

## Stable breast cancer prognosis

Xiaomei Li<sup>1</sup>, Lin Liu<sup>1</sup>, Jiuyong Li<sup>1</sup>, Thuc D. Le<sup>1,2\*</sup>

**1** UniSA STEM, University of South Australia, Adelaide, SA 5095, Australia

**2** Centre for Cancer Biology, University of South Australia, Adelaide, SA 5000, Australia

\* Thuc.Le@unisa.edu.au

### Abstract

Predicting breast cancer prognosis helps improve the treatment and management of the disease. In the last decades, many prediction models have been developed for breast cancer prognosis based on transcriptomic data. A common assumption made by these models is that the test and training data follow the same distribution. However, in practice, due to the heterogeneity of breast cancer and the different environments (e.g. hospitals) where data are collected, the distribution of the test data may shift from that of the training data. For example, new patients likely have different breast cancer stage distribution from those in the training dataset. Thus these existing methods may not provide stable prediction performance for breast cancer prognosis in situations with the shift of data distribution. In this paper, we present a novel stable prediction method for reliable breast cancer prognosis under data distribution shift. Our model, known as Deep Global Balancing Cox regression (DGBCox), is based on the causal inference theory. In DGBCox, firstly high-dimensional gene expression data is transferred to latent network-based representations by a deep auto-encoder neural network. Then after balancing the latent representations using a proposed causality-based approach, causal latent features are selected for breast cancer prognosis. Causal features have persistent relationships with survival outcomes even under distribution shift across different environments according to the causal inference theory. Therefore, the proposed DGBCox method is robust and stable for breast cancer prognosis. We apply DGBCox

to 12 test datasets from different breast cancer studies. The results show that DGBCox outperforms benchmark methods in terms of both prediction accuracy and stability. We also propose a permutation importance algorithm to rank the genes in the DGBCox model. The top 50 ranked genes suggest that the cell cycle and the organelle organisation could be the most relevant biological processes for stable breast cancer prognosis.

## Author summary

Various prediction models have been proposed for breast cancer prognosis. The prediction models usually train on a dataset and predict the survival outcomes of patients in new test datasets. The majority of these models share a common assumption that the test and training data follow the same distribution. However, as breast cancer is a heterogeneous disease, the assumption may be violated in practice. In this study, we propose a novel method for reliable breast cancer prognosis when the test data distribution shifts from that of the training data. The proposed model has been trained on one dataset and applied to twelve test datasets from different breast cancer studies. In comparison with the benchmark methods in breast cancer prognosis, our model shows better prediction accuracy and stability. The top 50 important genes in our model provide clues to the relationship between several biological mechanisms and clinical outcomes of breast cancer. Our proposed method in breast cancer can potentially be adapted to apply to other cancer types.

## Introduction

Breast cancer prognosis can help tailor treatments for patients and improve their survival outcomes. Traditional breast cancer prognosis mostly relies on the tumour-node-metastasis (TNM) staging system [1]. However, tumours with the same clinical characteristics can differ in both prognosis and treatment response because breast cancer is heterogeneous with variable molecular mechanisms of carcinogenesis and cancer development. Therefore, it is desirable to use transcriptomic data to conduct breast cancer prognosis at molecular level.

In the past decades, a large amount of transcriptomic data have been generated for breast cancer using next-generation sequencing and microarray technologies, and many computational methods have been developed to utilize the data for breast cancer prognosis. Most of the methods are based on Cox regression [2]. For example, rorS [3] trains a risk model by ridge penalized Cox regression based on the breast cancer subtype information obtained from gene expression data. TAMR13 [4] builds a Cox regression model based on the data of gene probe cluster centroids. Similarly, OncotypeDX [5], GENIUS [6] and EndoPredict [7] use Cox regression to determine the weights of the selected genes and then calculate the hazard risk as a weighted combination of the gene expression levels.

Some machine learning based methods have also been used for cancer prognosis, such as Cox regression regularized by elastic Net [8], Random Survival Forest (RSF) [9], Support Vector Regression for Censored Data (SVRc) [10], Cox proportional hazards deep neural network (DeepSurv) [11] and so on (see [12] for a review). These methods take advantage of machine learning techniques, to selecting vital features in high-dimension transcriptomic data and to model non-linear relationships between features (genes) and survival outcomes.

These methods assume no distribution shift from the training data to the test data (new unseen data). However, the assumption will be violated in real applications. For example, gene expression data may be produced using different platforms, by different labs and patients may come from different countries or have different breast cancer stages. Users do not know what future unseen data will be like. Inaccurate prognosis predictions will have resulted from the distribution shift even though the models perform well in the training and hold out validation datasets.

The computational methods need to have stable prediction performance since their predictions directly impact treatment decisions. Here the stability of a method means that the method performs consistently with the training data and the test data (i.e. unseen breast cancer patients) which may have different distributions. One strategy to improve the stability of a prediction method is to utilize samples from multiple sources to train the model. For example, several methods have been developed to select a set of robust features from multiple datasets obtained from the studies of the same disease, and use these features to build a more reliable prediction model [13,14]. Another

strategy is to employ transfer learning techniques to train a robust prediction model [15]. For example, the method in [16] transfers the weights of a variational auto-encoder trained with all the transcriptomic data of 20 cancer types (source domain) to the survival prediction model for patients from a single cancer type (target domain). The transfer learning based methods enhance the robustness of survival predictions in unseen cancer patients by extracting the shared genomic knowledge from the source domain and transfer the knowledge to the target domain.

From the above discussion, we can see that these existing stable prediction methods, when used for breast cancer prognosis, require cancer patient samples with transcriptomic data and/or survival outcomes coming from either different sources or multiple domains in the training phases. Both are challenging in practice because it is time-consuming to follow up cancer patients over long-term periods and collect a wide variety of data. Therefore, it is useful to develop survival prediction models that train on an available dataset and are robust to distribution shifts in test data from new environments.

In this paper, we focus on designing a model for stable survival prediction under new environments where the data distribution has been shifted from that in the training dataset. The new environment is agnostic to us and the distribution in the test data is unknown. In this case, we cannot use any information from the test data when we train a model. To learn a robust prediction model across different environments, the predictors (signatures) must be causal genes, i.e. genes causing the development of the disease [17]. The pinpointed causal genes in the model can be translated into disease mechanism and clinical treatment [18]. Therefore, we will model stable breast cancer prognosis predictions as a causal problem.

In the research area of causal inference, the causal effects are used for measuring the impact of a treatment on patients. Covariate balancing strategies are used for estimating the causal effect of a treatment variable on the outcome in the observational data. A recent global balancing strategy by Kuang *et al.* [19] is a covariate balancing strategy used to estimate the causal effects of covariates on the outcome when one does not know which covariates are causal and which ones are not. This can be used for our design because we do not know which genes are causal. However, three limitations prevent this method from being immediately applied to achieve a stable breast cancer

prognosis. Firstly, the method is for a classification problem with the categorical  
outcome not for a regression problem with the continuous outcome as in prognosis  
prediction. Secondly, this method is only applicable to data without censoring,  
unfortunately, the survival outcomes in breast cancer datasets are almost always  
censored. Thirdly, it cannot deal with high dimensional gene expression data used in  
building breast cancer prognosis models.

We propose a novel method, Deep Global Balancing Cox regression (DGBCox) for  
stable breast cancer prognosis based on censored survival data and gene expression data.  
With DGBCox, we have made two main contributions: (1) We leverage the strengths of  
both regularized Cox regression and global balancing approach and propose a novel  
method for stable cancer prognosis. To the best of our knowledge, this is the first cancer  
prognosis model which is aimed at stable predictions across different environments. (2)  
We have developed a new deep learning based framework to deal with high dimensional  
gene expression data.

We apply DGBCox to breast cancer datasets from different studies. The experiment  
results have demonstrated that DGBCox performed better than baseline methods and  
existing breast cancer prognosis methods with regard to accurate and stable breast  
cancer prognosis. We interpret the important genes for the stable breast cancer  
prognosis in our model. A high percentage of the top 50 important genes discovered by  
DGBCox has been selected by previous research for breast cancer prognosis. The  
important genes are enriched in several well-known biological processes involved in the  
development of cancer. These findings may help to understand the biological  
mechanisms of breast cancer development and progression.

