# SUPPLEMENTARY DATA

## ACRONYMS

| Acronym | Description |
|---|---|
| AOL | Adjuvant! Online. |
| AUC | Area Under the Curve. |
| AURKA | Proliferation gene having the gene symbol AURKA or STK6. |
| BC | Breast Cancer. |
| BSC | Brier SCore. |
| $C$-index | Concordance index. |
| CV | Cross-Validation. |
| ER | Estrogen Receptor. |
| GENE76 | Gene classifier from (Wang et al., 2005). |
| GEO | Gene Expression Omnibus. |
| GGI | Gene expression Grade Index. |
| HER2 | gene having the gene symbol HER2 or ERBB2. |
| HR | Hazard Ratio. |
| IAUC | Integrated Area Under the Curve. |
| IBSC | Integrated Brier SCore. |
| KM | Kaplan-Meier null model. |
| NPI | Nottingham Prognostic Index. |
| ROC | Receiver Operating Characteristic. |
| TBG | dataset from (Desmedt et al., 2007). |
| TTE | Time To Event. |
| TAM | dataset from (Loi et al., 2007; Haibe-Kains et al., 2008). |
| UPP | dataset from (Miller et al., 2005). |
| VDX | dataset from (Wang et al., 2005). |

Supplementary Table 1: List of acronyms.

## 1 DIMENSION REDUCTION USING CROSS-VALIDATION

In order to select a good signature size, i.e. a number of features used to fit the risk prediction model (see Section 2.2.1), we performed a cross-validated dimension reduction strategy composed of the following steps:

1. Ranking (univariate ranking or first principal components, see *Dimension reduction strategy* in Section 2.2) of all the features using all the samples. Although a nested cross-validation procedure, i.e. a new ranking generated within each fold of the cross-validation, reduce the bias in performance estimation (Varma and Simon, 2006), we computed a single ranking using all the samples because of computational cost. Indeed, due to the large number of features (22283 probes) and the large number of risk prediction methods (methods 6 to 11, see Table 1), a nested cross-validation procedure is not tractable.
2. Computation of a performance criterion using a cross-validation procedure for each signature size $k$. We used two performance criteria depending on the learning algorithm (see Section 2.2). For the WILCOXON learning algorithm, we used the $-\log_{10}$ p-value computed from a Wilcoxon rank sum test to test if the risk score is able to discriminate the patients with histological grade 1 and 3. For the COX and RCOX learning algorithms, we used the cross-validated partial likelihood (van Houwelingen et al., 2006) to estimate the partial likelihood of the risk score.
3. Selection of the best $k$ that is the best trade-off between performance and signature size. We select $k_{best}$ using the following formula

$$k_{best} = \underset{p}{\operatorname{argmax}} \left\{ p' - \alpha \frac{k}{t} \right\}$$

where $p$ is the performance estimate, $p' = \frac{p - \min(p)}{\max(p) - \min(p)}$, $\alpha$ is a penalty parameter fixed to 1 and $t$ is the total number of features, i.e. $t = 22283$ and $t =$ the number of samples in the training set, for RANK and PCA dimension reduction strategies respectively.

### 1.1 Signature Size

We estimated the performance using a 5-fold cross-validation for the signature size values from 1 to 45, and values from 50 to 200 spaced by 10. Due to the complexity of the learning algorithm RCOX, we did not estimate the performance for signature size values larger than 45.

| Model | Signature size |
|---|---|
| GW.RANKCV.COMBUNIV.WILCOXON.HG | 7 |
| GW.RANKCV.COMBUNIV.COX.SURV | 4 |
| GW.RANKCV.MULTIV.RCOX.SURV | 19 |
| GW.PCACV.COMBUNIV.WILCOXON.HG | 18 |
| GW.PCACV.COMBUNIV.COX.SURV | 34 |
| GW.PCACV.MULTIV.RCOX.SURV | 10 |

Supplementary Table 2: Signature size selected using cross-validated dimension reduction strategy for each method.

## 1.2 Specificity for Risk Score Prediction

| Model | VDX | TBG | TAM | UPP |
|---|---|---|---|---|
| GW.RANK.COMBUNIV.WILCOXON.HG | 0.258 | 0.373 | 0.277 | 0.227 |
| GW.RANKCV.COMBUNIV.WILCOXON.HG | 0.242 | 0.292 | 0.206 | 0.212 |
| GW.RANK.COMBUNIV.COX.SURV | 0.400 | 0.360 | 0.362 | 0.162 |
| GW.RANKCV.COMBUNIV.COX.SURV | 0.353 | 0.298 | 0.390 | 0.268 |
| GW.RANK.MULTIV.RCOX.SURV | 0.468 | 0.242 | 0.326 | 0.242 |
| GW.RANKCV.MULTIV.RCOX.SURV | 0.526 | 0.360 | 0.277 | 0.247 |
| GW.PCA.COMBUNIV.WILCOXON.HG | 0.147 | 0.298 | 0.067 | 0.091 |
| GW.PCACV.COMBUNIV.WILCOXON.HG | 0.147 | 0.280 | 0.110 | 0.116 |
| GW.PCA.COMBUNIV.COX.SURV | 0.426 | 0.379 | 0.450 | 0.217 |
| GW.PCACV.COMBUNIV.COX.SURV | 0.505 | 0.366 | 0.486 | 0.232 |
| GW.PCA.MULTIV.RCOX.SURV | 0.405 | 0.509 | 0.358 | 0.141 |
| GW.PCACV.MULTIV.RCOX.SURV | 0.300 | 0.373 | 0.319 | 0.106 |

AURKA and GGI models were not fitted on VDX. Therefore, this dataset can be considered as a validation set.

Supplementary Table 3: Specificity for a sensitivity of 90% for risk score prediction methods without (RANK and PCA) and with (RANKCV and PCACV) cross-validated dimension reduction strategy in the training set (VDX) and the three validation sets (TBG, TAM, and UPP).

## 1.3 Performance for Risk Score Prediction

| | $C$-index | | | | IAUC | | | | IBSC | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | VDX | TBG | TAM | UPP | VDX | TBG | TAM | UPP | VDX | TBG | TAM | UPP |
| GW.RANK.COMBUNIV.WILCOXON.HG | 0.619 | 0.624 | 0.691 | 0.653 | 0.639 | **0.617** | **0.684** | **0.662** | 0.182 | **0.141** | 0.131 | 0.146 |
| GW.RANKCV.COMBUNIV.WILCOXON.HG | 0.595 | 0.602 | 0.643 | 0.64 | 0.616 | 0.603 | 0.636 | **0.656** | 0.184 | 0.144 | 0.139 | 0.148 |
| GW.RANK.COMBUNIV.COX.SURV | **0.742** | **0.665** | 0.65 | 0.637 | **0.774** | **0.686** | 0.638 | **0.651** | **0.148** | 0.153 | 0.158 | 0.172 |
| GW.RANKCV.COMBUNIV.COX.SURV | **0.717** | 0.646 | 0.676 | 0.654 | **0.755** | **0.655** | 0.672 | **0.688** | **0.156** | **0.142** | 0.131 | **0.141** |
| GW.RANK.MULTIV.RCOX.SURV | **0.774** | 0.663 | 0.638 | 0.63 | **0.823** | **0.715** | 0.635 | **0.654** | **0.136** | 0.151 | 0.175 | 0.16 |
| GW.RANKCV.MULTIV.RCOX.SURV | **0.777** | **0.682** | 0.63 | 0.639 | **0.822** | **0.717** | 0.636 | **0.667** | **0.138** | **0.13** | 0.161 | 0.147 |
| GW.PCA.COMBUNIV.WILCOXON.HG | 0.586 | 0.591 | 0.566 | 0.579 | 0.617 | 0.616 | 0.565 | 0.561 | 0.186 | **0.14** | 0.136 | 0.148 |
| GW.PCACV.COMBUNIV.WILCOXON.HG | 0.596 | 0.583 | 0.575 | 0.588 | 0.63 | 0.605 | 0.577 | 0.57 | 0.185 | **0.142** | 0.137 | 0.15 |
| GW.PCA.COMBUNIV.COX.SURV | **0.726** | **0.676** | 0.695 | 0.594 | **0.749** | **0.705** | 0.672 | 0.589 | **0.154** | 0.147 | 0.153 | 0.177 |
| GW.PCACV.COMBUNIV.COX.SURV | **0.747** | **0.678** | 0.698 | 0.598 | **0.766** | **0.708** | 0.679 | 0.59 | **0.149** | 0.159 | 0.17 | 0.185 |
| GW.PCA.MULTIV.RCOX.SURV | **0.75** | **0.694** | 0.69 | 0.591 | **0.779** | **0.733** | 0.667 | 0.598 | **0.143** | 0.155 | 0.171 | 0.176 |
| GW.PCACV.MULTIV.RCOX.SURV | **0.686** | **0.684** | 0.657 | 0.566 | **0.721** | **0.721** | 0.629 | 0.564 | **0.161** | **0.141** | 0.148 | 0.176 |

AURKA and GGI models were not fitted on VDX. Therefore, this dataset can be considered as a validation set.

Supplementary Table 4: Performance for risk score prediction methods without (RANK and PCA) and with (RANKCV and PCACV) cross-validated dimension reduction strategy in the training set (VDX) and the three validation sets (TBG, TAM, and UPP). The accuracy measures in **bold** are significantly better than the accuracy of AURKA model. In case of IBSC, the accuracy measures of AURKA are in **bold** if they are significantly better than KM, the benchmark model.

