stratification of MB patients, we tested the performance of the classifier on the smaller TCGA dataset. To do this, we trained the discriminator using the entire MB data plus a fraction of the TCGA data from the merged dataset and generated predictions on an independent subset of TCGA patients (**Fig. 5A**), using five-fold cross-validation. The T-GAN-D correctly predicted 78% of the cases (**Fig. 5B, Supplementary Fig. 4**, **Supplementary Table 3)**. In contrast, when trained on the MB dataset alone, the T-GAN-D was not able to separate high and low risk patients (**Fig. 5C, Supplementary Fig. 4**), achieving an overall accuracy of only 43% (**Supplementary Table 3**). Therefore, the addition to the training set of a comparably small number of TCGA patients (n = 129) to the larger MB cohort (n = 1244) was sufficient to drastically improve the performance of the T-GAN-D predicting TCGA patient outcome. This demonstrates that even if the training set is largely dominated by patients belonging to one cohort, the introduction of a limited number of samples of a second, differently balanced dataset appears sufficient to possibly capture relevant patterns that contribute to achieving improved prognostic performance.

Discussion

The increasing availability and routine acquisition of large scale genomic data encourage the repurposing and application of AI to the field of oncology in order to identify novel means for improved and personalized prediction of prognosis [47]. In this study, we developed a DL-based tool to stratify high vs. low risk breast cancer patients according to full transcriptome profiles. Using the MB and TCGA cohorts as use cases, we converted expression data into images and used the trained discriminator of our GAN architecture as a standalone prognostic classifier. Our results show that the T-GAN-D performed better than classical outcome predictors and maintained robust performance when merging the two cohorts.

AI has already been applied to breast cancer based on different classes of data, to inform diagnosis, treatment planning and prognosis [48,49]. For example, pattern recognition and data augmentation proved to be promising approaches to assist in generating accurate diagnoses from mammography images [50,51]. Transcriptome data were also employed to develop ML-based analysis pipelines for breast cancer subtyping, diagnosis, patient stratification and identification of altered pathways [52], and these techniques may improve the accuracy of cancer prognosis in the future. However, shortcomings must be taken into account, as applicable also to currently available breast cancer datasets. When dealing with low sample size - high dimension datasets such as the MB and TCGA cohorts, common DL classification algorithms such as neural networks may be prone to overfitting [53]. Multi-gene signatures based on the expression of a lower number of transcripts may circumvent this problem, but are applicable only to subsets of patients with specific clinical characteristics [11–14]. To tackle these problems, we aimed at developing a more universally applicable algorithm that takes advantage of GAN's data augmentation and generalizing capability. In our training strategy, the T-GAN-D was exposed not only to a subset of original data, but also to the synthetic patients generated by the generator in each epoch. This approach for the augmentation of training data was demonstrated before to aid a discriminator in learning hidden features and correlations [54,55]. When compared to a classic CNN, the T-GAN-D showed comparable, yet slightly improved performance. Other GAN implementations have been applied to the MB or TCGA cohorts in the past, addressing different aims such as the generation of missing data [56], the identification of multi -omics signatures [57] and prognostication [58]. While showing encouraging results, these prior works limited the follow up time to 5 years and focused on death events only. Besides considering longer follow up times, the inclusion of progression or recurrence events in the class definition can be considered a more exhaustive assessment of a patient´s risk category, since OS or DSS alone may be insufficient especially in early stage screenings [59]. In addition, short follow up times were shown to affect the prognostication

performance of ML algorithms leading to low sensitivity, mostly due to the insufficient occurrence of recurrence or death events [60].

We demonstrated that the conversion of transcriptome profiles into images allowed the integration of independent transcriptome datasets. To date, the majority of gene expression databases cannot be directly integrated due to different sequencing technologies, protocols or batch effects, with the consequence of producing merely qualitative results in a meta-analysis fashion or unveiling evidences that remain cohort-specific [61]. To test if our conversion strategy could allow a straightforward integration of heterogenous datasets, we challenged the T-GAN-D in assessing the risk category of MB patients, training the network with a subset of MB patients plus the entire TCGA cohort. Introducing patients belonging to a different cohort improved the performance of the classifier, which in our case outperformed established clinical biomarkers and a published ROR-P signature [9] in uni- and multi-variate analyses. The T-GAN-D classifier also stratified early stage breast cancer patients into low and high risk groups, even though no additional factors such as treatment regimens, age, subtype or other clinical features were considered when composing the training datasets. Early stage patients expected to experience recurrence or progression may benefit from more frequent screenings, yet it remains to be assessed if the transcriptome-based classifier operates independently of or correlates with other established risk factors.

High accuracy in predicting the risk class of the smaller and imbalanced TCGA cohort was achieved when training the T-GAN-D with a subset of TCGA patients plus the whole MB dataset. Classical ML algorithms (SVM and random forest, among others) were also shown to benefit from the combination of TCGA RNA-Seq and MB microarray data, which in a previous study improved 5 years OS prognostication [62], but lead to misleadingly high accuracy due to highly imbalanced classes. Taken together, our results suggest that the T-GAN-D remains robust when merging cohorts differently balanced between positive and negative outcomes, and that the network is still able to capture relevant risk patterns when one

cohort is heavily underrepresented in the training dataset. Therefore, our classification framework may allow the integration of new, smaller datasets, lending itself as a suitable prototype for generating prospective personalized outcome predictions for scarce *de novo* data.