## Materials and methods

### Datasets

In this study, we use 13 genome-wide expression datasets containing in total 5668 breast  
cancer patients from different repositories (Table 1). Among these datasets, TCGA753  
and TCGA500 are subsets of TCGA1093 generated by The Cancer Genome Atlas  
(TCGA) program (<https://www.cancer.gov/tcga>). METABRIC is available at the

European Genome-phenome Archive (EGA) (<https://www.ebi.ac.uk/ega/>, accession number EGAS00000000083, access required). The MAINZ, TRANSBIG, UPP, UNT, and NKI datasets are from Bioconductor (<https://bioconductor.org/>) data packages, breastCancerMAINZ, breastCancerTRANSBIG, breastCancerUPP, breastCancerUNT, and breastCancerNKI, respectively. The remaining datasets are all downloaded from the Gene Expression Omnibus database (<https://www.ncbi.nlm.nih.gov/geo/>). More details about the datasets are in Section 1 in S1 File.

**Table 1. A summary of the datasets**

| Dataset  | Platform                            | #Samples | Min time <sup>1</sup> | Max time <sup>2</sup> | Event (%) <sup>3</sup> | Reference  |
|----------|-------------------------------------|----------|-----------------------|-----------------------|------------------------|------------|
| METABRIC | Illumina HumanHT-12 V3              | 1283     | 0                     | 29.60                 | 32.68                  | [20]       |
| TRANSBIG | Affymetrix Human Genome U133A       | 198      | 0.34                  | 24.95                 | 31.31                  | [21]       |
| UNT      | Affymetrix Human Genome U133A/B     | 137      | 0.17                  | 14.53                 | 21.05                  | [22]       |
| UPP      | Affymetrix Human Genome U133A/B     | 251      | 0.08                  | 12.75                 | 23.08                  | [23]       |
| MAINZ    | Affymetrix Human Genome U133A       | 200      | 0.08                  | 19.72                 | 23.00                  | [24]       |
| NKI      | Agilent Human Genome Microarray     | 337      | 0.02                  | 18.35                 | 34.17                  | [25, 26]   |
| GSE6532  | Affymetrix Human Genome U133 Plus 2 | 414      | 0.02                  | 16.85                 | 34.66                  | [27]       |
| GEO736   | Affymetrix Human Genome U133 Plus 2 | 736      | 0                     | 18.52                 | 47.42                  | [4, 28–31] |
| TCGA1093 | Illumina GA/HiSeq RNA-seq           | 1093     | 0                     | 23.44                 | 10.35                  | [32]       |
| TCGA753  | Illumina GA/HiSeq RNA-seq           | 753      | 0                     | 17.90                 | 9.57                   | [32]       |
| TCGA500  | Illumina GA/HiSeq RNA-seq           | 500      | 0.003                 | 17.90                 | 9.02                   | [32]       |
| UK       | Illumina humanRef-8 V1              | 207      | 0.39                  | 10.00                 | 37.20                  | [33]       |
| HEL      | Illumina HumanHT-12 V3              | 115      | 0                     | 5.00                  | 21.74                  | [34]       |

1: the minimum survival time in years in a dataset. 2: the maximum survival time in years in a dataset. 3: the percent of patients experiencing events in a dataset.

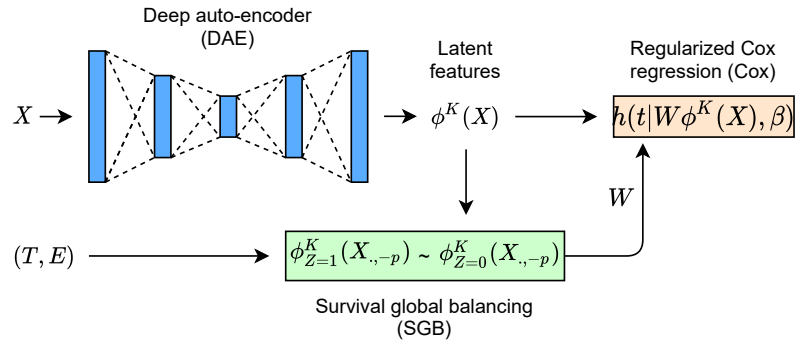
The clinical outcomes included in the datasets used in this study include relapse free survival time (UPP, GSE6532, GEO736, TCGA1093, TCGA753, TCGA500, METABRIC, and UK), distant metastasis free survival time (TRANSBIG, UNT, MAINZ, and NKI), and death or distant metastasis free survival time (HEL). In all the datasets, the unit of survival time is a year. The maximum survival time in the datasets ranges from 5 years to 29.6 years. The medical event in the datasets is breast cancer relapse, metastasis or death. The event percentage in the datasets ranges from 9.02% to 47.42%. These breast cancer datasets were collected from different batches of experiments with different RNA-sequencing platforms and different pipelines for RNA-seq data processing and have different ranges of survival time and different ratios of event occurrences. Therefore, it is reasonable to assume that the distributions of these datasets are different.

## Stable cancer prognosis using Deep Global Balancing Cox regression (DGBCox)

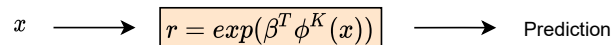
### Overview of DGBCox

As illustrated in Fig 1, DGBCox has three main components in the training phase, a deep auto-encoder neural network (DAE), a survival global balancing component (SGB) and a regularized Cox regression (Cox). DAE is used for extracting the latent patient representations from the original gene expression data of a patient. Cox is used to predicting a patient's hazard risk that indicates how likely the patient is to experience an event. SGB re-weights the training data samples so that it is possible to infer the causal relationships between latent features and the survival outcomes. The three components of DGBCox are trained using the input dataset  $D = (\mathbf{X}, T, E)$ , where  $\mathbf{X}$  is gene expression data,  $T$  is observed survival time and  $E$  is event status. In the test phase, DGBCox predicts the hazard risk of a patient only based on the trained DAE and Cox components. In the following sections, we explain the three components, the training and test phases of DGBCox.

#### A: Training phase of DGBCox



#### B: Test phase of DGBCox



**Fig 1. The framework of DGBCox.** **A.** DGBCox contains three main components in the training phase, i.e. DAE, SGB and Cox. SGB balances weighted covariate distributions in treated and control groups and then offers weights in  $W$  to Cox. DAE provides latent representations in  $\phi^K(\mathbf{X})$  to SGB and Cox. DAE, SGB and Cox are jointly trained by an optimization algorithm. **B.** For an unseen test patient with gene expression profile  $x$ , the survival prediction is an exponent of the linear combination of latent features, where the coefficients  $\beta$  (in Cox) and the encoder function  $\phi^K(\cdot)$  (in DAE) are determined in the training phase.

## Deep auto-encoder neural network (DAE) 136

The gene expression data of cancer patients are typical of high dimensionality, but the global balancing technique employed by DGBCox (for the SGB component) cannot deal with high dimensional data if it successively treats each gene as a treatment variable and learns weights to balance the dataset (details in the next section). Therefore we apply a deep auto-encoder neural network [35] to reduce the dimensionality of the gene expression data and feed the low dimensional latent features to SGB. 137  
138  
139  
140  
141  
142

DAE consists of two parts, the encoder and the decoder, which can be defined as transitions  $\phi : \mathbf{X} \rightarrow \phi(\mathbf{X})$  and  $\varphi : \phi(\mathbf{X}) \rightarrow \varphi(\mathbf{X})$ , respectively. Specifically, the latent representations for each hidden layer of the encoder can be given as: 143  
144  
145

$$\begin{aligned}\phi^1(\mathbf{X}) &= \sigma(A^1\mathbf{X} + b^1) \\ \phi^k(\mathbf{X}) &= \sigma(A^k\phi^{k-1}(\mathbf{X}) + b^k), k = 2, \dots, K\end{aligned}\tag{1}$$

where  $K$  is the number of hidden layers in the encoder.  $A^k$  and  $b^k$  are the weight matrix and bias vector on the  $k$ -th hidden layer of the encoder, respectively.  $\sigma(\cdot)$  is an activation function. 146  
147  
148

The decoder transitions the resulting latent representations  $\phi^K(\mathbf{X})$  to a reconstructed  $\hat{\mathbf{X}}$  with the same shape of  $\mathbf{X}$ : 149  
150

$$\begin{aligned}\varphi^1(\mathbf{X}) &= \sigma(\hat{A}^1\phi^K(\mathbf{X}) + \hat{b}^1) \\ \varphi^k(\mathbf{X}) &= \sigma(\hat{A}^k\varphi^{k-1}(\mathbf{X}) + \hat{b}^k), k = 2, \dots, K\end{aligned}\tag{2}$$

where  $\hat{A}^k$  and  $\hat{b}^k$  are the weight matrix and bias vector on the  $k$ -th layer of the decoder respectively. The reconstructed  $\hat{\mathbf{X}}$  is the  $K$ -th layer output  $\varphi^K(\mathbf{X})$ . DAE is trained to minimize reconstruction errors between the input gene expression data  $\mathbf{X}$  and the reconstructed  $\hat{\mathbf{X}}$ , such as  $\left\|\mathbf{X} - \hat{\mathbf{X}}\right\|_2^2$ . 151  
152  
153  
154