## 1.4 Sensitivity and Specificity for Risk Group Prediction

| | Sensitivity | | | | Specificity | | | |
|---|---|---|---|---|---|---|---|---|
| Model | VDX | TBG | TAM | UPP | VDX | TBG | TAM | UPP |
| GW.RANK.COMBUNIV.WILCOXON.HG | 0.812 | 0.892 | 0.840 | 0.833 | 0.400 | 0.379 | 0.358 | 0.359 |
| GW.RANKCV.COMBUNIV.WILCOXON.HG | 0.760 | 0.892 | 0.840 | 0.806 | 0.374 | 0.379 | 0.358 | 0.354 |
| GW.RANK.COMBUNIV.COX.SURV | 0.885 | 0.892 | 0.860 | 0.778 | 0.437 | 0.379 | 0.362 | 0.348 |
| GW.RANKCV.COMBUNIV.COX.SURV | 0.885 | 0.865 | 0.900 | 0.861 | 0.437 | 0.373 | 0.369 | 0.364 |
| GW.RANK.MULTIV.RCOX.SURV | 0.927 | 0.892 | 0.840 | 0.806 | 0.458 | 0.379 | 0.358 | 0.354 |
| GW.RANKCV.MULTIV.RCOX.SURV | 0.948 | 0.892 | 0.860 | 0.861 | 0.468 | 0.379 | 0.362 | 0.364 |
| GW.PCA.COMBUNIV.WILCOXON.HG | 0.740 | 0.838 | 0.760 | 0.750 | 0.363 | 0.366 | 0.344 | 0.343 |
| GW.PCACV.COMBUNIV.WILCOXON.HG | 0.760 | 0.811 | 0.740 | 0.778 | 0.374 | 0.360 | 0.340 | 0.348 |
| GW.PCA.COMBUNIV.COX.SURV | 0.896 | 0.892 | 0.940 | 0.778 | 0.442 | 0.379 | 0.376 | 0.348 |
| GW.PCACV.COMBUNIV.COX.SURV | 0.917 | 0.892 | 0.960 | 0.806 | 0.453 | 0.379 | 0.379 | 0.354 |
| GW.PCA.MULTIV.RCOX.SURV | 0.896 | 0.919 | 0.880 | 0.722 | 0.442 | 0.385 | 0.365 | 0.338 |
| GW.PCACV.MULTIV.RCOX.SURV | 0.823 | 0.892 | 0.880 | 0.722 | 0.405 | 0.379 | 0.365 | 0.338 |

AURKA and GGI models were not fitted on VDX. Therefore, this dataset can be considered as a validation set.

Supplementary Table 5: Sensitivity and specificity for risk group prediction methods without (RANK and PCA) and with (RANKCV and PCACV) cross-validated dimension reduction strategy in the training set (VDX) and the three validation sets (TBG, TAM, and UPP).

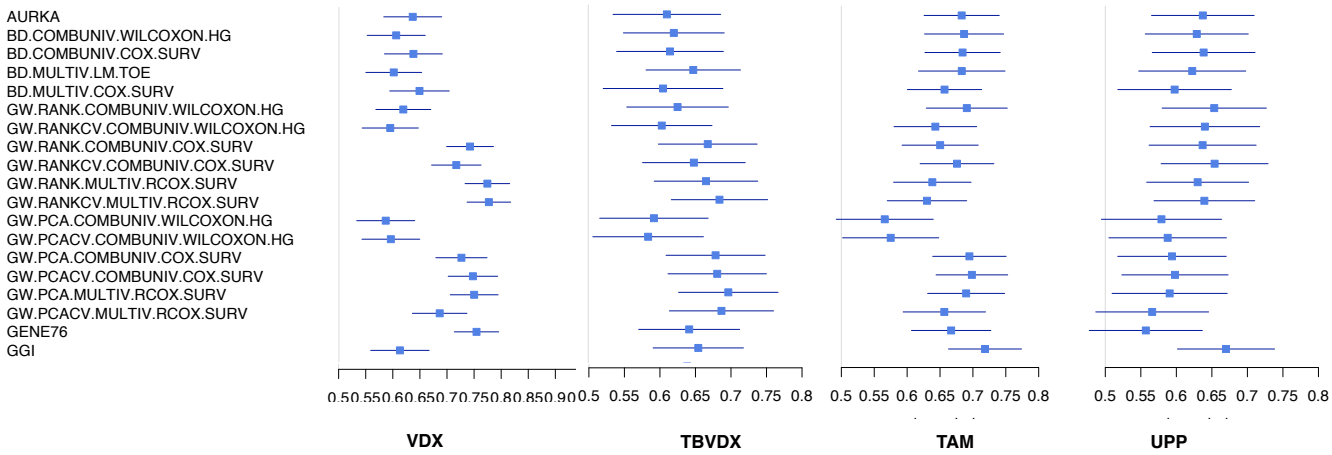
## 1.5 Performance for Risk Group Prediction

| | $C$-index | | | | HR | | | | IBSC | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | VDX | TBG | TAM | UPP | VDX | TBG | TAM | UPP | VDX | TBG | TAM | UPP |
| GW.RANK.COMBUNIV.WILCOXON.HG | 0.694 | 0.785 | 0.775 | 0.733 | 1.99 | 3.61 | 3.42 | 2.42 | 0.182 | **0.139** | 0.136 | **0.146** |
| GW.RANKCV.COMBUNIV.WILCOXON.HG | 0.642 | 0.749 | 0.728 | 0.703 | 1.64 | 2.77 | 2.7 | 2.05 | 0.185 | 0.14 | 0.137 | **0.146** |
| GW.RANK.COMBUNIV.COX.SURV | **0.836** | 0.77 | 0.778 | 0.632 | **4.69** | 2.96 | 3.53 | 1.53 | **0.168** | 0.143 | 0.139 | 0.156 |
| GW.RANKCV.COMBUNIV.COX.SURV | **0.824** | 0.727 | 0.816 | 0.718 | **4.04** | 2.57 | 4.26 | 2.13 | **0.17** | 0.146 | 0.133 | 0.149 |
| GW.RANK.MULTIV.RCOX.SURV | **0.906** | 0.765 | 0.749 | 0.696 | **9.62** | 3.28 | 3 | 2.18 | **0.159** | 0.15 | 0.147 | 0.157 |
| GW.RANKCV.MULTIV.RCOX.SURV | **0.938** | 0.793 | 0.766 | 0.735 | **14.1** | 3.61 | 3.05 | 2.52 | **0.155** | 0.151 | 0.148 | 0.158 |
| GW.PCA.COMBUNIV.WILCOXON.HG | 0.616 | 0.69 | 0.589 | 0.586 | 1.46 | 1.94 | 1.3 | 1.37 | 0.187 | 0.142 | 0.14 | 0.15 |
| GW.PCACV.COMBUNIV.WILCOXON.HG | 0.633 | 0.652 | 0.582 | 0.617 | 1.55 | 1.67 | 1.25 | 1.64 | 0.186 | 0.143 | 0.141 | 0.15 |
| GW.PCA.COMBUNIV.COX.SURV | **0.843** | 0.734 | 0.909 | 0.63 | **5.13** | 2.62 | 9.5 | 1.53 | **0.167** | 0.147 | 0.133 | 0.174 |
| GW.PCACV.COMBUNIV.COX.SURV | **0.868** | 0.745 | 0.907 | 0.691 | **6.4** | 2.67 | 7.52 | 1.82 | **0.164** | 0.148 | 0.137 | 0.173 |
| GW.PCA.MULTIV.RCOX.SURV | **0.826** | 0.749 | 0.818 | 0.564 | **4.3** | 2.6 | 4.64 | 1.15 | **0.169** | 0.142 | 0.136 | 0.177 |
| GW.PCACV.MULTIV.RCOX.SURV | 0.754 | 0.722 | 0.789 | 0.592 | 2.97 | 2.34 | 3.69 | 1.46 | **0.178** | 0.141 | 0.136 | 0.154 |

AURKA and GGI models were not fitted on VDX. Therefore, this dataset can be considered as a validation set.

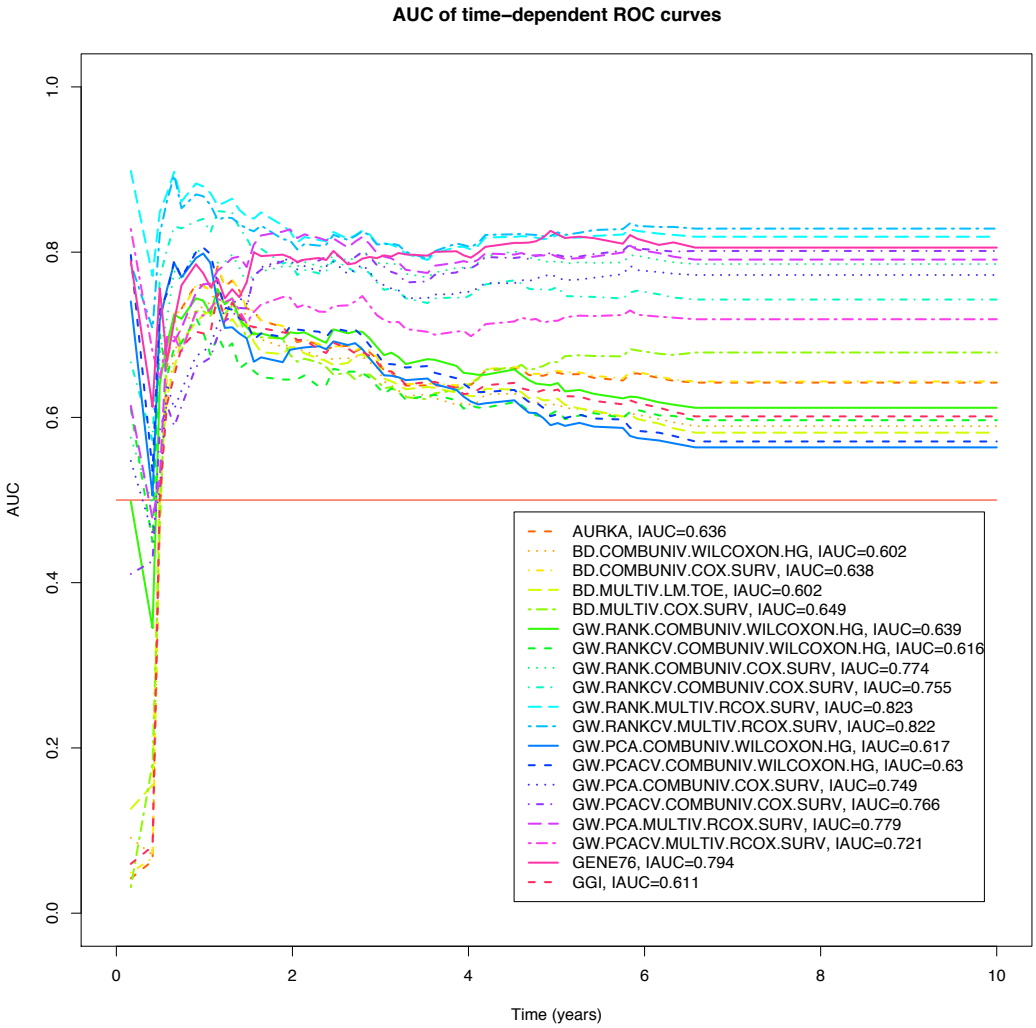
Supplementary Table 6: Performance for risk group prediction methods without (RANK and PCA) and with (RANKCV and PCACV) cross-validated dimension reduction strategy in the training set (VDX) and the three validation sets (TBG, TAM, and UPP). The accuracy measures in **bold** are significantly better than the accuracy of AURKA model. In case of IBSC, the accuracy measures of AURKA are in **bold** if they are significantly better than KM, the benchmark model.