In conclusion, our proof-of-concept study represents an avenue for developing a scalable data augmentation-based tool that could be a stepping stone towards individualized prognosis in the future. Molecular high throughput techniques are increasing in quality, resolution and amount of data produced and are more and more commonly captured in clinical research and diagnostic environments. It was estimated that within the next decade, between 2 and 40 exabytes of genomic data will be generated every year [63], with large quantities being related to human health and disease. GAN-based approaches therefore could become a meaningful approach to exploit such data for the benefit of patients. In addition, -omics domains other than transcriptomics likewise have the potential to enter the clinical arena as part of routine analytical practice, including proteome, metabolome or lipidome data. Such data classes can readily be integrated with clinical-pathological information [64], and could be processed with the assistance of GAN based approaches to improve patient-tailored interventions or prognostication.

Acknowledgements

MR and CG receive funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2075 – 390740016 and acknowledge the support by the Stuttgart Center for Simulation Science (SimTech).

References

1. Breast cancer. Available at: https://www.who.int/news-room/fact-sheets/detail/breast-cancer. (Accessed: 30th August 2022)

2. Yersal, O. & Barutca, S. Biological subtypes of breast cancer: Prognostic and therapeutic implications. *World J. Clin. Oncol.* **5**, 412 (2014).

3. Turashvili, G. & Brogi, E. Tumor heterogeneity in breast cancer. *Front. Med.* **4**, 227 (2017).

4. Cardoso, F. *et al.* 70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer. *N. Engl. J. Med.* **375**, 717–729 (2016).

5. Syed, Y. Y. Oncotype DX Breast Recurrence Score®: A Review of its Use in Early-Stage Breast Cancer. *Mol. Diagnosis Ther.* **24**, 621–632 (2020).

6. Van't Veer, L. J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nat. 2002 4156871* **415**, 530–536 (2002).

7. Arc, M. *et al.* A Gene-Expression Signature as a Predictor of Survival in Breast Cancer. *https://doi.org/10.1056/NEJMoa021967* **347**, 1999–2009 (2002).

8. Bernard, P. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27**, 1160–1167 (2009).

9. Xia, Y., Fan, C., Hoadley, K. A., Parker, J. S. & Perou, C. M. Genetic determinants of the molecular portraits of epithelial cancers. *Nat. Commun. 2019 101* **10**, 1–13 (2019).

10. Buus, R. *et al.* Validation of the OncoMASTR risk score in estrogen receptor–positive/HER2-negative patients: A TransATAC study. *Clin. Cancer Res.* **26**, 623–631 (2020).

11. Ross, J. S., Hatzis, C., Symmans, W. F., Pusztai, L. & Hortobágyi, G. N. Commercialized Multigene Predictors of Clinical Outcome for Breast Cancer. *Oncologist* **13**, 477–493 (2008).

12. Yao, K., Tong, C. Y. & Cheng, C. A framework to predict the applicability of Oncotype DX, MammaPrint, and E2F4 gene signatures for improving breast cancer prognostic prediction. *Sci. Reports 2022 121* **12**, 1–11 (2022).

13. Kelly, C. M. *et al.* Comparison of the prognostic performance between OncoMasTR

and OncotypeDX multigene signatures in hormone receptor-positive, HER2-negative, lymph node-negative breast cancer. *https://doi.org/10.1200/JCO.2018.36.15_suppl.12074* **36**, 12074–12074 (2018).

14. Jensen, M. B. *et al.* The Prosigna gene expression assay and responsiveness to adjuvant cyclophosphamide-based chemotherapy in premenopausal high-risk patients with breast cancer. *Breast Cancer Res.* **20**, (2018).

15. Tran, K. A. *et al.* Deep learning in cancer diagnosis, prognosis and treatment selection. *Genome Med. 2021 131* **13**, 1–17 (2021).

16. Sharma, A., Vans, E., Shigemizu, D., Boroevich, K. A. & Tsunoda, T. DeepInsight: A methodology to transform a non-image data to an image for convolution neural network architecture. *Sci. Reports 2019 91* **9**, 1–7 (2019).

17. Liu, R. & Gillies, D. F. Overfitting in linear feature extraction for classification of high-dimensional image data. *Pattern Recognit.* **53**, 73–86 (2016).

18. Barandela, R., Valdovinos, R. M., Salvador Sánchez, J. & Ferri, F. J. The imbalanced training sample problem: under or over sampling? *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **3138**, 806–814 (2004).

19. Raghu, V. K., Ge, X., Chrysanthis, P. K. & Benos, P. V. Integrated Theory- and Data-driven Feature Selection in Gene Expression Data Analysis. *Proceedings. Int. Conf. Data Eng.* **2017**, 1525 (2017).

20. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **16**, 321–357 (2011).

21. Goodfellow, I. *et al.* Generative Adversarial Networks. *Commun. ACM* **63**, 139–144 (2014).

22. Shorten, C. & Khoshgoftaar, T. M. A survey on Image Data Augmentation for Deep Learning. *J. Big Data* **6**, 1–48 (2019).