The output of the DAE component is the low-dimensional latent features (in  $\phi^K(\mathbf{X})$ ) which are the non-linear combinations of the input gene expression data. For the sake of simplicity, the superscript  $K$  is dropped when the context is clear.  $\phi(\mathbf{X})$  is used for input data of SGB and Cox instead of  $\mathbf{X}$  as described in the following sections. 155  
156  
157  
158



## Survival global balancing (SGB) 159

It is known that causal relationships between covariates are persistent and stable across 160  
different environments whereas noncausal relationships or purely correlations may 161  
variant across different environments [36]. Therefore to achieve stable predictions for 162  
cancer prognosis, for our proposed DGBCox, we aim to use the latent features which are 163  
causally related to the survival outcomes (called causal latent features in this paper) as 164  
the predictors in the prediction model (i.e. the Cox regression model) for prognosis. To 165  
find the causal features, we need to estimate the causal effect of a latent feature on the 166  
survival outcomes. Covariate balancing is a way to estimate causal effects [37]. Here we 167  
extend the global balancing method [19] to censored survival data and propose the 168  
survival global balancing (SGB) component for DGBCox. SGB re-weights the training 169  
data samples, which enables estimation of the causal effect of each latent feature and 170  
thus to identify the causal features. 171

In the following, we start with introducing the general idea of covariate balancing 172  
and the global balancing method [19] for non-censored data. Then we propose our 173  
method to extend the covariate balancing approach and the global balancing method to 174  
censored survival data. Lastly, we incorporate the extended global balancing approach 175  
with DAE for censored data. 176

Let  $Z_i \in \{0, 1\}$  denote treatment assignment, with  $Z_i = 1$  indicating the  $i$ -th patient 177  
is treated and  $Z_i = 0$  indicating the opposite.  $\mathbf{X}_i$  and  $Y_i$  describe the features and the 178  
observed survival time for the  $i$ -th patient respectively. The subscript  $i$  will be dropped 179  
when the context is clear.  $Y^{(z)}$  ( $z \in \{0, 1\}$ ) denote the potential survival outcomes of a 180  
patient under the treatment assignment  $Z$  and  $Y = ZY^{(1)} + (1 - Z)Y^{(0)}$ . The observed 181  
dataset is composed of i.i.d. samples  $(Y_i, Z_i, \mathbf{X}_i)$ , for  $i = 1, 2, \dots, N$ , where  $N$  is the 182  
sample size of the training dataset. In causal inference, to estimate of average treatment 183  
effect from observed data with non-random treatment assignment, covariate balancing is 184  
a commonly used method to balance the distribution of the covariates between treated 185  
and control groups [19]. Under the assumption of unconfoundedness 186  
 $(Z \perp\!\!\!\perp (Y^{(1)}, Y^{(0)}) | \mathbf{X})$ , the general idea of the covariate balancing approach is to 187  
re-weight the training samples, i.e. re-weight the  $i$ -th sample with the weight  $W_i$ , and 188

the weight vector  $W = \{W_1, W_2, \dots, W_N\}$  is obtained as follows: 189

$$\arg \min_W \left\| \frac{\sum_{i:Z_i=1} W_i \cdot \mathbf{X}_i}{\sum_{i:Z_i=1} W_i} - \frac{\sum_{i:Z_i=0} W_i \cdot \mathbf{X}_i}{\sum_{i:Z_i=0} W_i} \right\|_2^2 \quad (3)$$

Through the sample re-weighting, the average treatment effect can be estimated using 190  
the difference of the expected weighted survival time of patients in treated and control 191  
groups [19]. 192

When the causal features (regarded as treatments) are unknown, Kuang *et al.* 193  
proposed a Global Balancing (GB) method to identify causal features from the 194  
observational data [19]. GB successively takes each covariate as a treatment variable 195  
and balance the distribution of all the other covariates between treatment and control. 196  
As a result, GB learns global sample weights in  $W$  satisfied with 197

$$\arg \min_W \sum_{p=1}^{\hat{P}} \left\| \frac{\mathbf{X}_{\cdot, -p}^T \cdot (W \odot \mathbf{X}_{\cdot, p})}{W^T \cdot \mathbf{X}_{\cdot, p}} - \frac{\mathbf{X}_{\cdot, -p}^T \cdot (W \odot (1 - \mathbf{X}_{\cdot, p}))}{W^T \cdot (1 - \mathbf{X}_{\cdot, p})} \right\|_2^2 \quad (4)$$

where  $\hat{P}$  is the dimension of  $\mathbf{X}$ ,  $\mathbf{X}_{\cdot, p}$  is the  $p$ -th covariate or the binary of the  $p$ -th 198  
covariate (for non-binary variable) in  $\mathbf{X}$ , and  $\mathbf{X}_{\cdot, -p} = \mathbf{X} \setminus \mathbf{X}_{\cdot, p}$  holds all the remaining 199  
covariates except the  $p$ -th covariate in  $\mathbf{X}$ .  $\odot$  refers to element-wise multiplication. 200

Now we extend Formula 3 to right censored survival data. Right censoring occurs 201  
when a patient does not experience an event but she is lost to follow-up or the medical 202  
study ends. Therefore, when right censoring occurs to a patient, instead of observing  $Y$ , 203  
one observes  $T = ZT^{(1)} + (1 - Z)T^{(0)}$ , and  $T^{(z)} = \min(Y^{(z)}, C)$ , where  $C$  denotes the 204  
censoring time. Similarly, one observes the event status  $E = ZE^{(1)} + (1 - Z)E^{(0)}$ , 205  
where  $E^{(z)} = I(C \geq Y^{(z)})$  and  $I(\cdot)$  is an indication function.  $E_i = 1$  indicates an event 206  
has been observed for the  $i$ -th patient and the realized survival time of the patient has 207  
been completely observed, while  $E_i = 0$  indicating the opposite. The censored survival 208  
data are thus samples of i.i.d. random vectors  $(T, E, E \cdot Y, Z, \mathbf{X})$ . Similar to data 209  
without censoring, we assume the treatment variable is independent of censored and 210  
uncensored survival time given the gene expression data, the unconfoundedness 211  
assumption is now extended to  $(T^{(1)}, T^{(0)}, Y^{(1)}, Y^{(0)}) \perp\!\!\!\perp Z | \mathbf{X}$ . In addition, we assume 212  
the censoring is independent of survival time and gene expression data given the 213  
treatment variable  $(T^{(1)}, T^{(0)}, Y^{(1)}, Y^{(0)}, \mathbf{X}) \perp\!\!\!\perp C | Z$ . Inspired by estimating of average 214

treatment effect for censored survival data [38], we extend the covariate balancing approach to censored survival data by:

$$\arg \min_W \left\| \frac{\sum_{i:Z_i=1} \frac{W_i \cdot E_i \cdot \mathbf{X}_i}{F_1(T_i, Z_i)} - \frac{\sum_{i:Z_i=0} \frac{W_i \cdot E_i \cdot \mathbf{X}_i}{F_0(T_i, Z_i)}}{\sum_{i:Z_i=1} \frac{W_i \cdot E_i}{F_1(T_i, Z_i)} - \frac{\sum_{i:Z_i=0} \frac{W_i \cdot E_i}{F_0(T_i, Z_i)}}} \right\|_2^2 \quad (5)$$

where  $F_z(T, Z) = Pr(C > T | Z = z)$  denotes the conditional censoring distribution which is estimated by the Kaplan–Meier method [39, 40].

With the extended covariate balancing approach, we propose SGB to learn global sample weight vector  $W$  from censored survival data:

$$\arg \min_W \sum_{p=1}^{\hat{P}} \left\| \frac{\mathbf{X}_{\cdot, -p}^T \cdot (W \odot \hat{\pi}_p \odot \mathbf{X}_{\cdot, p})}{(W \odot \hat{\pi}_p)^T \cdot \mathbf{X}_{\cdot, p}} - \frac{\mathbf{X}_{\cdot, -p}^T \cdot (W \odot \hat{\pi}_p \odot (1 - \mathbf{X}_{\cdot, p}))}{(W \odot \hat{\pi}_p)^T \cdot (1 - \mathbf{X}_{\cdot, p})} \right\|_2^2 \quad (6)$$

where  $\hat{\pi}_p = E \odot F_z(T, \mathbf{X}_{\cdot, p})$ , and  $\odot$  refers to element-wise division.