## 2 CONCORDANCE INDEX FOR RISK SCORE PREDICTION

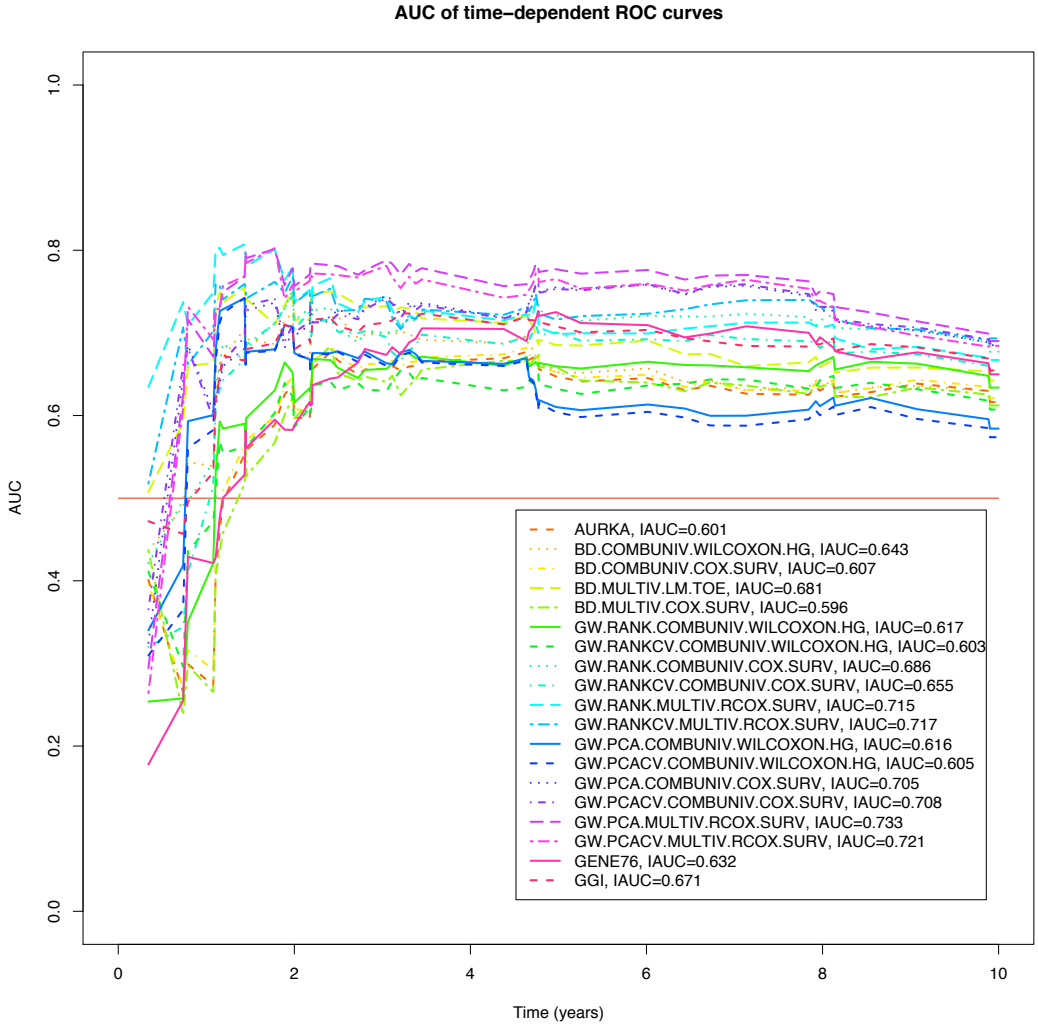



Supplementary Figure 1: Forest plot of the concordance indices for the risk scores predicted by all the methods in the training set (VDX) and in the three validation sets (TBG, TAM, and UPP). AURKA and GGI models were not fitted on VDX which can be considered as a validation set for these models.

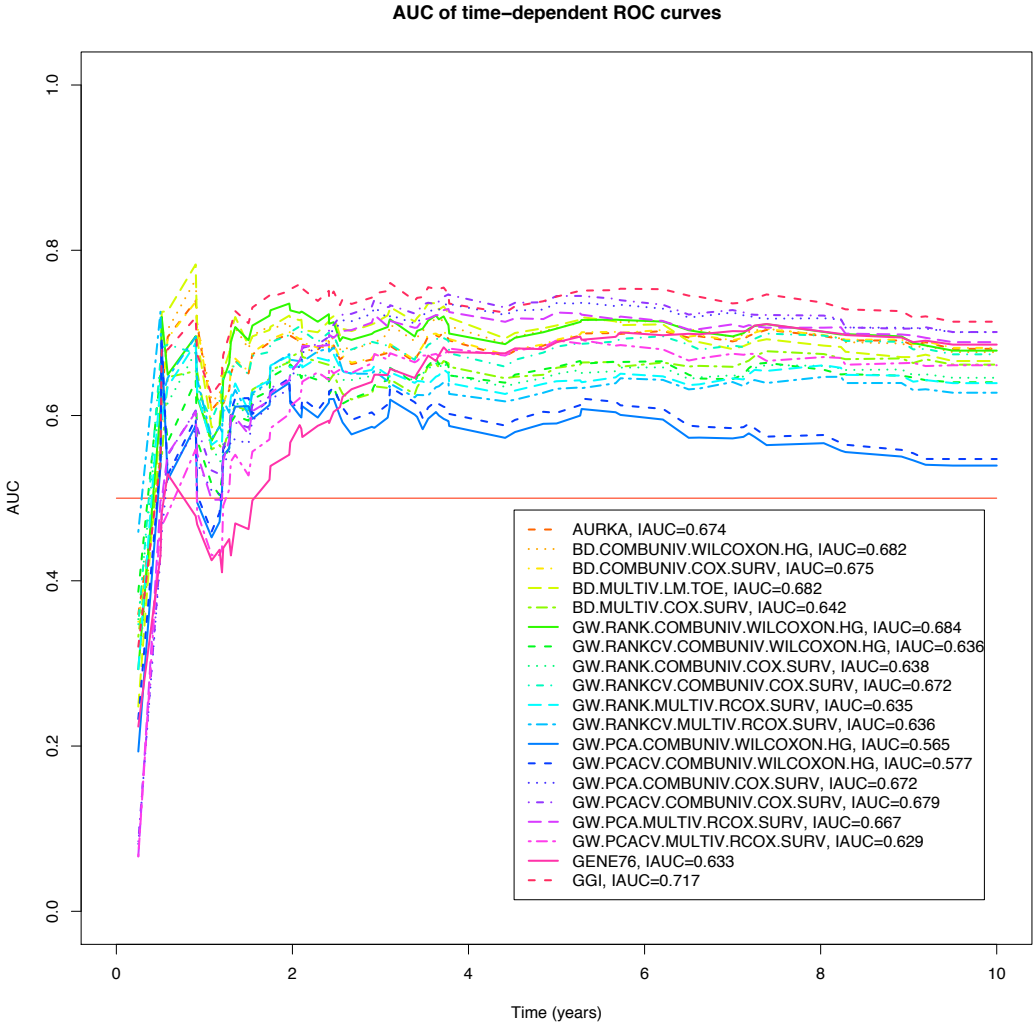
## 3 TIME-DEPENDENT ROC CURVE FOR RISK SCORE PREDICTION



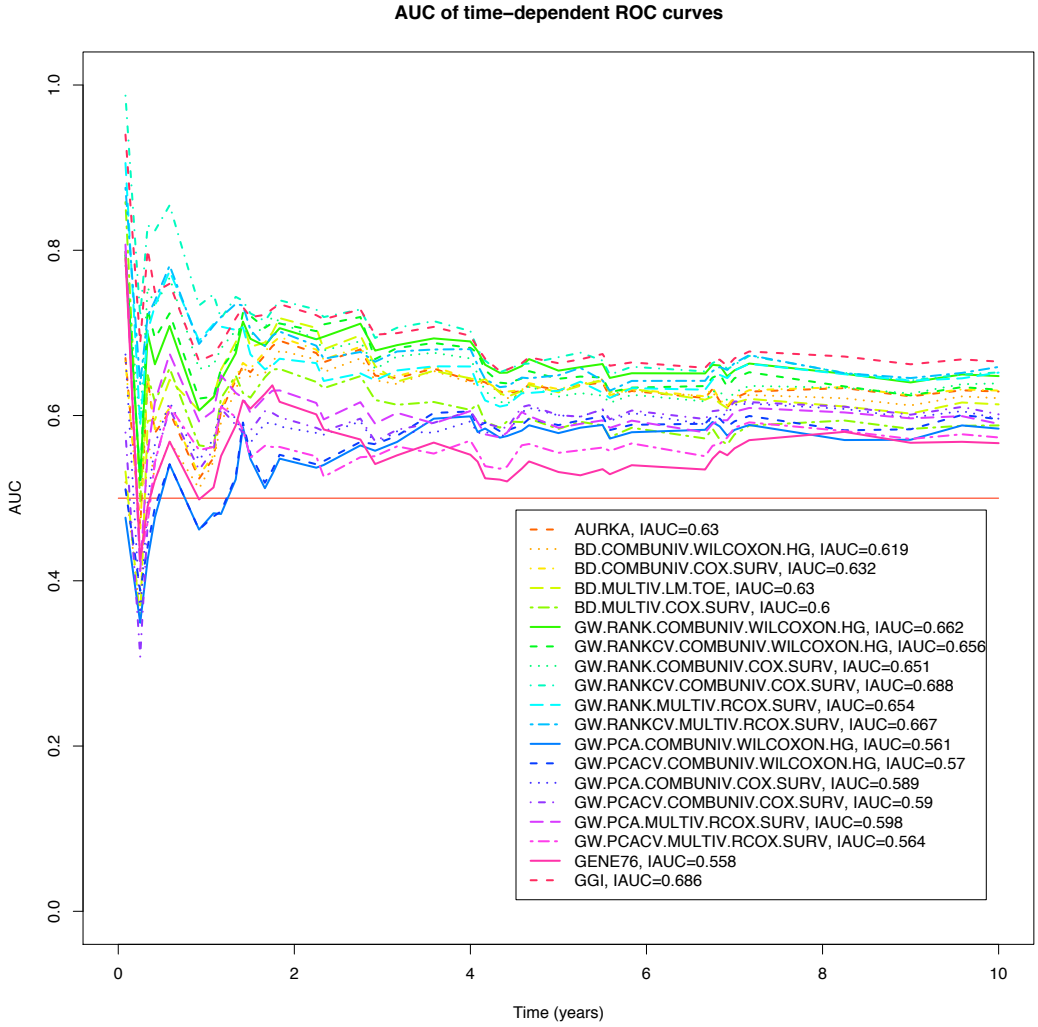

Supplementary Figure 2: AUC of the time-dependent ROC curve with respect to the time for the risk scores predicted by all the methods in the training set (VDX). AURKA and GGI models were not fitted on VDX which can be considered as a validation set for these models. The IAUC is given in the legend.