23. Bowles, C. *et al.* GAN Augmentation: Augmenting Training Data using Generative

443 Adversarial Networks. (2018). doi:10.48550/arxiv.1810.10863

444 24. Li, X. *et al.* When medical images meet generative adversarial network: recent
445 development and research opportunities. *Discov. Artif. Intell. 2021 11* **1**, 1–20 (2021).

446 25. Xiao, Y., Wu, J. & Lin, Z. Cancer diagnosis using generative adversarial networks
447 based on deep learning from imbalanced data. *Comput. Biol. Med.* **135**, 104540 (2021).

448 26. Wei, K., Li, T., Huang, F., Chen, J. & He, Z. Cancer classification with data
449 augmentation based on generative adversarial networks. *Front. Comput. Sci. 2022 162*
450 **16**, 1–11 (2021).

451 27. Kwon, C. H., Park, S., Ko, S. & Ahn, J. Increasing prediction accuracy of pathogenic
452 staging by sample augmentation with a GAN. *PLoS One* **16**, e0250458 (2021).

453 28. Yang, H., Chen, R., Li, D. & Wang, Z. Subtype-GAN: a deep learning approach for
454 integrative cancer subtyping of multi-omics data. *Bioinformatics* **37**, 2231–2237
455 (2021).

456 29. Mukherjee, A. *et al.* Associations between genomic stratification of breast cancer and
457 centrally reviewed tumour pathology in the METABRIC cohort. *npj Breast Cancer*
458 *2018 41* **4**, 1–9 (2018).

459 30. The Cancer Genome Atlas Program - NCI. Available at:
460 https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga.
461 (Accessed: 30th August 2022)

462 31. Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles
463 using the cBioPortal. *Sci. Signal.* **6**, (2013).

464 32. Cerami, E. *et al.* The cBio cancer genomics portal: an open platform for exploring
465 multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–4 (2012).

466 33. Rueda, O. M. *et al.* Dynamics of breast-cancer relapse reveal late-recurring ER-positive
467 genomic subgroups. *Nature* **567**, 399–404 (2019).

468 34. Liu, J. *et al.* An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-

Quality Survival Outcome Analytics. *Cell* **173**, 400-416.e11 (2018).

35. Gendoo, D. M. A. *et al.* MetaGxData: Clinically Annotated Breast, Ovarian and Pancreatic Cancer Datasets and their Use in Generating a Multi-Cancer Gene Signature. *Sci. Rep.* **9**, (2019).

36. Guttà, C., Morhard, C. & Rehm, M. T-GAN-D: a GAN-based classifier for breast cancer prognostication. (2022). doi:10.5281/ZENODO.7151831

37. Troester, M. A. *et al.* Racial Differences in PAM50 Subtypes in the Carolina Breast Cancer Study. *JNCI J. Natl. Cancer Inst.* **110**, 176–182 (2018).

38. Sweeney, C. *et al.* Intrinsic subtypes from PAM50 gene expression assay in a population-based breast cancer cohort: Differences by age, race, and tumor characteristics. *Cancer Epidemiol. Biomarkers Prev.* **23**, 714 (2014).

39. Arjovsky, M., Chintala, S. & Bottou, L. Wasserstein GAN. (2017). Available at: https://arxiv.org/abs/1701.07875v3. (Accessed: 1st March 2022)

40. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V. & Courville, A. Improved Training of Wasserstein GANs. *Adv. Neural Inf. Process. Syst.* **2017-December**, 5768–5778 (2017).

41. Odena, A., Olah, C. & Shlens, J. Conditional Image Synthesis With Auxiliary Classifier GANs. *34th Int. Conf. Mach. Learn. ICML 2017* **6**, 4043–4055 (2016).

42. Mirza, M. & Osindero, S. Conditional Generative Adversarial Nets. (2014). Available at: https://arxiv.org/abs/1411.1784v1. (Accessed: 1st March 2022)

43. Kodali, N., Abernethy, J., Hays, J. & Kira, Z. On Convergence and Stability of GANs. (2017). Available at: https://arxiv.org/abs/1705.07215v5. (Accessed: 1st March 2022)

44. Radford, A., Metz, L. & Chintala, S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *4th Int. Conf. Learn. Represent. ICLR 2016 - Conf. Track Proc.* (2015).

45. Ioffe, S. & Szegedy, C. Batch Normalization: Accelerating Deep Network Training by

Reducing Internal Covariate Shift. *32nd Int. Conf. Mach. Learn. ICML 2015* **1**, 448–456 (2015).

46. Craven, K. E., Gökmen-Polar, Y. & Badve, S. S. CIBERSORT analysis of TCGA and METABRIC identifies subgroups with better outcomes in triple negative breast cancer. *Sci. Reports 2021 111* **11**, 1–19 (2021).

47. Wallis, C. How Artificial Intelligence Will Change Medicine. *Nature* **576**, S48 (2019).

48. Zhang, C. *et al.* Cancer diagnosis with DNA molecular computation. *Nat. Nanotechnol. 2020 158* **15**, 709–715 (2020).