In gene expression data  $\mathbf{X}$ , the dimension  $\hat{P}$  is large. There may be no sufficient samples and computational resources to estimate  $W$  by optimizing Formula 6.

Therefore we successively regard each latent feature in a low-dimensional space from DAE as a treatment variable and learn  $W$  by:

$$\arg \min_W \sum_{p=1}^P \left\| \frac{\Phi_{\cdot, -p}^T \cdot (W \odot \pi_p \odot Z_{\cdot, p})}{(W \odot \pi_p)^T \cdot Z_{\cdot, p}} - \frac{\Phi_{\cdot, -p}^T \cdot (W \odot \pi_p \odot (1 - Z_{\cdot, p}))}{(W \odot \pi_p)^T \cdot (1 - Z_{\cdot, p})} \right\|_2^2 \quad (7)$$

where  $\Phi = \phi(\mathbf{X})$ ,  $P$  is the dimension of  $\Phi$ , and  $Z_{\cdot, p}$  is the  $p$ -th treatment variable via dichotomizing  $\Phi_{\cdot, p}$  by its mean value.  $\Phi_{\cdot, p}$  is the  $p$ -th covariate in  $\Phi$ , and  $\Phi_{\cdot, -p} = \Phi \setminus \Phi_{\cdot, p}$ .  $\pi_p = E \odot F_z(T, Z_{\cdot, p})$ .

### Regularized Cox regression (Cox)

In survival analysis literature, a hazard function is used for modelling the instant probability of an individual will experience an event at time  $t$  given that she has already survived up to time  $t$ . One can estimate the hazard risk score(s) using the hazard function for each individual based on their baseline data. Cox regression [2] is a commonly used survival analysis method. Here we first introduce the general Cox regression model build on original gene expression data. The hazard function of Cox

regression is defined as:

$$h(t|\mathbf{X}, \hat{\beta}) = h_0(t) \exp(\hat{\beta}^T \mathbf{X}) \quad (8)$$

where  $h_0(t)$  is the baseline hazard function which is not specified in Cox regression. The prediction of Cox regression is the ratio  $\frac{h(t|\mathbf{X})}{h_0(t)} = \exp(\hat{\beta}^T \mathbf{X})$  which represents the relative hazard risk of a patient.  $\hat{\beta}$  holds the coefficients of the predictors and can be estimated from training data, and this is often done through minimizing the following loss function (called average negative log partial likelihood):

$$- \sum_{i: E_i=1} (\hat{\beta}^T \mathbf{X}_i - \log \sum_{j \in \mathfrak{R}(T_i)} \exp(\hat{\beta}^T \mathbf{X}_j)) \quad (9)$$

where  $\mathfrak{R}(T_i) = \{j : T_j \geq T_i\}$  is the set of patients who are still at risk of event at time  $T_i$ .

However, Cox regression for high-dimensional data is challenging and often does not behave well [8]. Thus we build a regularized Cox regression model based on the latent features found by DAE instead of the original gene expression data. Moreover, to estimate the causal effects of latent features on survival outcomes, we need to re-weighting training samples by the weights obtained by SGB (Formula 7). Therefore, we use the following loss function of our modified Cox regression:

$$- \sum_{i: E_i=1} W_i \cdot (\beta^T \phi(\mathbf{X}_i) - \log \sum_{j \in \mathfrak{R}(T_i)} \exp(\beta^T \phi(\mathbf{X}_j))) \quad (10)$$

$$s.t. \|\beta\|_1 \leq \lambda_1, \|\beta\|_2^2 \leq \lambda_2.$$

where the L1 regularization (also known as lasso regularization) term ( $\|\beta\|_1 \leq \lambda_1$ ) and the L2 regularization term ( $\|\beta\|_2^2 \leq \lambda_2$ ) help to avoid overfitting of our modified Cox regression model. After survival global balancing, causal latent features are more likely selected as predictors, which leads to stable performance for cancer prognosis.

### Joint optimization of the three components in DGBCox

A joint optimization strategy is used for tradeoffs between the prediction accuracy (by Cox), the reconstruction error (by DAE) and the global balancing regularization (by

SGB) in DGBCox. The loss function of DGBCox is defined as:

257

$$\begin{aligned}
 & - \sum_{i:E_i=1} W_i \cdot (\beta^T \phi(\mathbf{X}_i) - \log \sum_{j \in \mathfrak{R}(T_i)} \exp(\beta^T \phi(\mathbf{X}_j))) \\
 & \quad \text{s.t. } \|\beta\|_1 \leq \lambda_1, \|\beta\|_2^2 \leq \lambda_2, \\
 & \quad \|\mathbf{X} - \hat{\mathbf{X}}\|_2^2 \leq \lambda_3, \sum_{k=1}^K (\|A^k\|_2^2 + \|\hat{A}^k\|_2^2) \leq \lambda_4, \\
 & \quad \sum_{p=1}^P \left\| \frac{\Phi_{\cdot, -p}^T \cdot (W \odot \pi_p \odot Z_{\cdot, p})}{(W \odot \pi_p)^T \cdot Z_{\cdot, p}} - \frac{\Phi_{\cdot, -p}^T \cdot (W \odot \pi_p \odot (1 - Z_{\cdot, p}))}{(W \odot \pi_p)^T \cdot (1 - Z_{\cdot, p})} \right\|_2^2 \leq \lambda_5, \\
 & \quad W_i \geq 0, \|W\|_2^2 \leq \lambda_6, \sum_{i=1}^N (W_i - 1)^2 \leq \lambda_7.
 \end{aligned} \tag{11}$$

The condition of  $\sum_{i=1}^N (W_i - 1)^2 \leq \lambda_7$  helps to avoid all the weights to be 0.

258

### Cancer prognosis with DGBCox

259

The relative hazard risk score for a new patient is calculated by  $r = \exp(\beta^T \phi(x))$ . The patient with a low (hazard) risk score is expected to have a better prognosis than the one with a high risk score.

260

261

262

### Evaluation metrics

263

We first introduce the traditional evaluation metrics for cancer prognosis methods, and then define the metrics for evaluating the stability of a prognosis method.

264

265

**Concordance index** C-index [41] is a commonly used measure for evaluating the accuracy of the risk score prediction of a cancer prognosis method. Let  $Y_i$  and  $r_i$  be the potential survival time and the risk score for a patient  $i$ , respectively. C-index is defined as the concordance probability  $Pr(r_j > r_i | Y_j < Y_i)$  for all randomly selected pairs of patients  $(i, j)$ . The concordance means that the patient with a high risk score survives shorter than the one with a low risk score. In other words, the predicted risk score should be negatively associated with the length of survival. However, the potential survival time  $Y$  cannot be observed in the right censored survival data. The observed survival outcomes contain observed survival time  $T$  and event status  $E$ . For the right censored survival data, the concordance probability is extended to

266

267

268

269

270

271

272

273

274

275

$Pr(r_j > r_i | T_j < T_i, T_i \leq C_i)$ , where  $C_i$  is the right censoring time.  $T_i \leq C_i$  indicates  
that the patient  $i$  experienced an event during the follow-up time and  $E_i = 1$ .  
Therefore, C-index can be computed with the following formula:

$$\frac{\sum_i \sum_j I(T_j > T_i) \cdot I(r_j < r_i) \cdot E_i}{\sum_i \sum_j I(T_j > T_i) \cdot E_i} \quad (12)$$

C-index ranges from 0 to 1. If the C-index of a method equals 0.5, that means this  
method is no better than a random guess model. A model with a C-index of 1 indicates  
that the model's predictions and survival outcomes are perfectly concordant. On the  
contrary, a model with a C-index of 0 means that the model's predictions and survival  
outcomes are discordant. In this case, the predicted risk score is positively associated  
with the length of survival which is not true in cancer prognosis.