Supplementary Figure 3: AUC of the time-dependent ROC curve with respect to the time for the risk scores predicted by all the methods in the TBG validation set. The IAUC is given in the legend.

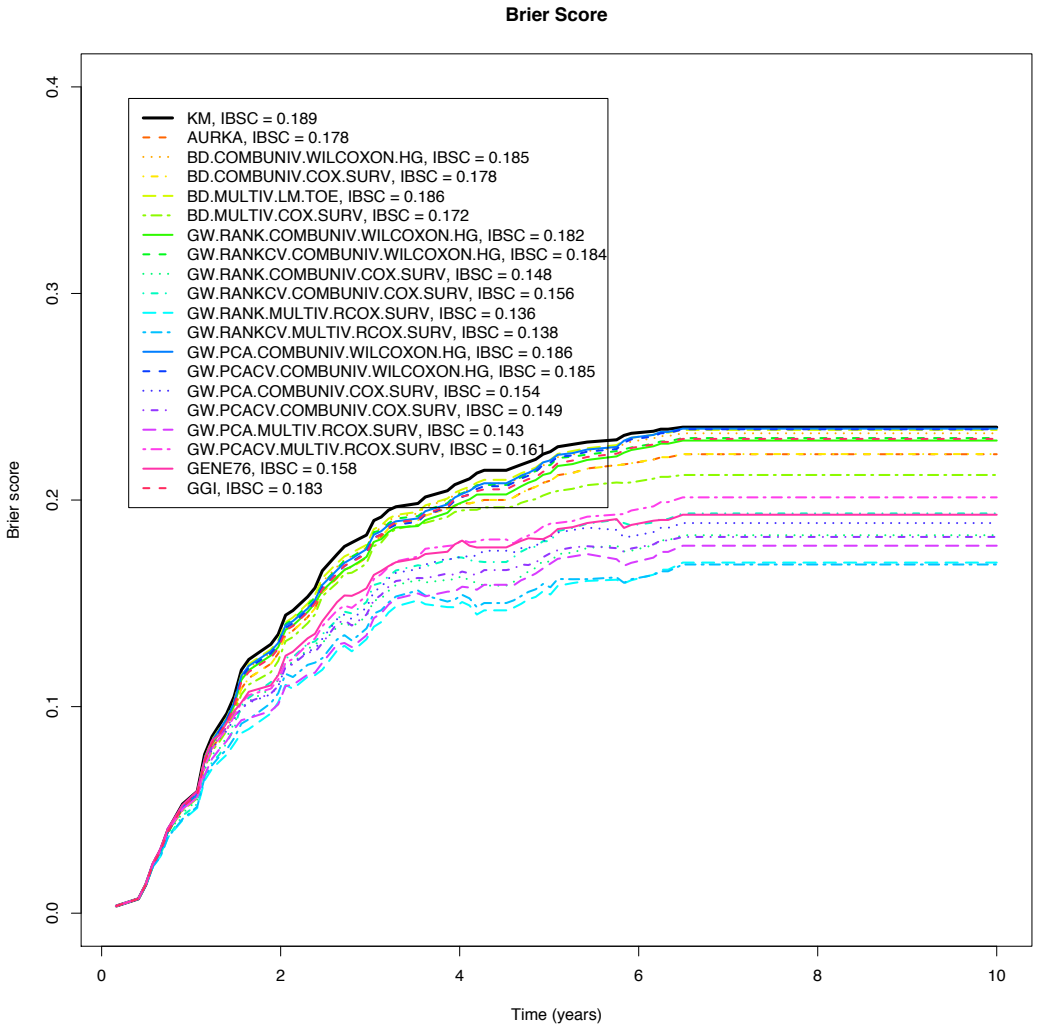



Supplementary Figure 4: AUC of the time-dependent ROC curve with respect to the time for the risk scores predicted by all the methods in the TAM validation set. The IAUC is given in the legend.

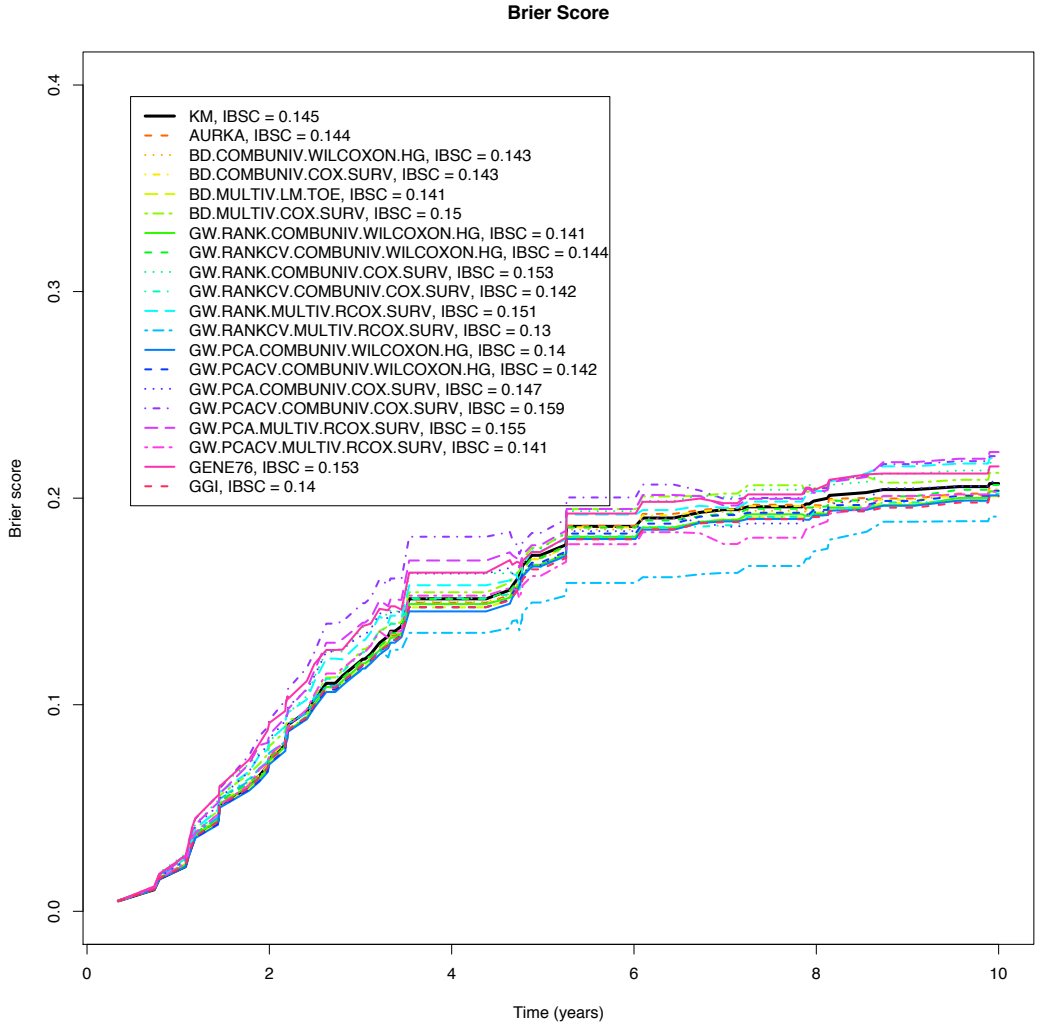



Supplementary Figure 5: AUC of the time-dependent ROC curve with respect to the time for the risk scores predicted by all the methods in the UPP validation set. The IAUC is given in the legend.