49. Jia, D. *et al.* Breast Cancer Case Identification Based on Deep Learning and Bioinformatics Analysis. *Front. Genet.* **12**, 767 (2021).

50. McKinney, S. M. *et al.* International evaluation of an AI system for breast cancer screening. *Nat. 2020 5777788* **577**, 89–94 (2020).

51. Desai, S. D., Giraddi, S., Verma, N., Gupta, P. & Ramya, S. Breast Cancer Detection Using GAN for Limited Labeled Dataset. *Proc. - 2020 12th Int. Conf. Comput. Intell. Commun. Networks, CICN 2020* 34–39 (2020). doi:10.1109/CICN49253.2020.9242551

52. Liñares-Blanco, J., Pazos, A. & Fernandez-Lozano, C. Machine learning analysis of TCGA cancer data. *PeerJ Comput. Sci.* **7**, 1–47 (2021).

53. Liu, B., Wei, Y., Zhang, Y. & Yang, Q. Deep neural networks for high dimension, low sample size data. in *IJCAI International Joint Conference on Artificial Intelligence* 2287–2293 (2017). doi:10.24963/ijcai.2017/318

54. He, K., Zhang, X., Ren, S. & Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. in *Proceedings of the IEEE International Conference on Computer Vision* **2015 Inter**, 1026–1034 (2015).

55. Shams, S., Platania, R., Zhang, J., Kim, J. & Park, S. J. Deep generative breast cancer screening and diagnosis. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **11071**

LNCS, 859–867 (Springer Verlag, 2018).

56. Arya, N. & Saha, S. Generative Incomplete Multi-View Prognosis Predictor for Breast Cancer: GIMPP. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 1–1 (2021). doi:10.1109/TCBB.2021.3090458

57. Kim, M., Oh, I. & Ahn, J. An Improved Method for Prediction of Cancer Prognosis by Network Learning. *Genes (Basel).* **9**, 1. – 11 (2018).

58. Hsu, T. C. & Lin, C. Generative Adversarial Networks for Robust Breast Cancer Prognosis Prediction with Limited Data Size. *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS* **2020-July**, 5669–5672 (2020).

59. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V. & Fotiadis, D. I. Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* **13**, 8–17 (2015).

60. Boeri, C. *et al.* Machine Learning techniques in breast cancer prognosis prediction: A primary evaluation. *Cancer Med.* **9**, 3234 (2020).

61. Carnielli, C. M. *et al.* Combining discovery and targeted proteomics reveals a prognostic signature in oral cancer. *Nat. Commun.* **9**, 3598 (2018).

62. Dubourg-Felonneau, G. *et al.* A Framework for Implementing Machine Learning on Omics Data. (2018). doi:10.48550/arxiv.1811.10455

63. Stephens, Z. D. *et al.* Big Data: Astronomical or Genomical? *PLoS Biol.* **13**, (2015).

64. Karczewski, K. J. & Snyder, M. P. Integrative omics for health and disease. *Nat. Rev. Genet.* **19**, 299 (2018).

Figure legends

Fig. 1. MB and TCGA patient demographics and survival

(**A**) Patients demographics of the MB subcohort. (**B**) Patients demographics of the TCGA

subcohort. (**C**) Overall and (**D**) relapse-free, progression-free or disease specific survival of the MB and TCGA cohorts. (**E**) Kaplan Meier curves comparing low vs high risk patients of the MB and (**F**) the TCGA cohorts.

Fig. 2. The T-GAN-D robustly stratifies low and high risk breast cancer patients

(**A**) Workflow of the data processing, including the schematics of the generator network and its adversary, the discriminator network. Together these result in an AC-WGAN-GP architecture. After the conversion of patient transcriptome profiles into images, 4/5 of the MB dataset was used to train the GAN's discriminator. After 1000 epochs, the trained discriminator was used as a standalone classifier to separate the remaining 1/5 patients of the dataset into low and high risk categories. (**B**) Kaplan-Meier curves separating low vs. high risk patients as predicted with the T-GAN-D (iteration 1 of the 5-fold CV shown as representative). (**C**) Kaplan-Meier curves generated pooling the category predictions obtained for all patients of the MB dataset after five independent CV runs. (**D**) Separation of low vs. high risk patients predicted with a classical CNN on the same subset used in **B** and (**E**) comparison obtained pooling the predictions of five independent CV runs. The area between the curves (ABC) between Low risk (blue dashed line) and Predicted low risk (solid blue line), Predicted low risk and Predicted high risk (solid red line), Predicted high risk and High risk groups (dashed red line) are shown top to bottom in **D** and **E**.

Fig. 3. Introducing the independent TCGA cohort improves MB patient classification

(**A**) Schematic representing the training strategy: rescaled data from the entire TCGA cohort were merged with 4/5 of the MB cohort to train the T-GAN-D, which was subsequently used to predict the risk class of the remaining 1/5 of MB patients. The process was iterated 5 times. (**B**) Kaplan-Meier curves based on the pooled predictions of the T-GAN-D trained on both cohorts. (**C**) Kaplan-Meier curves separating low vs. high risk patients predicted with the

CNN that was trained after merging the MB and the TCGA cohorts. The area between the curves (ABC) between Low risk (blue dashed line) and Predicted low risk (solid blue line), Predicted low risk and Predicted high risk (solid red line), Predicted high risk and high risk groups (dashed red line) are shown top to bottom in **B** and **C**.