**Hazard ratio** A commonly used metric for evaluating risk group predictions is  
hazard ratio [42]. We obtain the predicted groups  $G$  for patients via dichotomizing the  
predicted hazard risk scores by their median value. If a patient's risk score is bigger  
than the median predicted hazard risk score of the cohort, the patient is put into the  
high-risk group, otherwise, the patient is put into the low-risk group. Then the risk  
groups  $G$  is fitted in a Cox regression model:

$$h(t|G, \mu) = h_0(t) \exp(\mu G) \quad (13)$$

where  $h_0(t)$  is the same as that in Eq 8. The quantity  $\exp(\mu)$  is defined as hazard ratio  
which indicates the risk difference between the two groups of patients. The method with  
the high hazard ratio performs better than the one with the low hazard ratio.

**The Log-rank test** We use the Log-rank test [43] to assess whether a method  
successfully stratified patients into two risk groups with statistically significant different  
survival distributions. In the Log-rank test, the null hypothesis is that there is no  
difference between the survival distribution in the two risk groups at any time point. If  
the p-value of the Log-rank test for a method is less than 0.05, we reject the null  
hypothesis and consider that the method successfully stratified patients into two risk  
groups with distinct survival patterns.

**Metrics for stability evaluation** Given a set of test datasets  $(D^1, D^2, \dots, D^M)$ ,  $CI^m$  and  $HR^m$  are the C-index and hazard ratio on dataset  $D^m$ ,  $m \in \{1, 2, \dots, M\}$ , where  $M$  is the number of test datasets. To assess whether a cancer prognosis model is a stable and accurate model, we define a set of stability evaluation metrics as below.

$$\begin{aligned}Stability_{ci} &= mean(\{CI^m\}_{m=1}^M) - sd(\{CI^m\}_{m=1}^M) \\Stability_{hr} &= mean(\{HR^m\}_{m=1}^M) - sd(\{HR^m\}_{m=1}^M)\end{aligned}\tag{14}$$

where  $mean(\cdot)$  and  $sd(\cdot)$  returns the mean and standard deviation of a given vector that holds C-indices or hazard ratios of a method on all the  $M$  test datasets. The larger  $Stability_{ci}$  or  $Stability_{hr}$  is, the more stable the cancer prognosis method will be. In this study, a cancer prognosis model with high stability implies that the model has not only a high average accuracy but also a low standard deviation of accuracy across different environments.

## Prioritizing stable features

To interpret the DGBCox method, we propose a permutation importance approach to measure how much DGBCox's stability depends on the information in each covariate (gene). The permutation importance algorithm calculates the decrease in prediction stability when a covariate is permuted (randomly shuffled). The pseudocode of the permutation importance algorithm is illustrated in Algorithm 1 in S1 File.

## Experimental setting

To evaluate the effectiveness of DGBCox, we create and implement the following baselines to investigate the contribution of the components (in particular DAE and SGB) proposed for DGBCox.

- A Cox regression model with elastic net regularization (Coxnet)
- Deep Auto-Encoder Cox regression (DAECox)

The first baseline (Coxnet) is to evaluate the effect of both DAE and SGB, and the second baseline (DAECox) is to evaluate the effect of SGB. Only integrating SGB and a regularized Cox regression is impractical to the high dimensional gene expression data,

therefore, we do not compare DGBCox with such a method. The details of the two  
baselines are described in Section 2 in S1 File.

Since the proposed method aims at stable prediction of the hazard risks of new  
breast cancer patients from unknown environments, we train the proposed method and  
baselines based on METABRIC and test the models on the other 12 independent  
datasets. C-index is the primary metric to evaluate the accuracy for risk score  
prediction of a cancer prognosis method. Thus, in this study, the hyper-parameters of  
the methods are determined by maximizing  $Stability_{ci}$  on validation datasets by a  
Bayesian hyper-parameters search [44]. More information about the hyper-parameters  
and the training data partition is in Section 3 in S1 File.

We also compare our methods with the state-of-the-art methods, including  
DeepSurv [11], Robust [13], AURKA [45], ESR1 [45], ERBB2 [45], GGI [22],  
GENIUS [6], Endopredict [7], OncotypeDx [5], TAMR13 [4], PIK3CAGS [27],  
GENE70 [25] and rorS [3]. We implement DeepSurv and Robust as follows. The optimal  
hyper-parameters of DeepSurv are selected based on cross validation on METABRIC by  
the Bayesian hyper-parameters Optimization. Robust is trained based on METABRIC  
to determine the parameters of its predictive model. The remaining methods have  
selected gene signatures (predictors) based on prior knowledge and clinical information.  
These methods have been implemented by R package CancerSubtypesPrognosis  
(<https://github.com/XiaomeiLi1/CancerSubtypesPrognosis>).

## Results

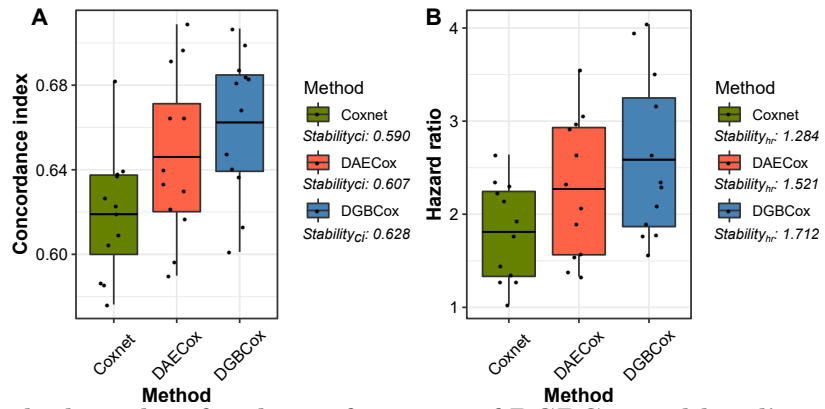
### Effectiveness of DGBCox

Fig 2A shows the risk score prediction performance of DGBCox and the two baselines  
(Coxnet and DAECox). Fig 2B shows the risk group prediction performance of the three  
methods. Fig 2 also shows the stability results based on  $Stability_{ci}$  and  $Stability_{hr}$ .  
Please note that the higher the better for all the evaluation metrics.

From the results shown in Fig 2, we have the following observations:

- Compared with Coxnet, DAECox achieves better performance in terms of  
 $Stability_{ci}$  and  $Stability_{hr}$ . This indicates that the DAE component improves the



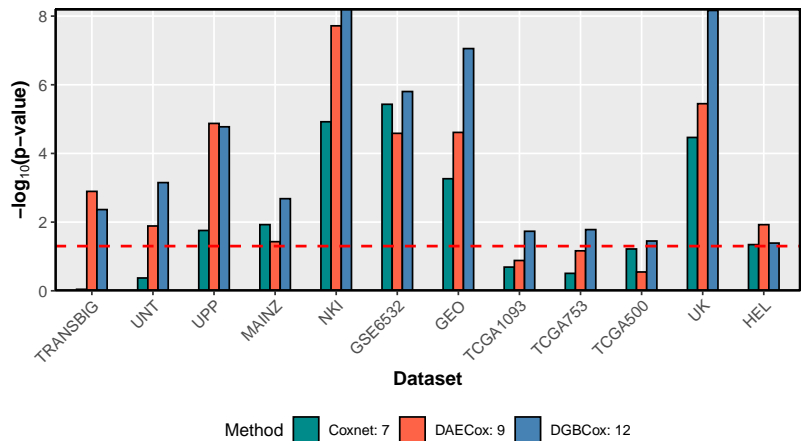


**Fig 2.** The box plots for the performance of DGBCox and baselines on the independent test datasets. **A:** the box plot for the concordance indices. **B:** the box plot for the hazard ratios. The horizontal line in each box indicates the mean value.

performance of Coxnet for both risk score and risk group predictions. This is because Coxnet only models the linear relationship between genes and survival outcomes, while DAECox can capture the non-linear relationship between genes and survival outcomes.

- In comparison with DAECox and Coxnet, DGBCox makes the most stable risk score and risk group prediction. This is because the SGB component ensures approximate estimation of the causal effect of features, therefore, causal features can be selected in the prediction model, which leads to accurate and stable prediction across different environments. Moreover, DGBCox obtains the highest average C-index of 0.662 and the highest average *HR* of 2.586, outperforming the other two methods.

Furthermore, we use the Log-rank test to analyze whether the survival pattern in one risk group is significantly different from that in another one. The results are shown in Fig 3. From the figure, we see that DGBCox outperforms Coxnet and DAECox by 5 and 3 respectively in terms of the number of datasets where a method obtains small p-values (less than 0.05). The results confirm that DGBCox is a viable approach for improving the performance in stratifying breast cancer patients into two risk groups that have distinct survival patterns.