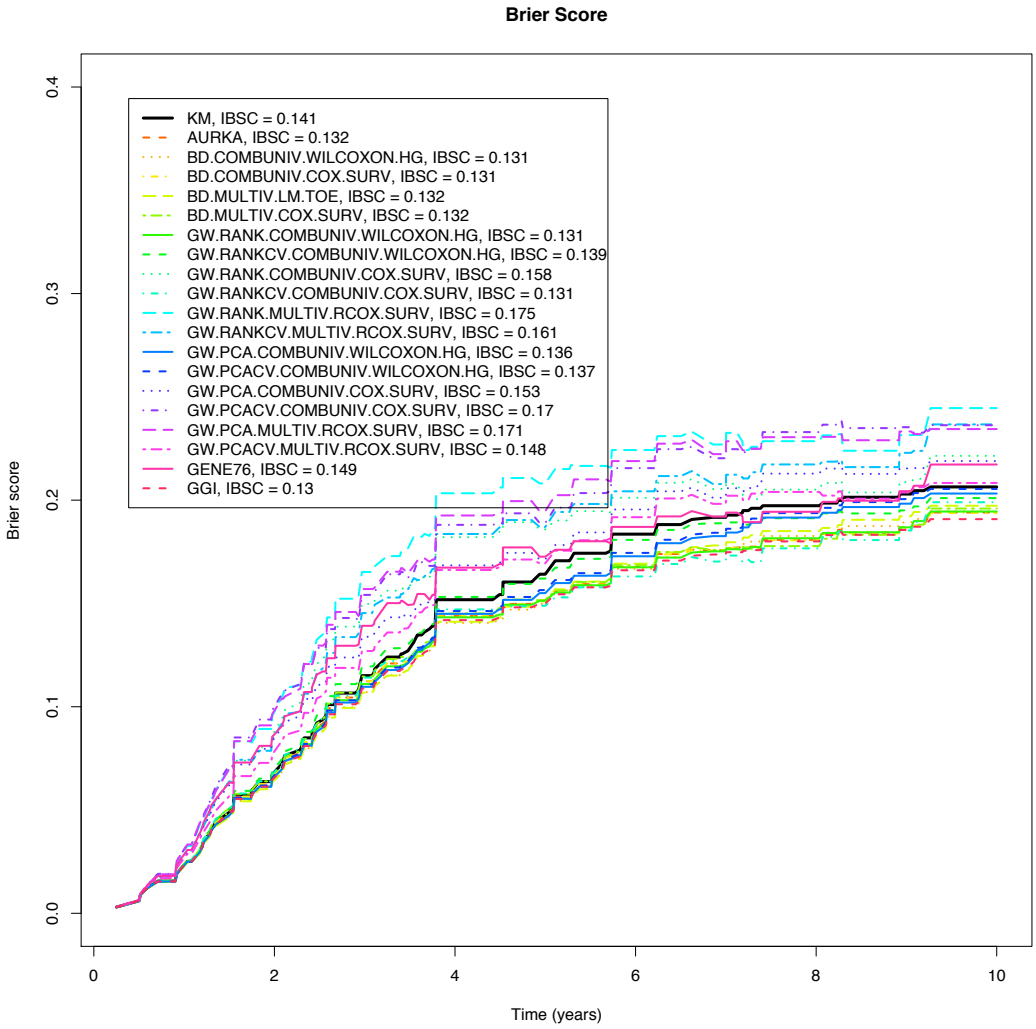
## 4 BRIER SCORE FOR RISK SCORE PREDICTION



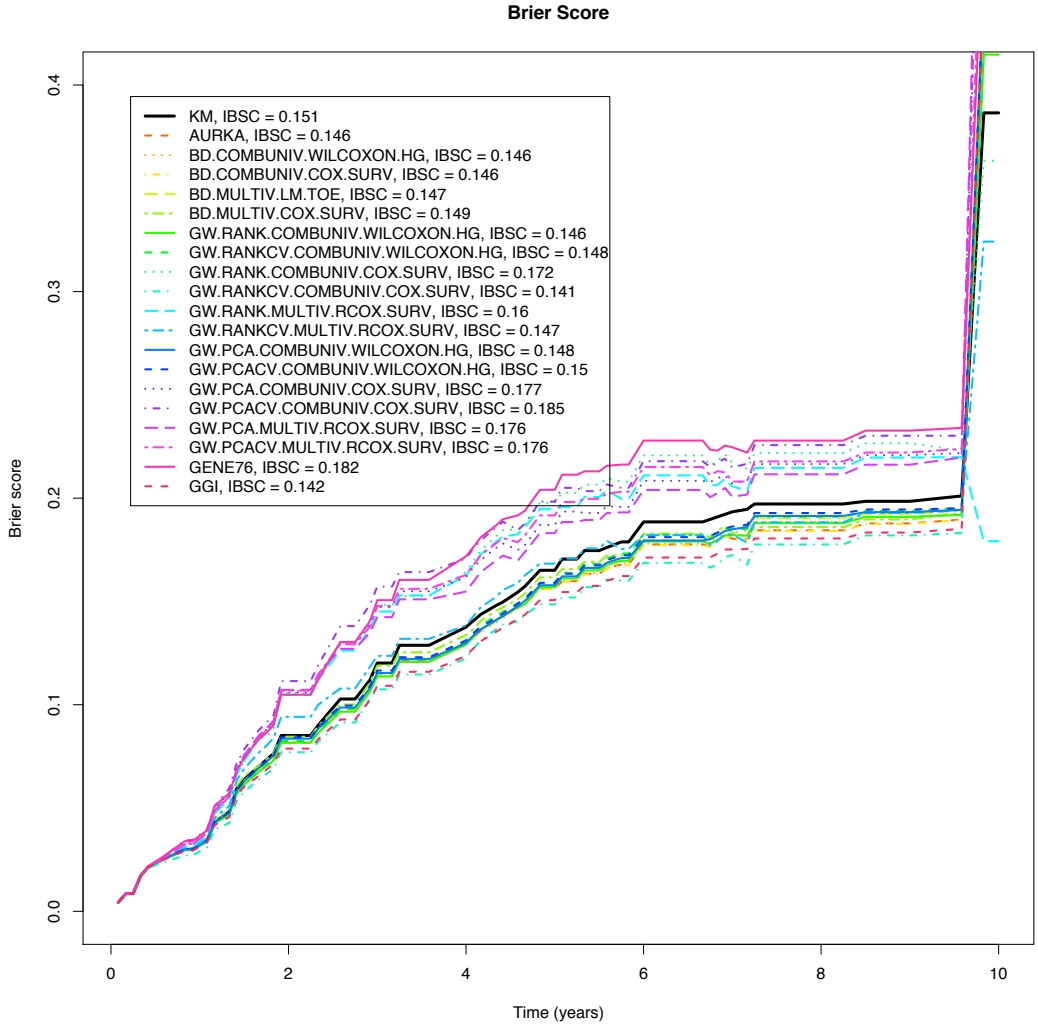

Supplementary Figure 6: Brier scores with respect to the time for the risk scores predicted by all the methods in the training set (VDX). AURKA and GGI models were not fitted on VDX which can be considered as a validation set for these models. The IBSC is given in the legend.




Supplementary Figure 7: Brier scores with respect to the time for the risk scores predicted by all the methods in the TBG validation set. The IBSC is given in the legend.

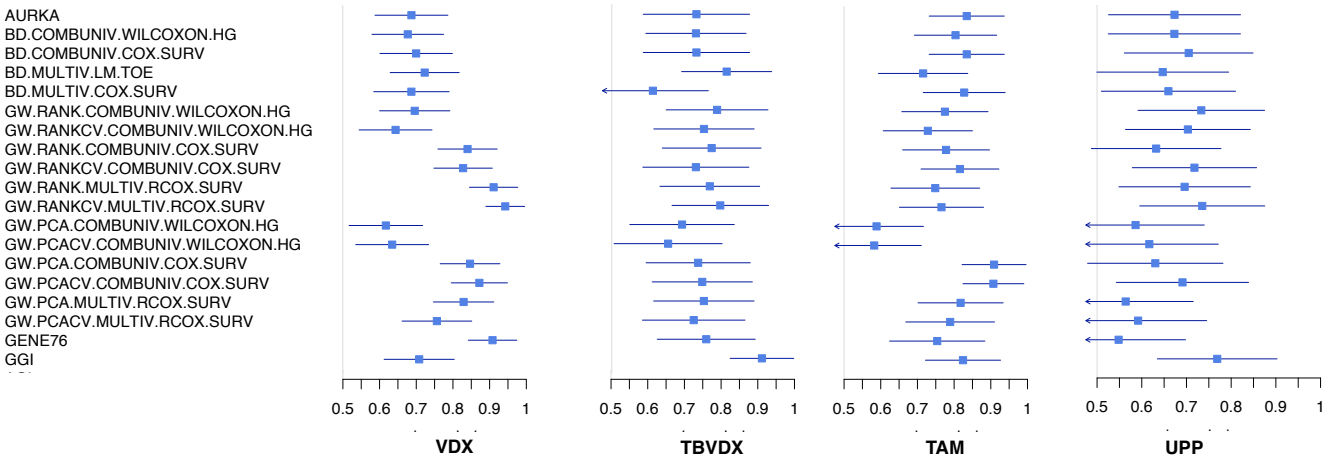



Supplementary Figure 8: Brier scores with respect to the time for the risk scores predicted by all the methods in the TAM validation set. The IBSC is given in the legend.



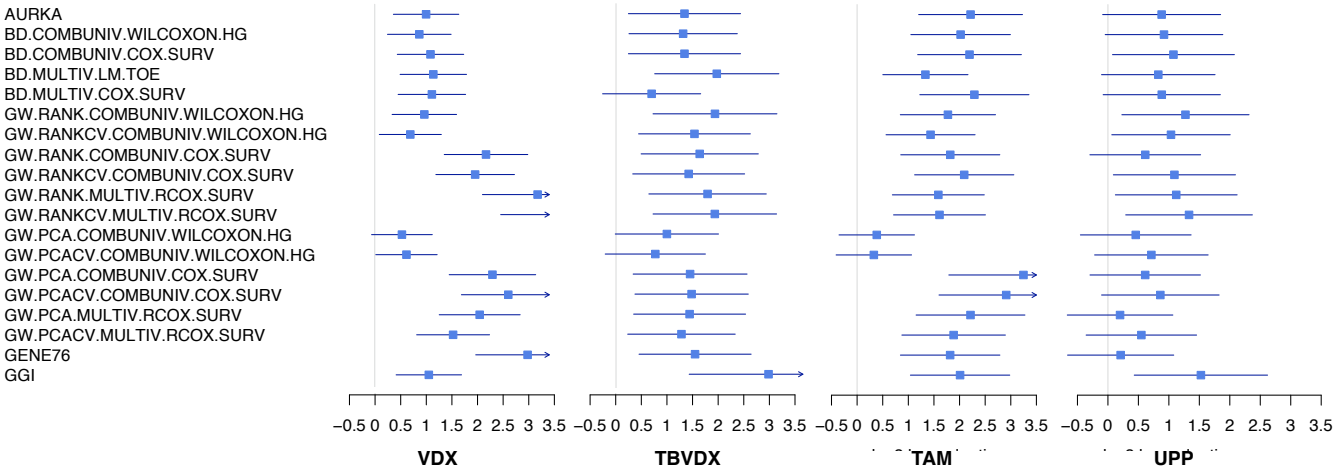

Supplementary Figure 9: Brier scores with respect to the time for the risk scores predicted by all the methods in the UPP validation set. The IBSC is given in the legend.