Fig. 4 The T-GAN-D outperforms classical biomarkers after merging the MB and TCGA cohorts and significantly stratifies early stage MB patients

(**A**) Comparison of the hazard ratios (Cox model, univariate) of a multi-transcript signature (ROR-P) and established prognostic biomarkers (ER, HER2, PR) vs. the CNN and the T-GAN-D before and after cohort merging. (**B**) Multivariate Cox hazard ratio of the T-GAN-D compared to ROR-P and receptor status and (**C**) disease stage. (**D**) Kaplan -Meier curves of Stage I and (**E**) Stage II patients stratified by the T-GAN-D into low and high risk categories.

Fig. 5 The T-GAN-D stratifies TCGA patients despite these being scarcely represented in the merged training set

(**A**) Schematic representing the training strategy: rescaled data from the entire MB cohort were merged with 4/5 of the TCGA cohort to train the T-GAN-D, which was subsequently used to predict the risk class of the remaining 1/5 of TCGA patients. The process was iterated 5 times. (**B**) Stratification of the TCGA patients by T-GAN-D trained on the merged dataset and (**C**) the MB dataset alone. Kaplan-Meier curves were generated pooling the predictions of all iterations of the 5-fold CV. The area between the curves (ABC) between Low risk (blue dashed line) and Predicted low risk (solid blue line), Predicted low risk and Predicted high risk (solid red line), Predicted high risk and High risk groups (dashed red line) are shown top to bottom in **B** and **C**.

Fig. 1

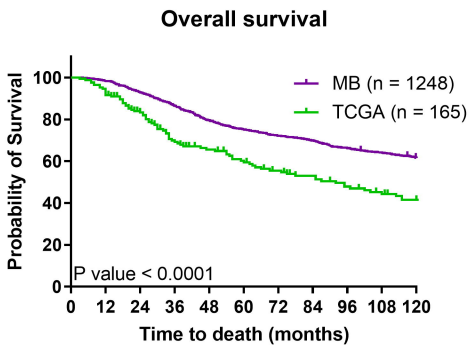
A

METABRIC sub-cohort (n = 1248)		
Age at diagnosis (years)	**# cases**	**%**
<40	83	6.7
40-49	188	15.1
50-59	288	23.1
60-69	392	31.4
70-79	235	18.8
80+	62	5.0
Disease Stage	**# cases**	**%**
Stage I	303	24.3
Stage II	514	41.2
Stage III	88	7.1
Stage IV	8	0.6
Not available	335	26.8
Survival (months)	**Median**	
Overall survival	163.2	
Relapse or disease specific death	160.2	

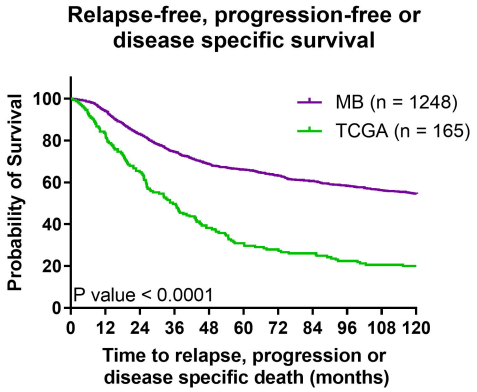
B

BRCA-TCGA sub-cohort (n = 165)		
Age at diagnosis (years)	**# cases**	**%**
<40	20	12.1
40-49	38	23.0
50-59	29	17.6
60-69	44	26.7
70-79	22	13.3
80+	12	7.3
Disease Stage	**# cases**	**%**
Stage I	22	13.3
Stage II	72	43.6
Stage III	50	30.3
Stage IV	14	8.5
Not available	7	4.2
Survival (months)	**Median**	
Overall survival	92	
Progression or disease specific death	35.5	