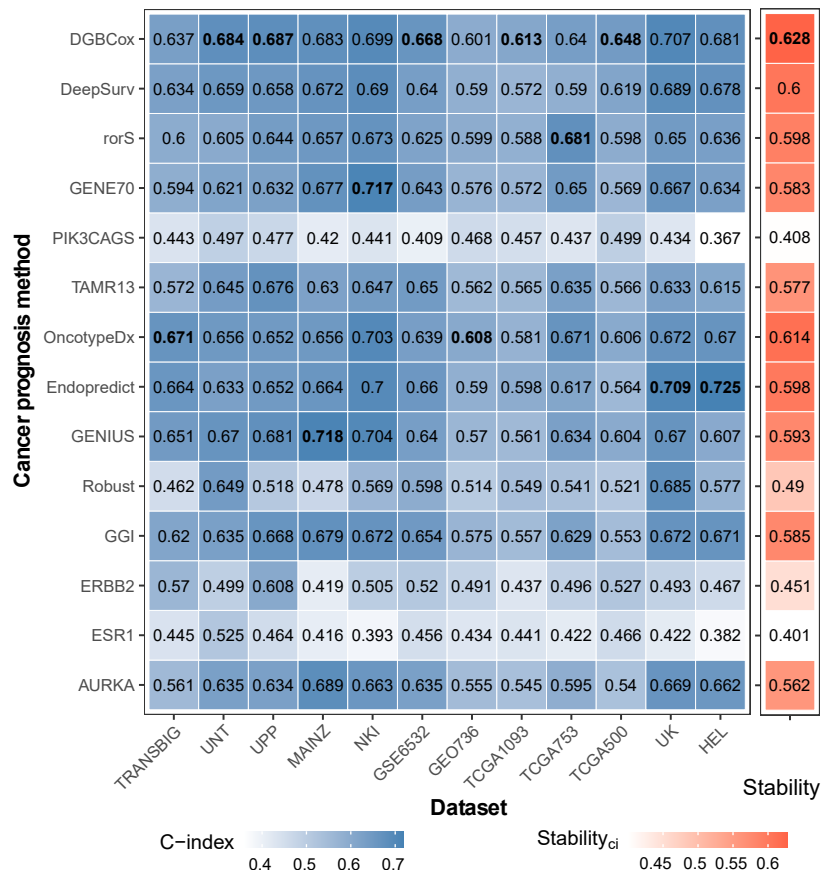
**Fig 3. The bar plot for the performance of DGBCox and baselines on 12 test datasets.** The x-axis stands for the dataset, and the y-axis shows the log transformation of the p-value of the Log-rank test. The red horizontal line indicates the threshold for significantly results i.e. the results over the red line are statistically significant, while the results under the line are not.

### Comparison with the state-of-the-art methods

Fig 4 shows the C-indices of DGBCox and each of the 13 state-of-the-art methods compared on each of the 12 test datasets, and the  $Stability_{ci}$  of each method across all test datasets. As shown in Fig 4, DGBCox outperforms the state-of-the-art methods on five datasets (UNT, UPP, GSE6532, TCGA1093, TCGA500) in terms of C-index. Meanwhile, the state-of-the-art methods only display the best results on up to two datasets. For a model which is capable of making predictions correctly, the C-index of the model on each test dataset should be higher than 0.5. From Fig 4, we can see that most methods are capable of making predictions correctly on most test datasets, but DGBCox achieves the most notable performance with the C-indices over 0.6 on all the datasets. We observed that the C-indices of some methods (ESR1, ERBB2, Robust, and PIK3CAGS) are less than 0.5 on many datasets, which means the correlations between the predicted risk scores by those methods and the survival outcomes of patients in different cohorts are contradictory. These incoherent associations may be attributable to that some signatures used in these methods vary across different environments, while the Cox regression method cannot correct the distribution shift.

When assessing the stability of each method across different environments for the risk score prediction, DGBCox also outperforms all the existing methods because it achieves the highest  $Stability_{ci}$  (0.628) as shown in Fig 4. A good method should have a

high average C-index and a low standard deviation of C-indices across test datasets. Conversely, even ESR1 has a low standard deviation (0.038) of C-indices, it performs poorly in most datasets and has a low average C-index (0.439).



**Fig 4. Comparison of different methods for risk score prediction.** The left block holds the C-indices of methods on each test dataset. The x-axis stands for the dataset, and the y-axis stands for the method. The C-indices are coloured from white to steel blue. The best result for each data is highlighted in bold. The right block displays  $Stability_{ci}$  of methods across the test datasets. The values of  $Stability_{ci}$  are coloured from white to tomato colour.

We further conduct the comparison of different methods for risk group prediction. As shown in S1 Fig, DGBCox outperforms all the state-of-the-art methods for risk group prediction based on  $Stability_{hr}$  and the number of datasets for which a method has the best  $HR$ . All the  $HR$ s of most methods (except ESR1, ERBB2, Robust, and PIK3CAGS) are larger than 1, which means the patients in the predicted high risk group have higher risks of an event when compared to those in the predicted low risk group. These results are consistent with the results of the risk score prediction.

We also use the Log-rank test to evaluate the methods for risk group prediction. The

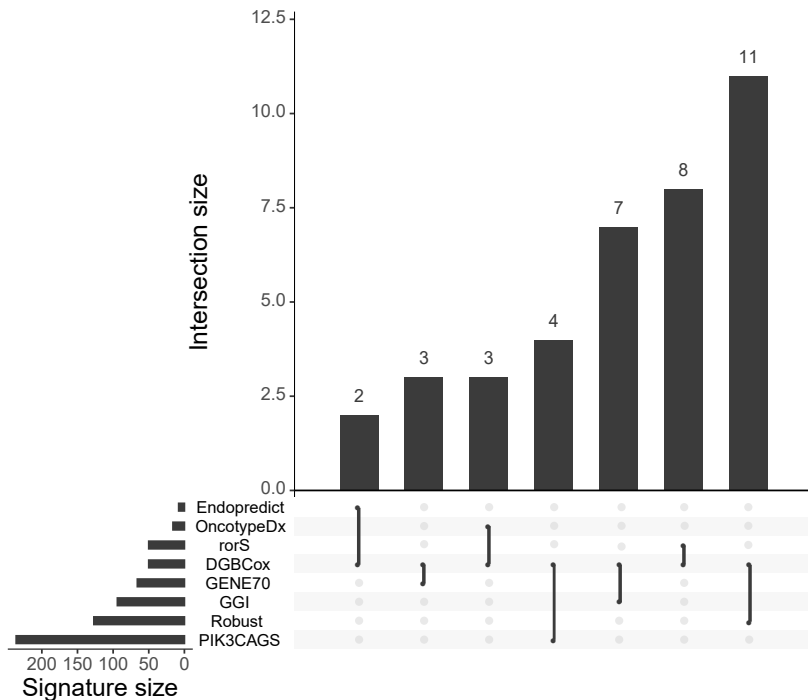
results show that DGBCox successfully stratifies patients in all the twelve test datasets 403  
into two risk groups that have distinct survival patterns (S2 Fig). In the meantime, the 404  
state-of-the-art methods only succeed on up to nine datasets. 405

## Important genes for stable prognosis 406

To further understand which genes are influential in DGBCox, we use the permutation 407  
importance algorithm to prioritize genes based on their contributions to prediction 408  
stability. Since the signatures discovered by the existing breast cancer prognosis 409  
methods (including AURKA, ESR1, ERBB2, GGI, Robust, GENIUS, Endopredict, 410  
OncotypeDx, TAMR13, PIK3CAGS, GENE70 and rorS) have been validated as breast 411  
cancer related genes, we compare the top 50 ranked genes in DGBCox with these 412  
signatures. We map the probe identifiers of signatures to gene symbols if needed. As 413  
shown in Fig 5, there are two to three genes in common between the top 50 ranked 414  
genes and the breast signature genes identified by Endopredict, OncotypeDx and 415  
GENE70 respectively. Four of the 50 genes are confirmed as PIK3CA 416  
mutation-associated gene signatures (by PIK3CAGS). Seven of the 50 genes are related 417  
to breast cancer histologic grade (by GGI). A significant portion of the top 50 ranked 418  
genes of DGBCox is overlapped with the 50 gene signatures of rorS (hypergeometric 419  
p-value:  $6e-11$ ). Eleven genes are common in the top 50 ranked genes of DGBCox and 420  
the 127 genes identified by Robust. Based on the curated list of breast cancer signatures 421  
by Huang *et al.* [46], 25 of the top 50 ranked genes have been used for breast cancer 422  
prognosis in previous studies. 423