## 5 CONCORDANCE INDEX FOR RISK GROUP PREDICTION



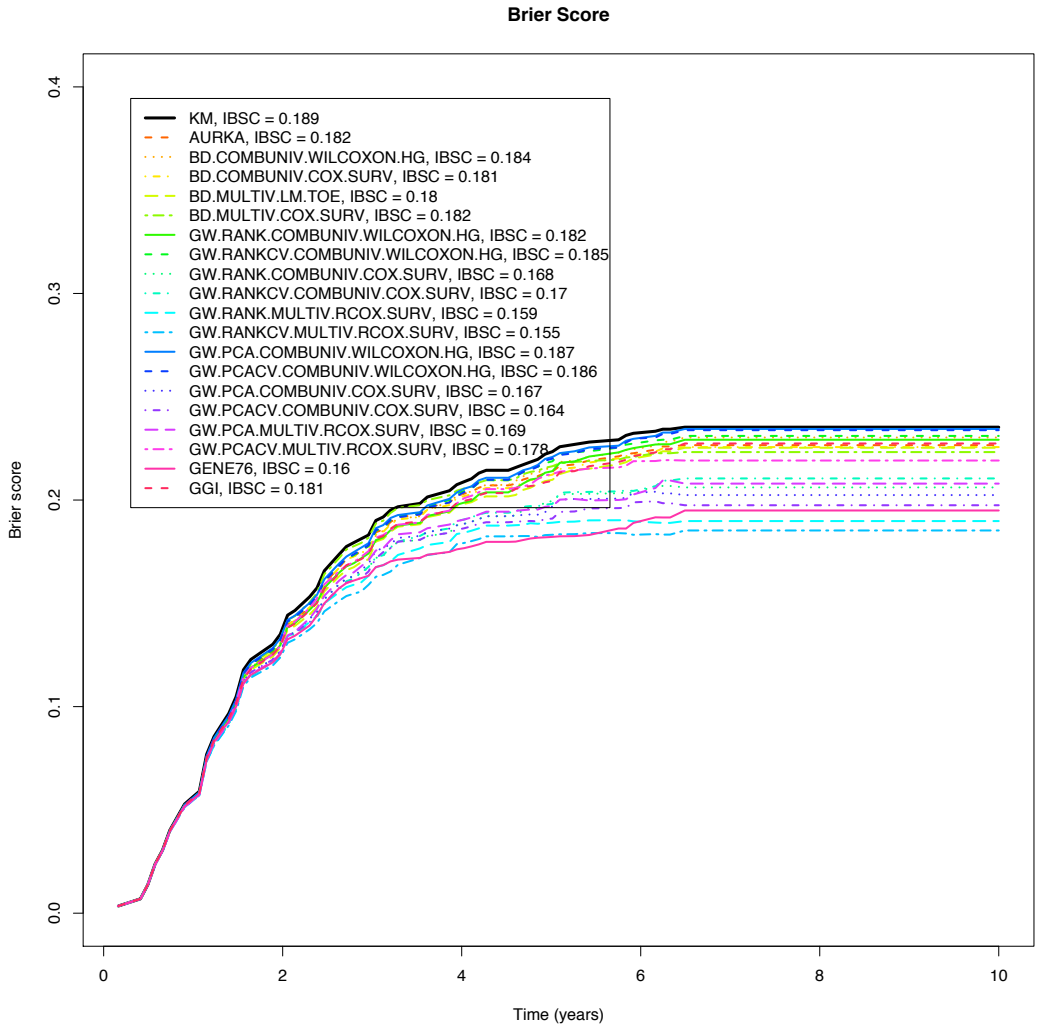

Supplementary Figure 10: Forest plot of the concordance indices for the risk groups predicted by all the methods in the training set (VDX) and in the three validation sets (TBG, TAM, and UPP). AURKA and GGI models were not fitted on VDX which can be considered as a validation set for these models.

## 6 HAZARD RATIO FOR RISK GROUP PREDICTION

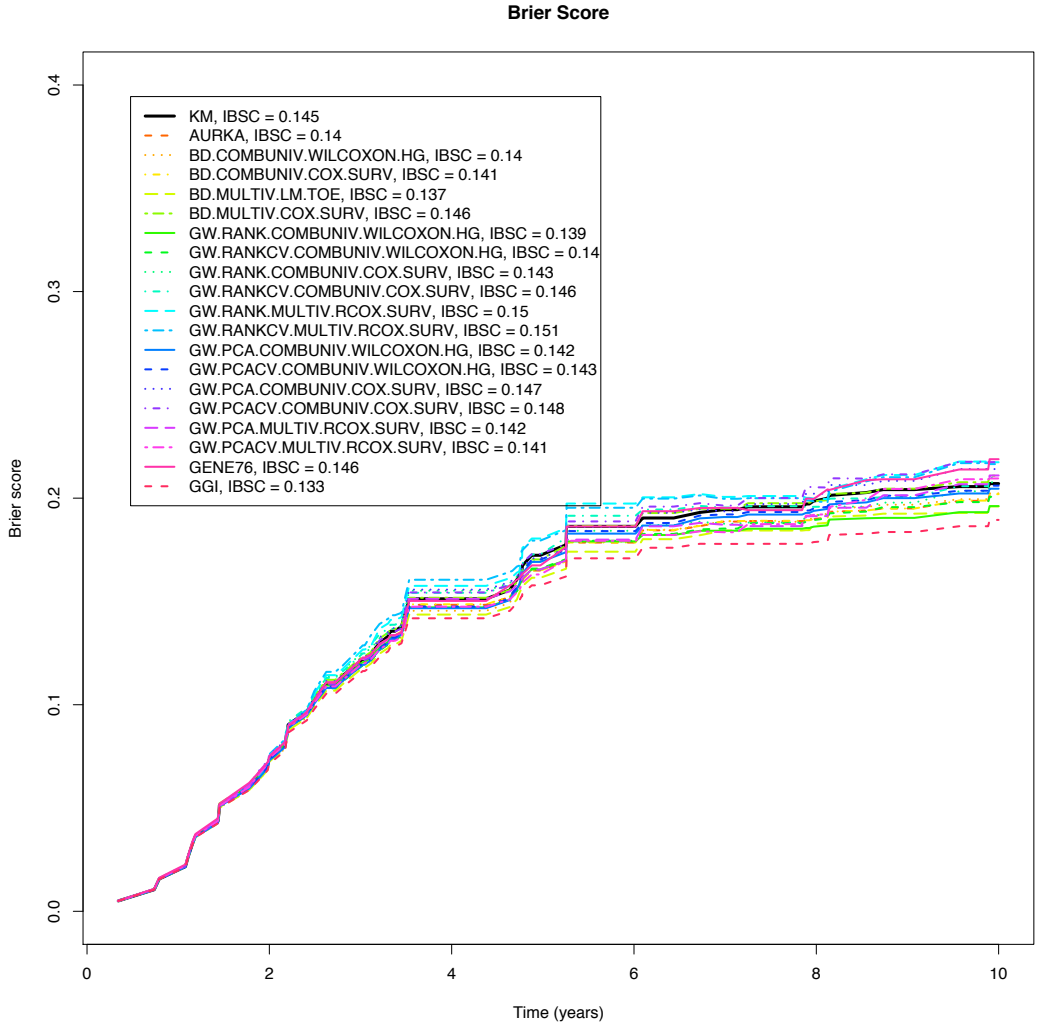



Supplementary Figure 11: Forest plot of the $\log_2$ hazard ratios for the the risk groups predicted by all the methods in the training set (VDX) and in the three validation sets (TBG, TAM, and UPP). AURKA and GGI models were not fitted on VDX which can be considered as a validation set for these models.