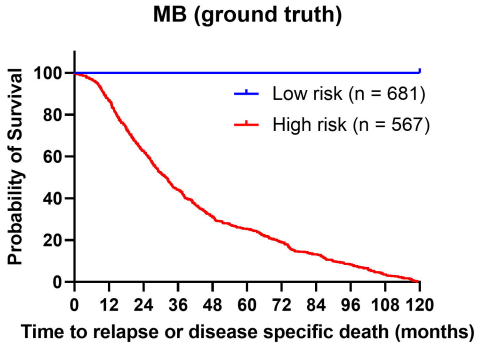
C



D



E



F

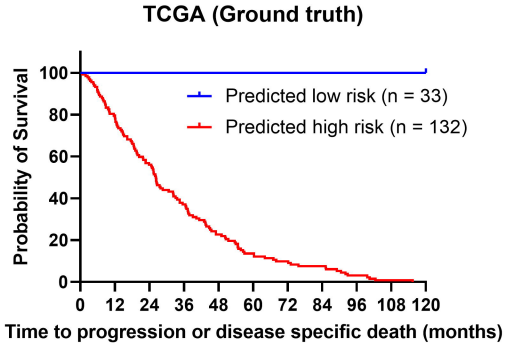
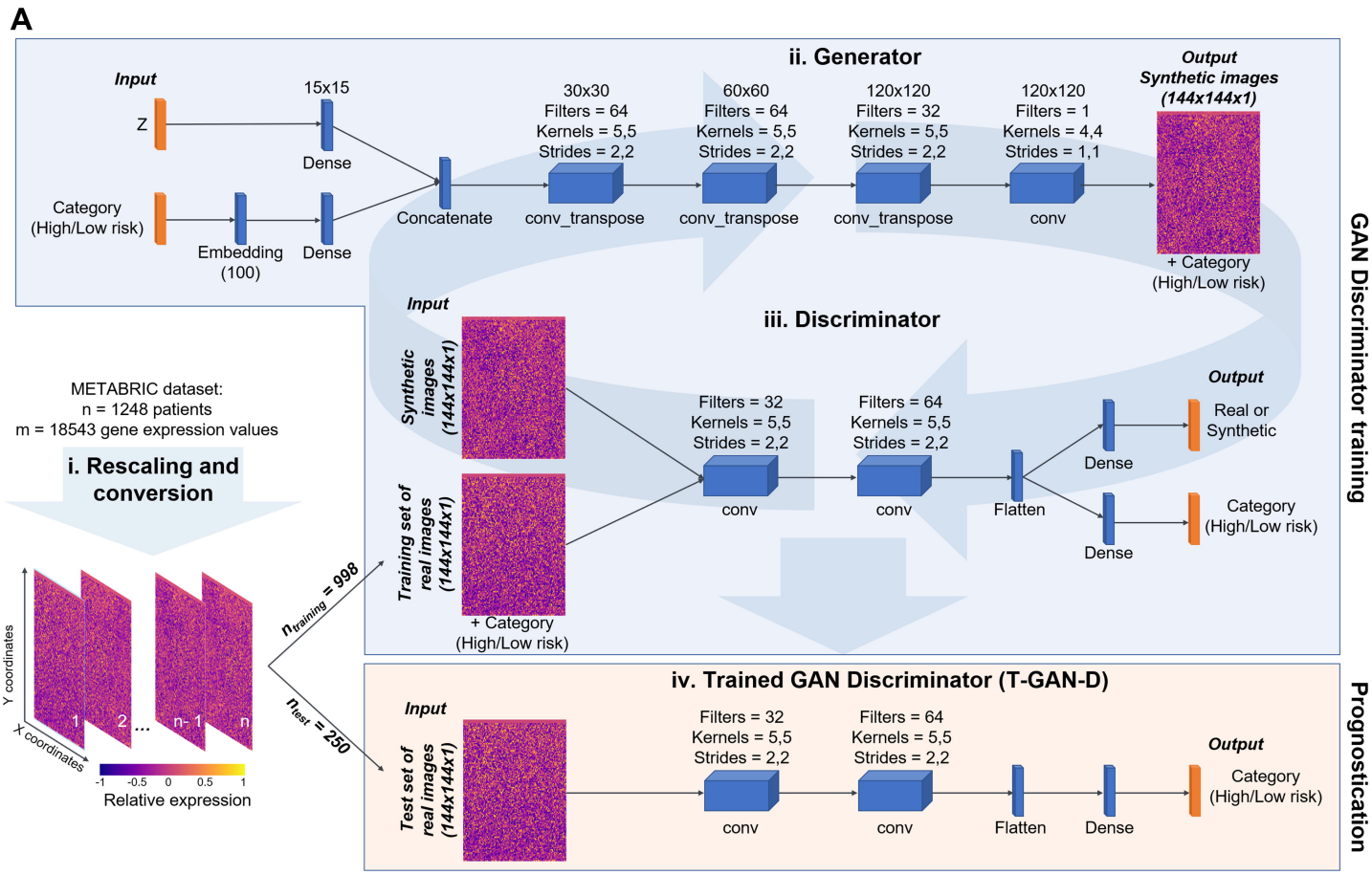
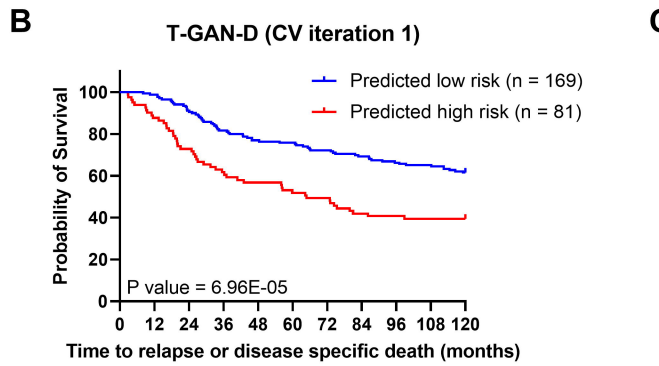
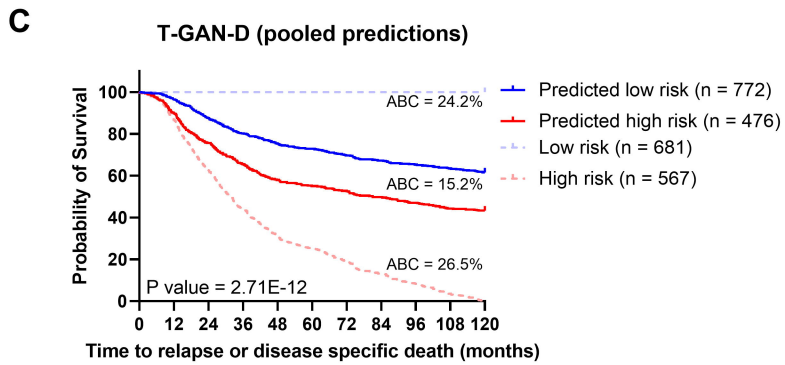
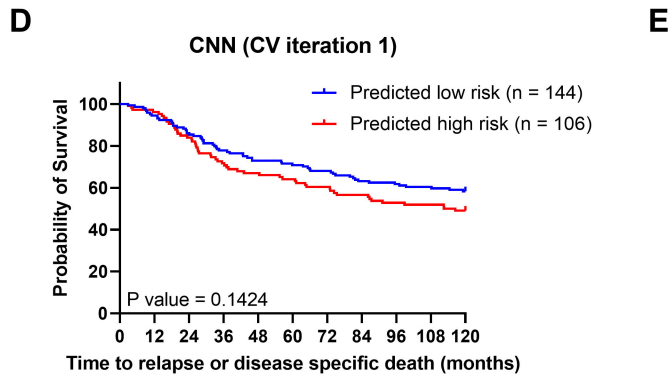