We then compare the top 50 ranked genes of DGBCox with cancer causal genes to 424  
demonstrate the quality of DGBCox and its findings. It has been confirmed that cancer 425  
development and progression are caused by gene mutations [47], and these genes are 426  
called cancer causal genes or cancer drivers. We observe that three of the top 50 ranked 427  
genes of DGBCox appear in the Cancer Gene Census (CGC) [47], which is a manually 428  
curated list of likely cancer drivers. The full list of the top 50 ranked genes and their 429  
overlapping with signatures and cancer drivers can be found in S1 Table. 430

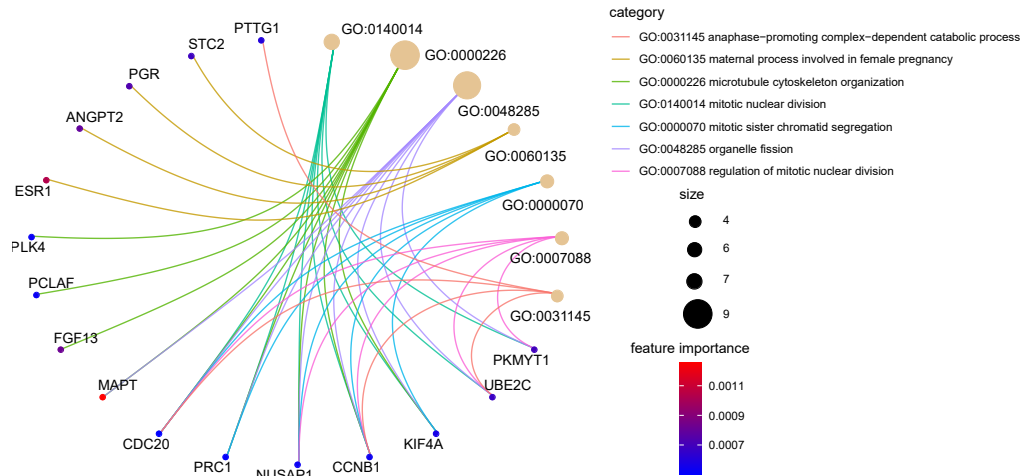
To investigate which biological mechanisms are important for stable breast cancer 431  
prognosis, we conduct Gene Ontology (GO) and Kyoto Encyclopedia of Genes and 432



**Fig 5. Overlapping genes between DGBCox and the existing methods.** The bottom left bar shows the number of signatures in each method. The dotted lines and the diagram on top show that the interaction overlaps between DGBCox and other methods.

Genomes (KEGG) pathway enrichment analysis (detailed in Section 5 in S1 File). As 433  
 shown in Fig 6, 16 out of the top 50 ranked genes of DGBCox are mainly enriched in 434  
 the cell cycle (GO:0140014, GO:0000070 and GO:0007088) and the organelle 435  
 organization (GO:0000226 and GO:0048285) processes, which are consistent with the 436  
 functions of most current breast cancer signatures [46]. We also observed some genes 437  
 are functionally enriched the maternal process involved in female pregnancy 438  
 (GO:0060135) or the anaphase-promoting complex-dependent catabolic process 439  
 (GO:0031145). These genes have been reported to be associated with breast cancer 440  
 development or metastasis. For example, *ESR1* and *PGR* are breast cancer biomarkers 441  
 that are commonly used for breast cancer subtype diagnosis [48]. *STC2* is reported to 442  
 be associated with breast cancer cell development [49]. *ANGPT2* is reported to be 443  
 capable of promoting breast cancer metastasis [50]. *PTTG1* is known as an oncogene 444  
 that promotes cancer cell proliferation [51], migration, and invasion [52] as well as 445  
 associated with the epithelial-mesenchymal transition (EMT) in breast cancer [53]. The 446  
 top 50 ranked genes of DGBCox are significantly enriched (adjusted p-value less than 447

0.01) in the oocyte meiosis pathway (KEGG ID: hsa04114, as shown in S3 Fig. Several studies have suggested that this pathway plays an important role in the development of breast cancer [54,55], however, the underlying biological mechanism needs to be verified by further laboratory and clinical researches.



**Fig 6. The non-redundant GO terms contained 16 of the top 50 ranked genes.** The genes (illustrated by dots) are coloured by their values of permutation feature importance. The size of the circle for a GO term is proportional to the number of genes annotated to this term.

## Conclusion

In this study, we proposed DGBCox to improve the stability of cancer prognosis across unknown environments. To the best of our knowledge, this is the first method designed for the stable cancer prognosis task. Results on 12 independent test datasets demonstrated that DGBCox is effective and stable for breast cancer prognosis across different environments.

We further analyzed which genes are important for the stable breast cancer prognosis. Genes were ranked by the proposed permutation importance algorithm. Interestingly, 25 of the top 50 ranked genes have been used for breast cancer prognosis in previous studies. These 50 genes are significantly enriched in seven non-redundant GO terms and one KEGG pathway which have been reported to be related to the development of breast cancer. The important genes help us understand the molecular mechanism of breast cancer development and also potentially be exploited to improve precision medicine for breast cancer treatment. We believe DGBCox is a useful method

in predicting breast cancer prognosis and it will aid researchers in other cancer research fields. 466  
467

## Supporting information 468

**S1 File. Supplementary information.** 469

**S1 Table. The top 50 ranked genes in DGBCox and their overlapping with other studies.** 1: An Entrez ID is a gene's identifier at the National Center for Biotechnology Information (NCBI)'s database. 2: The gene has been included in Cancer Gene Census (CGC) database [47], which indicates the dysfunction of this gene may cause cancer. 3: The gene has been used for breast cancer prognosis in previous studies based on the list curated by Huang *et al.* [46]. 470  
471  
472  
473  
474  
475

**S1 Fig. Comparison of different methods for risk group prediction.** The left block holds the hazard ratios of methods on each test dataset. The x-axis stands for the dataset, and the y-axis stands for the method. The hazard ratios are coloured from white to steel blue. The best result for each data is highlighted in bold. The right block displays  $Stability_{hr}$  of methods across the test datasets. The x-axis shows the  $Stability_{hr}$ , and the y-axis stands for the method. The values of  $Stability_{hr}$  are coloured from white to tomato colour. 476  
477  
478  
479  
480  
481  
482

**S2 Fig. Comparison of different methods based on the p-values by the Log-rank tests.** The x-axis stands for the dataset, and the y-axis stands for the method. The steel blue tile means that the p-value of a method is less than 0.05, which indicates that the method can successfully stratify patients in the dataset into two risk groups with distinct survival patterns. The grey tile indicates the opposite. 483  
484  
485  
486  
487

**S3 Fig. The oocyte meiosis pathway (hsa04114) contains six important genes for DGBCox.** These six important genes are highlighted by red colour. 488  
489

## Acknowledgments

490

This study makes use of data generated by the Molecular Taxonomy of Breast Cancer International Consortium. Funding for the project was provided by Cancer Research UK and the British Columbia Cancer Agency Branch.

491

492

493

## Funding

494

TDL was supported by the ARC DECRA Grant (Grant Number: DE200100200). This work has been partially supported by the ARC Discovery Project DP170101306.

495

496

## References

1. Park Y, Lee S, Cho E, Choi YL, Lee J, Nam S, et al. Clinical relevance of TNM staging system according to breast cancer subtypes. *Annals of Oncology*. 2011;22(7):1554–1560.
2. Cox DR. Regression models and life-tables. In: *Breakthroughs in statistics*. Springer; 1992. p. 527–541.
3. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology*. 2009;27(8):1160.
4. Loi S, Haibe-Kains B, Desmedt C, Wirapati P, Lallemand F, Tutt AM, et al. Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen. *BMC genomics*. 2008;9(1):239.
5. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New England Journal of Medicine*. 2004;351(27):2817–2826.
6. Haibe-Kains B, Desmedt C, Rothé F, Piccart M, Sotiriou C, Bontempi G. A fuzzy gene expression-based computational approach improves breast cancer prognostication. *Genome biology*. 2010;11(2):R18.