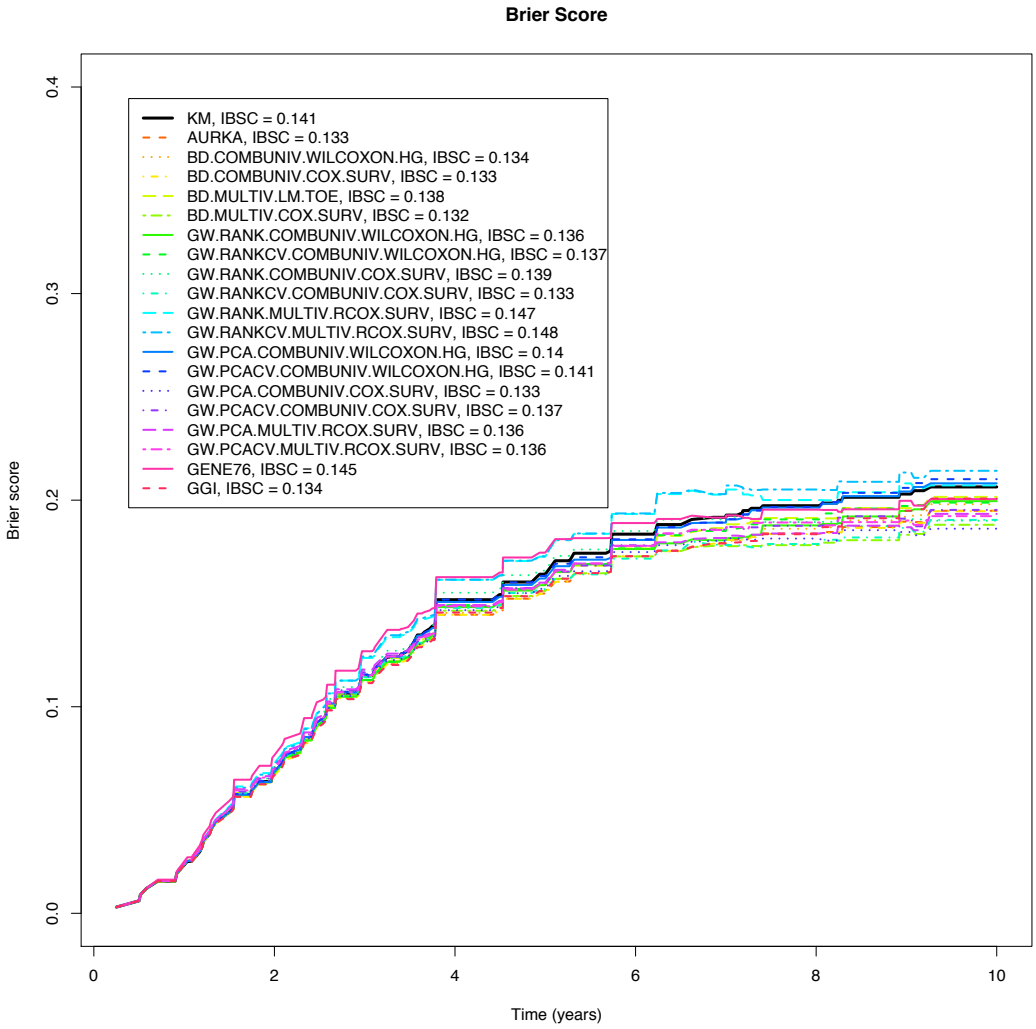
## 7 BRIER SCORE FOR RISK GROUP PREDICTION



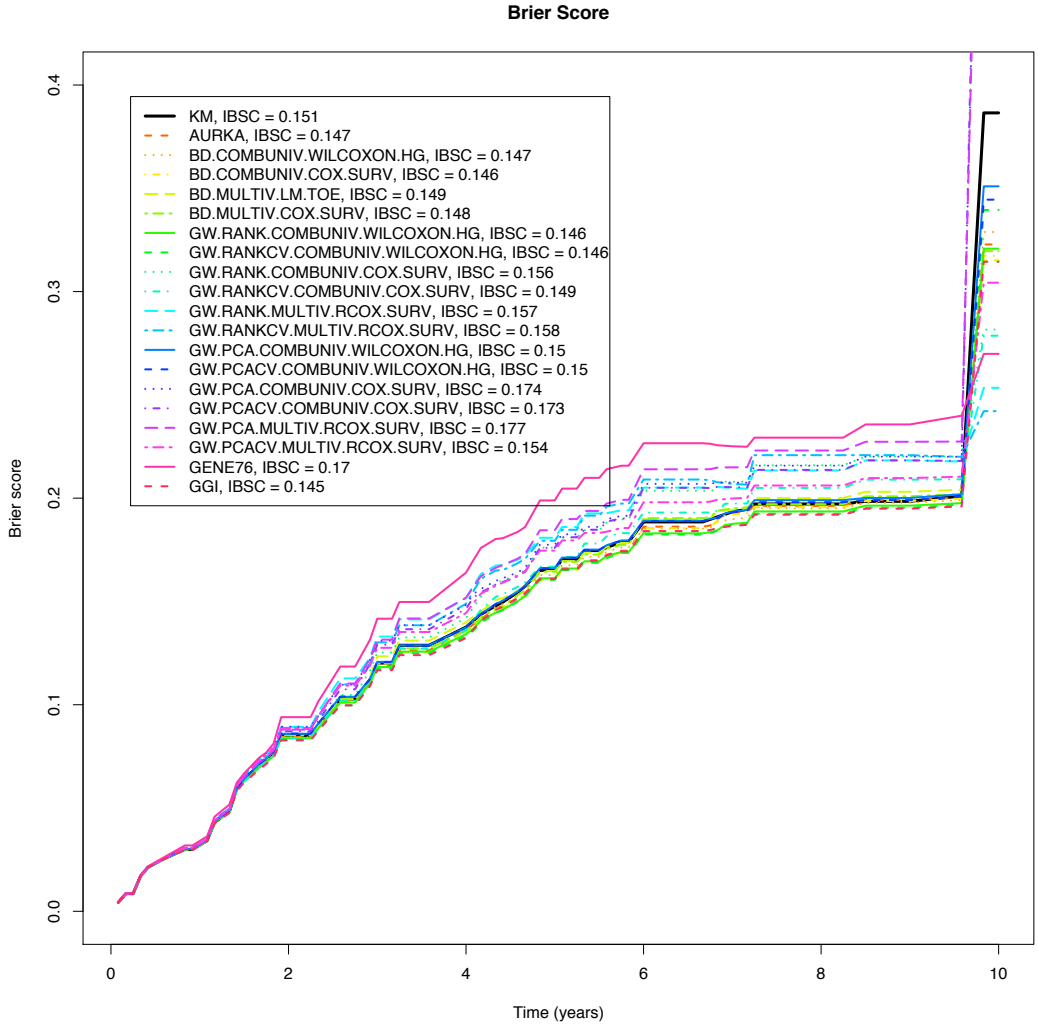

Supplementary Figure 12: Brier scores with respect to the time for the risk groups predicted by all the methods in the training set (VDX). AURKA and GGI models were not fitted on VDX which can be considered as a validation set for these models. The IBSC is given in the legend.




Supplementary Figure 13: Brier scores with respect to the time for the risk groups predicted by all the methods in the TBG validation set. The IBSC is given in the legend.

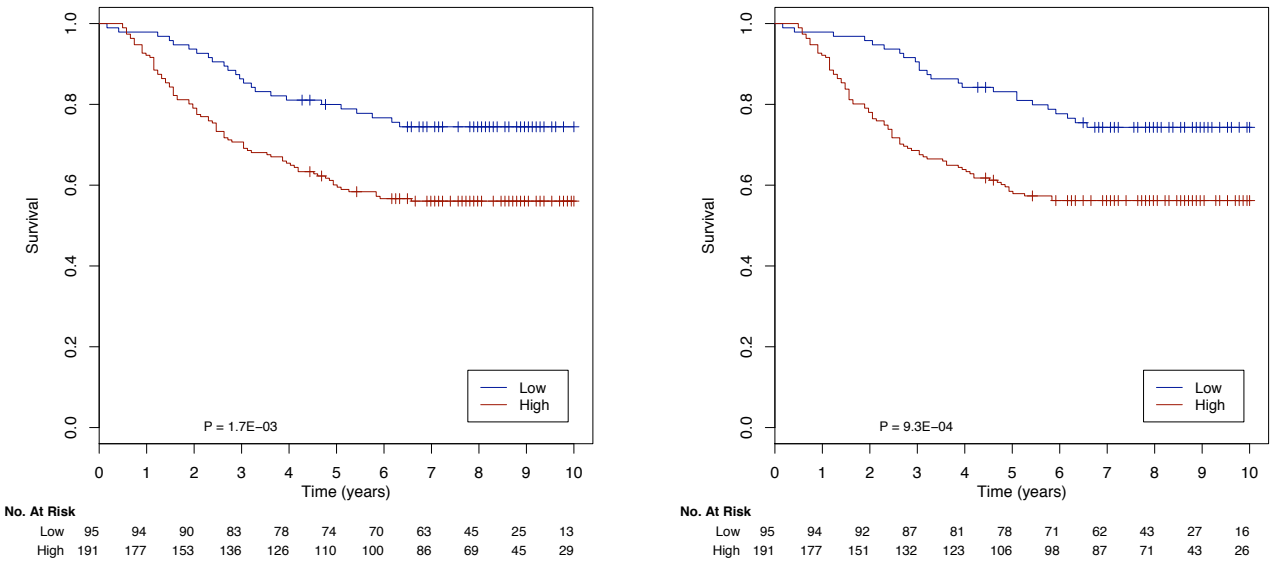



Supplementary Figure 14: Brier scores with respect to the time for the risk groups predicted by all the methods in the TAM validation set. The IBSC is given in the legend.

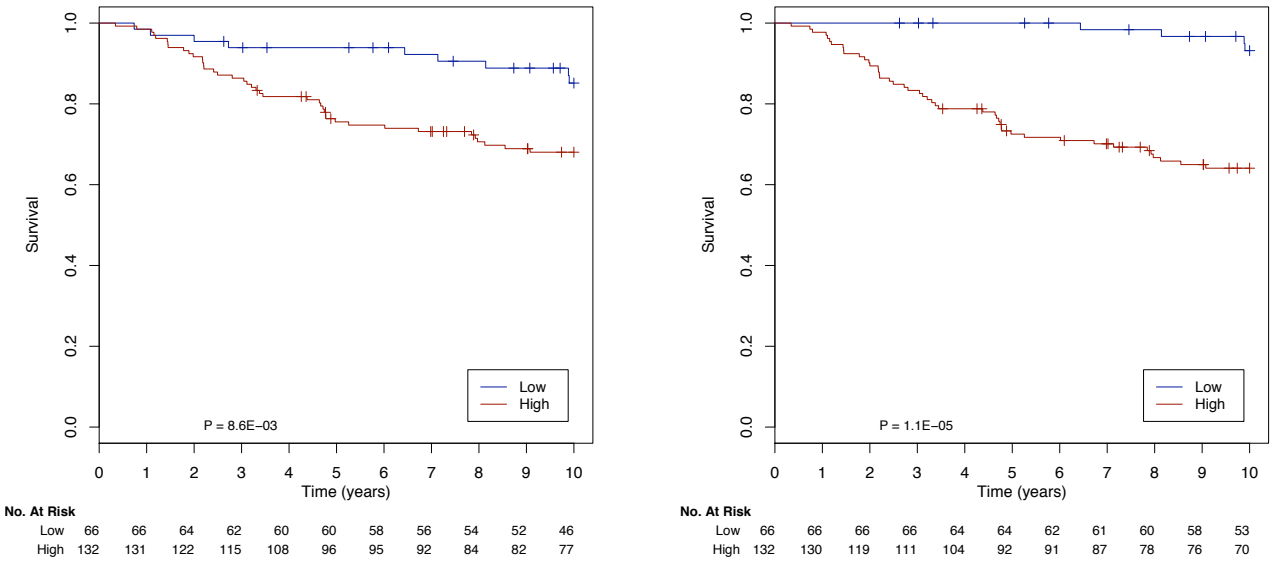



Supplementary Figure 15: Brier scores with respect to the time for the risk groups predicted by all the methods in the UPP validation set. The IBSC is given in the legend.

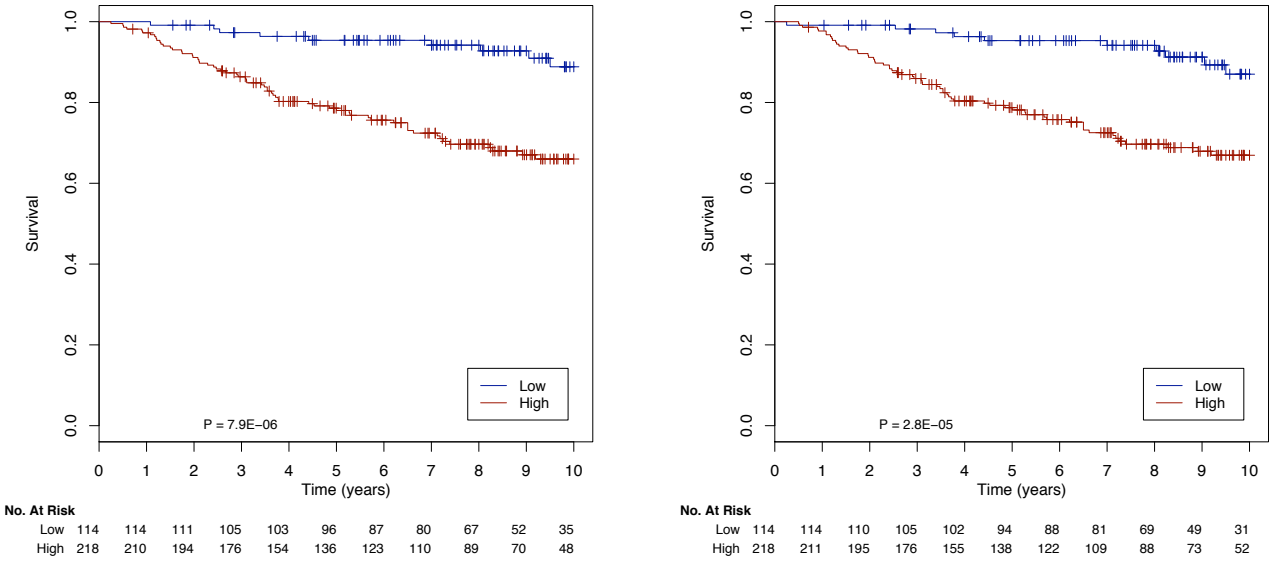
## 8 SURVIVAL CURVES FOR RISK GROUP PREDICTIONS COMPUTED BY AURKA AND GGI MODELS