Fig. 2









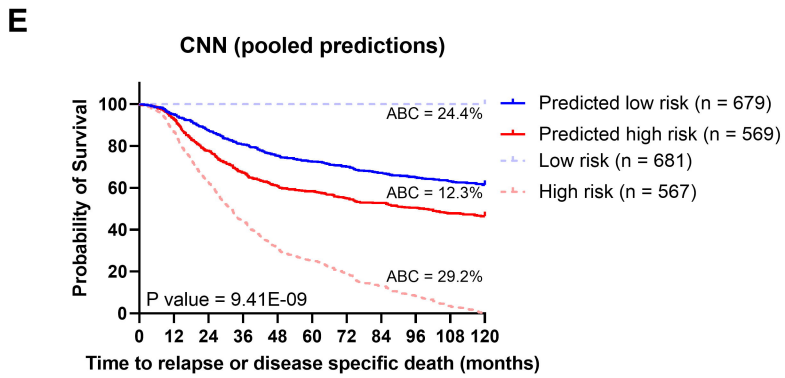
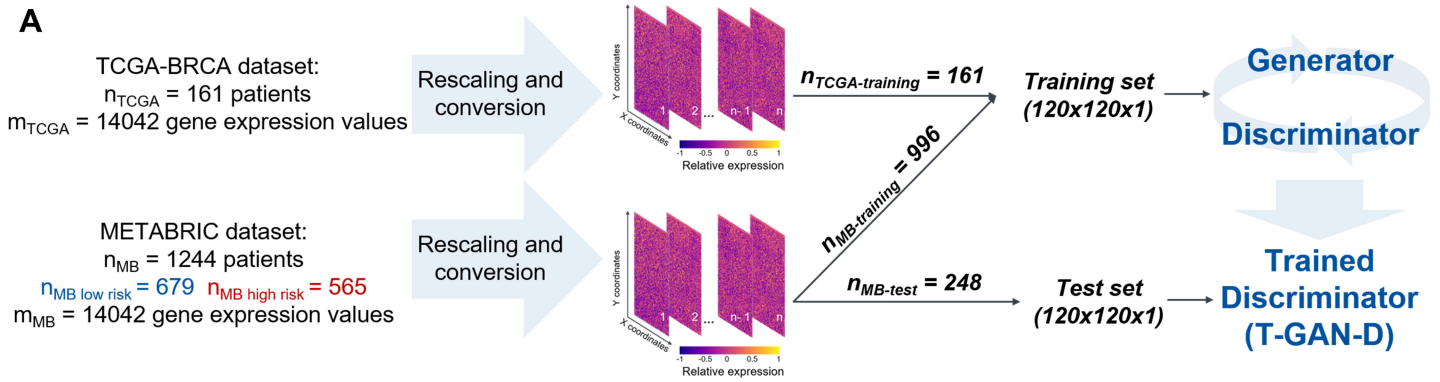
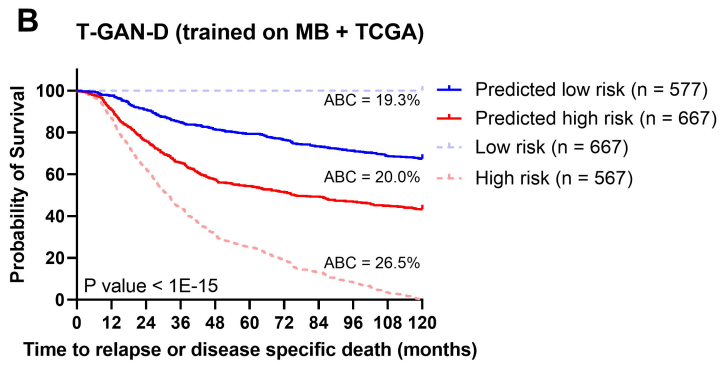