7. Filipits M, Rudas M, Jakesz R, Dubsy P, Fitzal F, Singer CF, et al. A new molecular predictor of distant recurrence in ER-positive, HER2-negative breast cancer adds independent information to conventional clinical risk factors. *Clinical Cancer Research*. 2011;17(18):6012–6020.
8. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox’s proportional hazards model via coordinate descent. *Journal of statistical software*. 2011;39(5):1.
9. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS, et al. Random survival forests. *The annals of applied statistics*. 2008;2(3):841–860.
10. Khan FM, Zubek VB. Support vector regression for censored data (SVRc): a novel tool for survival analysis. In: 2008 Eighth IEEE International Conference on Data Mining. IEEE; 2008. p. 863–868.
11. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC medical research methodology*. 2018;18(1):1–12.
12. Wang P, Li Y, Reddy CK. Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)*. 2019;51(6):1–36.
13. van Vliet MH, Reyal F, Horlings HM, van de Vijver MJ, Reinders MJ, Wessels LF. Pooling breast cancer datasets has a synergetic effect on classification performance and improves signature stability. *BMC genomics*. 2008;9(1):375.
14. Abraham G, Kowalczyk A, Loi S, Haviv I, Zobel J. Prediction of breast cancer prognosis using gene set statistics provides signature stability and biological context. *BMC bioinformatics*. 2010;11(1):277.
15. Li Y, Wang L, Wang J, Ye J, Reddy CK. Transfer learning for survival analysis via efficient l2, 1-norm regularized cox regression. In: 2016 IEEE 16th International Conference on Data Mining (ICDM). IEEE; 2016. p. 231–240.
16. Kim S, Kim K, Choe J, Lee I, Kang J. Improved survival analysis by learning shared genomic information from pan-cancer data. *Bioinformatics*. 2020;36(Supplement\_1):i389–i398.

17. Li J, Liu L, Le TD, Liu J. Accurate data-driven prediction does not mean high reproducibility. *Nature Machine Intelligence*. 2020;2(1):13–15.
18. Ma C, Gu C, Huo Y, Li X, Luo XJ. The integrated landscape of causal genes and pathways in schizophrenia. *Translational psychiatry*. 2018;8(1):1–14.
19. Kuang K, Cui P, Athey S, Xiong R, Li B. Stable prediction across unknown environments. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*; 2018. p. 1617–1626.
20. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*. 2012;486(7403):346.
21. Desmedt C, Piette F, Loi S, Wang Y, Lallemand F, Haibe-Kains B, et al. Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clinical cancer research*. 2007;13(11):3207–3214.
22. Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, Smeds J, et al. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *Journal of the National Cancer Institute*. 2006;98(4):262–272.
23. Miller LD, Smeds J, George J, Vega VB, Vergara L, Ploner A, et al. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proceedings of the National Academy of Sciences*. 2005;102(38):13550–13555.
24. Schmidt M, Böhm D, Von Törne C, Steiner E, Puhl A, Pilch H, et al. The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer research*. 2008;68(13):5405–5413.
25. Van't Veer LJ, Dai H, Van De Vijver MJ, He YD, Hart AA, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. *nature*. 2002;415(6871):530–536.

26. Van De Vijver MJ, He YD, Van't Veer LJ, Dai H, Hart AA, Voskuil DW, et al. A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*. 2002;347(25):1999–2009.
27. Loi S, Haibe-Kains B, Majjaj S, Lallemand F, Durbecq V, Larsimont D, et al. PIK3CA mutations associated with gene signature of low mTORC1 signaling and better outcomes in estrogen receptor–positive breast cancer. *Proceedings of the National Academy of Sciences*. 2010;107(22):10208–10213.
28. Bos PD, Zhang XHF, Nadal C, Shu W, Gomis RR, Nguyen DX, et al. Genes that mediate breast cancer metastasis to the brain. *Nature*. 2009;459(7249):1005–1009.
29. Li Y, Zou L, Li Q, Haibe-Kains B, Tian R, Li Y, et al. Amplification of LPTM4B and YWHAZ contributes to chemotherapy resistance and recurrence of breast cancer. *Nature medicine*. 2010;16(2):214–218.
30. Dedeurwaerder S, Desmedt C, Calonne E, Singhal SK, Haibe-Kains B, Defrance M, et al. DNA methylation profiling reveals a predominant immune component in breast cancers. *EMBO molecular medicine*. 2011;3(12):726–741.
31. Sabatier R, Finetti P, Cervera N, Lambaudie E, Esterni B, Mamessier E, et al. A gene expression signature identifies two prognostic subgroups of basal breast cancer. *Breast cancer research and treatment*. 2011;126(2):407–420.
32. Levine DA. Integrated genomic characterization of endometrial carcinoma. *Nature*. 2013;497(7447):67–73.
33. Buffa FM, Camps C, Winchester L, Snell CE, Gee HE, Sheldon H, et al. microRNA associated progression pathways and potential therapeutic targets identified by integrated mRNA and microRNA expression profiling in breast cancer. *Cancer research*. 2011; p. canres–0489.
34. Muranen TA, Greco D, Fagerholm R, Kilpivaara O, Kämpjärvi K, Aittomäki K, et al. Breast tumors from CHEK2 1100delC-mutation carriers: genomic landscape and clinical implications. *Breast Cancer Research*. 2011;13(5):R90.
35. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. *science*. 2006;313(5786):504–507.

36. Peters J, Bühlmann P, Meinshausen N. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*. 2016; p. 947–1012.
37. Imai K, Ratkovic M. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2014;76(1):243–263.
38. Anstrom KJ, Tsiatis AA. Utilizing propensity scores to estimate causal treatment effects with censored time-lagged data. *Biometrics*. 2001;57(4):1207–1218.
39. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*. 1958;53(282):457–481.
40. Gerds TA, Schumacher M. Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical Journal*. 2006;48(6):1029–1040.
41. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*. 1996;15(4):361–387.
42. Haibe-Kains B, Desmedt C, Sotiriou C, Bontempi G. A comparative study of survival models for breast cancer prognostication based on microarray data: does a single gene beat them all? *Bioinformatics*. 2008;24(19):2200–2208.
43. Bland JM, Altman DG. The logrank test. *Bmj*. 2004;328(7447):1073.
44. Bergstra J, Yamins D, Cox D. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In: *International conference on machine learning*. PMLR; 2013. p. 115–123.
45. Desmedt C, Haibe-Kains B, Wirapati P, Buyse M, Larsimont D, Bontempi G, et al. Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. *Clinical cancer research*. 2008;14(16):5158–5165.
46. Huang S, Murphy L, Xu W. Genes and functions from breast cancer signatures. *BMC cancer*. 2018;18(1):473.

47. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, et al. A census of human cancer genes. *Nature reviews cancer*. 2004;4(3):177–183.
48. Balic M, Thomssen C, Würtle R, Gnant M, Harbeck N. St. Gallen/Vienna 2019: a brief summary of the consensus discussion on the optimal primary breast cancer treatment. *Breast Care*. 2019;14(2):103–110.
49. Raulic S, Ramos-Valdes Y, DiMattia GE. Stanniocalcin 2 expression is regulated by hormone signalling and negatively affects breast cancer cell viability in vitro. *Journal of endocrinology*. 2008;197(3):517–530.
50. Imanishi Y, Hu B, Jarzynka MJ, Guo P, Elishaev E, Bar-Joseph I, et al. Angiopoietin-2 stimulates breast cancer metastasis through the  $\alpha 5\beta 1$  integrin-mediated pathway. *Cancer research*. 2007;67(9):4254–4263.
51. Huang S, Liao Q, Li L, Xin D. PTTG1 inhibits SMAD3 in prostate cancer cells to promote their proliferation. *Tumor Biology*. 2014;35(7):6265–6270.
52. Li H, Yin C, Zhang B, Sun Y, Shi L, Liu N, et al. PTTG1 promotes migration and invasion of human non-small cell lung cancer cells and is modulated by miR-186. *Carcinogenesis*. 2013;34(9):2145–2155.
53. Yoon CH, Kim MJ, Lee H, Kim RK, Lim EJ, Yoo KC, et al. PTTG1 oncogene promotes tumor malignancy via epithelial to mesenchymal transition and expansion of cancer stem cell population. *Journal of Biological Chemistry*. 2012;287(23):19516–19527.
54. Wu D, Han B, Guo L, Fan Z. Molecular mechanisms associated with breast cancer based on integrated gene expression profiling by bioinformatics analysis. *Journal of Obstetrics and Gynaecology*. 2016;36(5):615–621.
55. Liu H, Ye H. Screening of the prognostic targets for breast cancer based co-expression modules analysis. *Molecular Medicine Reports*. 2017;16(4):4038–4044.