Fig. 3





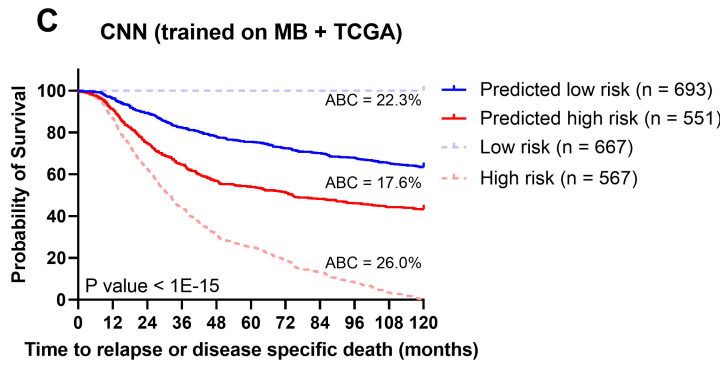


Fig. 4

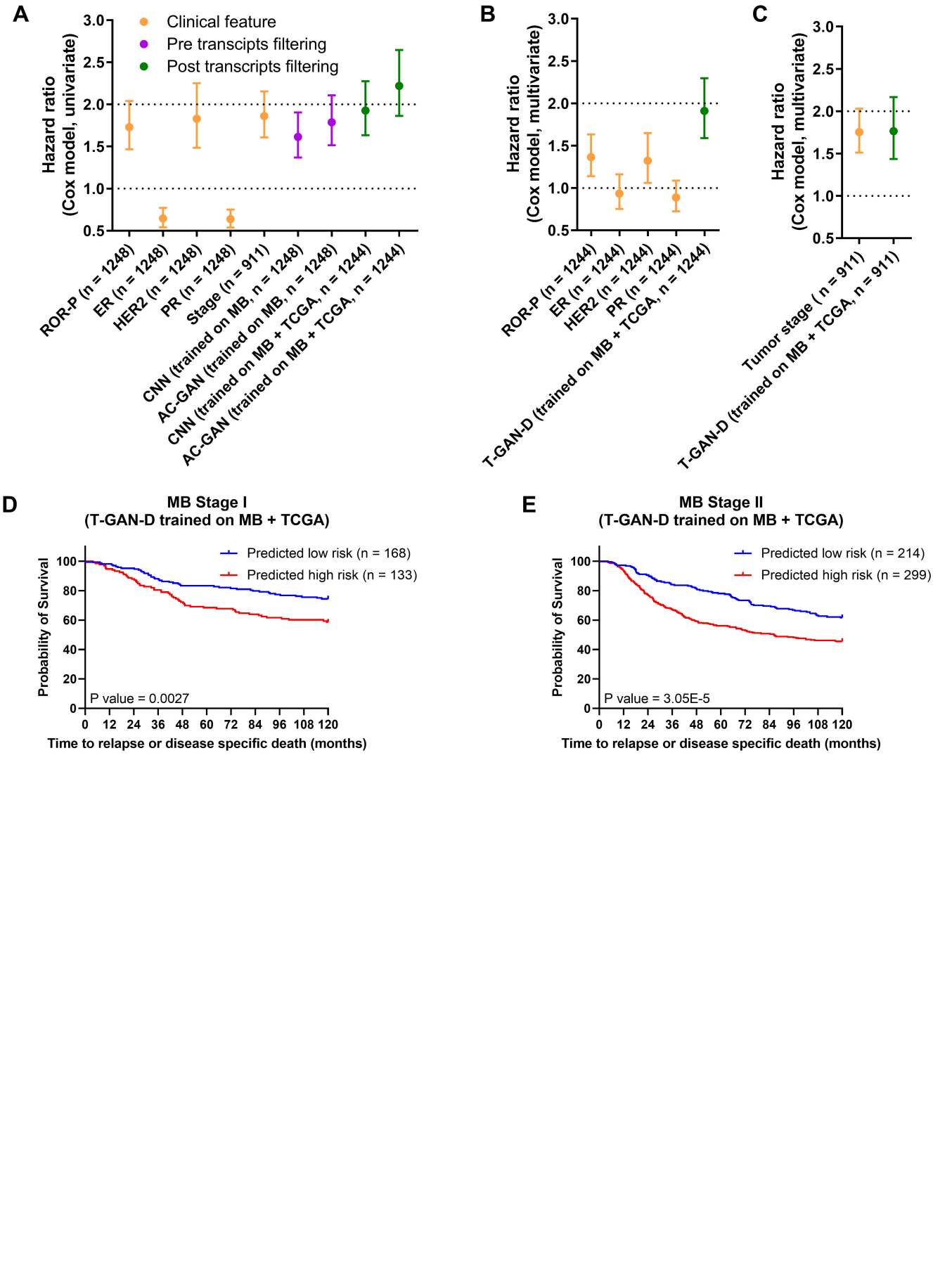


Fig. 5