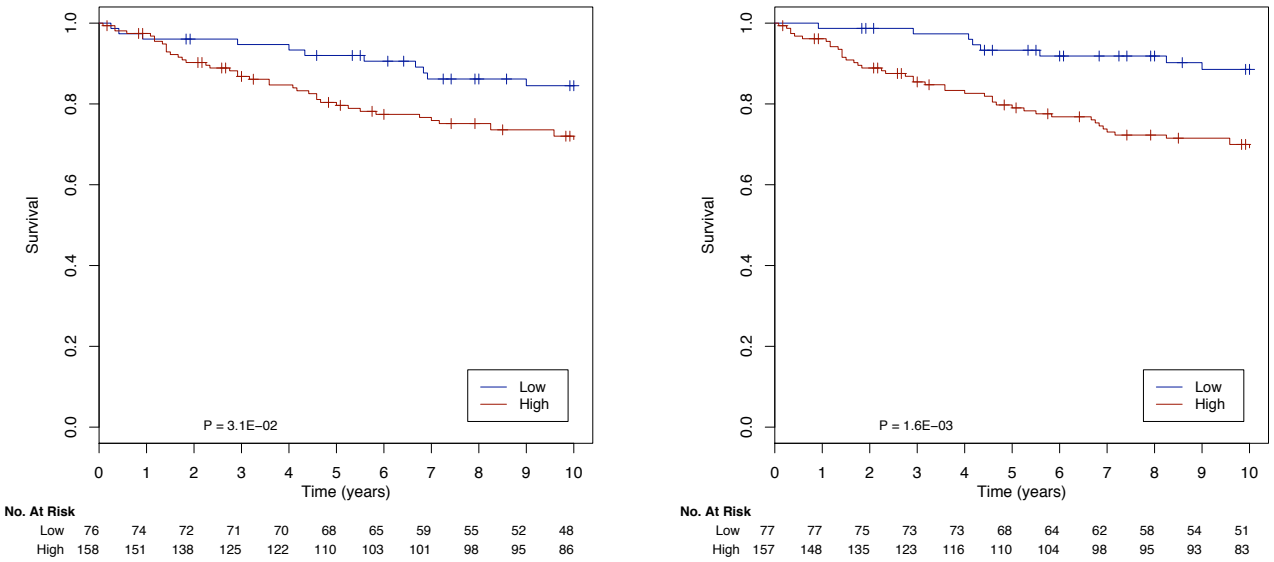

Supplementary Figure 16: Kaplan-Meier survival curves for the risk groups as computed by AURKA (left) and GGI (right) in VDX.




Supplementary Figure 17: Kaplan-Meier survival curves for the risk groups as computed by AURKA (left) and GGI (right) in TBG.




Supplementary Figure 18: Kaplan-Meier survival curves for the risk groups as computed by AURKA (left) and GGI (right) in TAM.

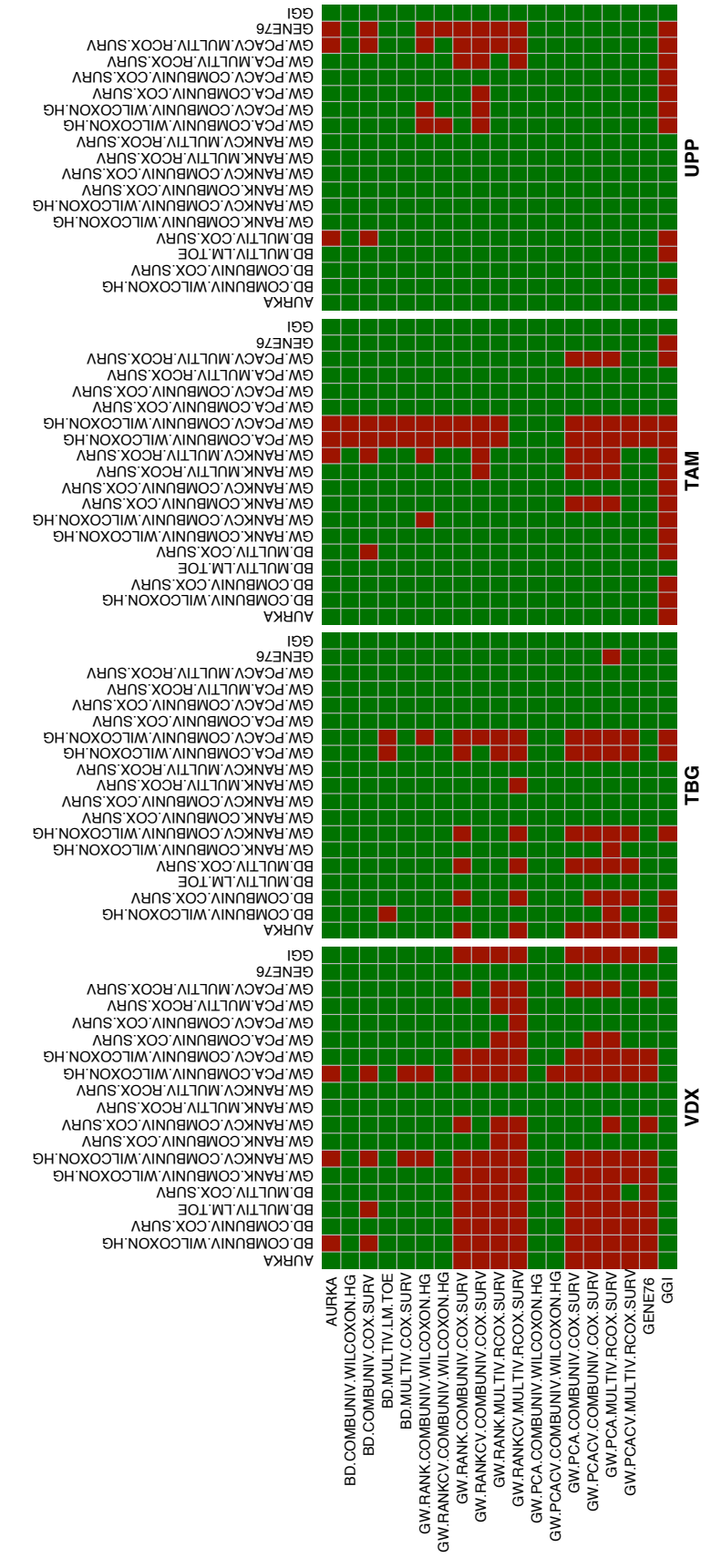



Supplementary Figure 19: Kaplan-Meier survival curves for the risk groups as computed by AURKA (left) and GGI (right) in UPP.

## 9 PAIRWISE PERFORMANCE COMPARISONS

The p-values computed from the pairwise performance comparisons between all the risk prediction methods are represented using heatmaps. The methods in rows are compared to the methods in columns using a one-tailed test of superiority. Each p-value $< 0.05$ is represented by a red box, otherwise by a green box. Therefore, methods whose the row is mostly red are better than most methods and inversely, methods whose the column is mostly red are poorer than most methods.

### 9.1 Risk Score Prediction

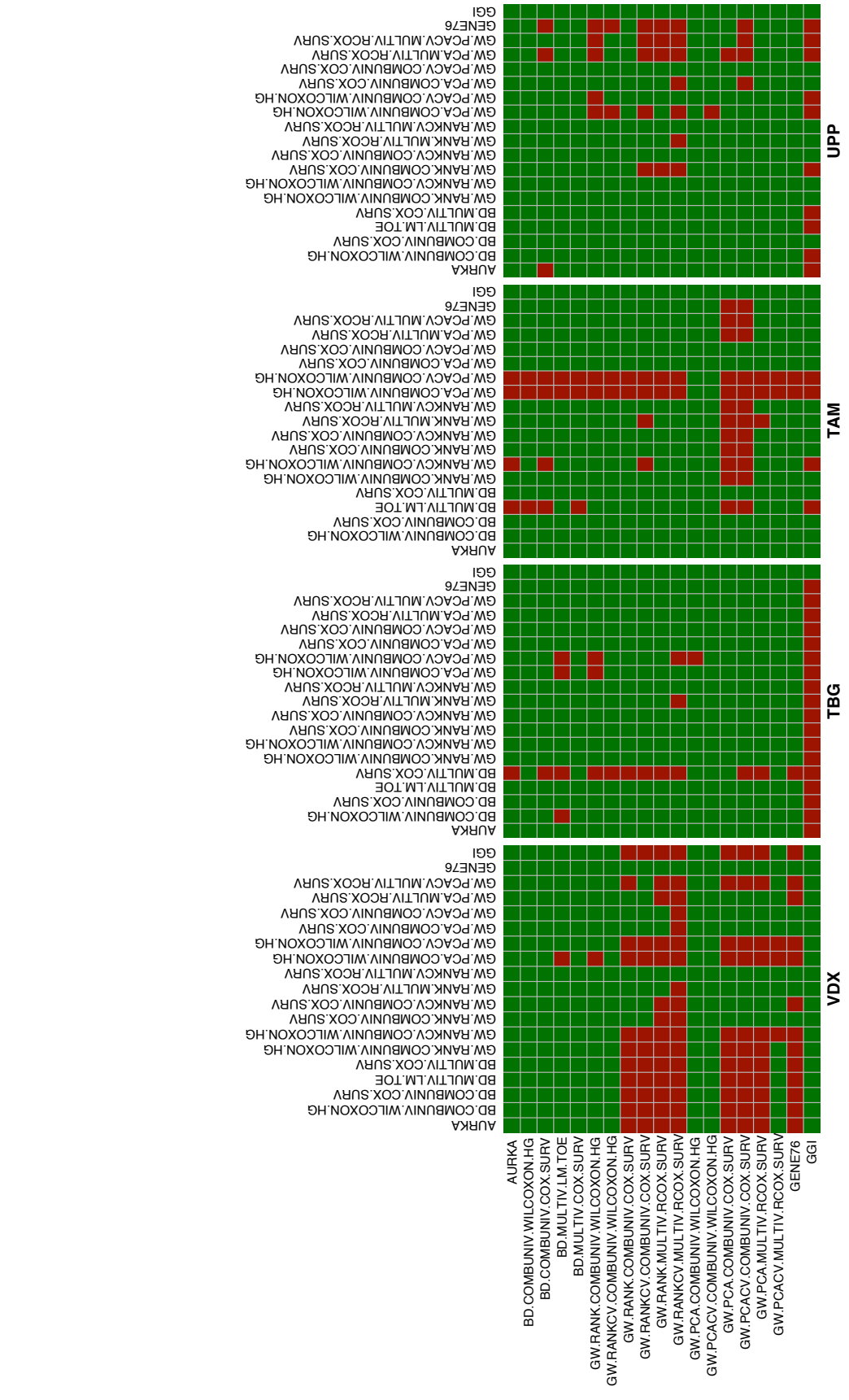



Supplementary Figure 20: Heatmap for all the pairwise comparisons between the concordance indices for the risk scores predicted by all the risk prediction methods in the training set (VDX) and the three validation sets (TBG, TAM, and UPP). AURKA and GGI models were not fitted on VDX which can be considered as a validation set for these models.

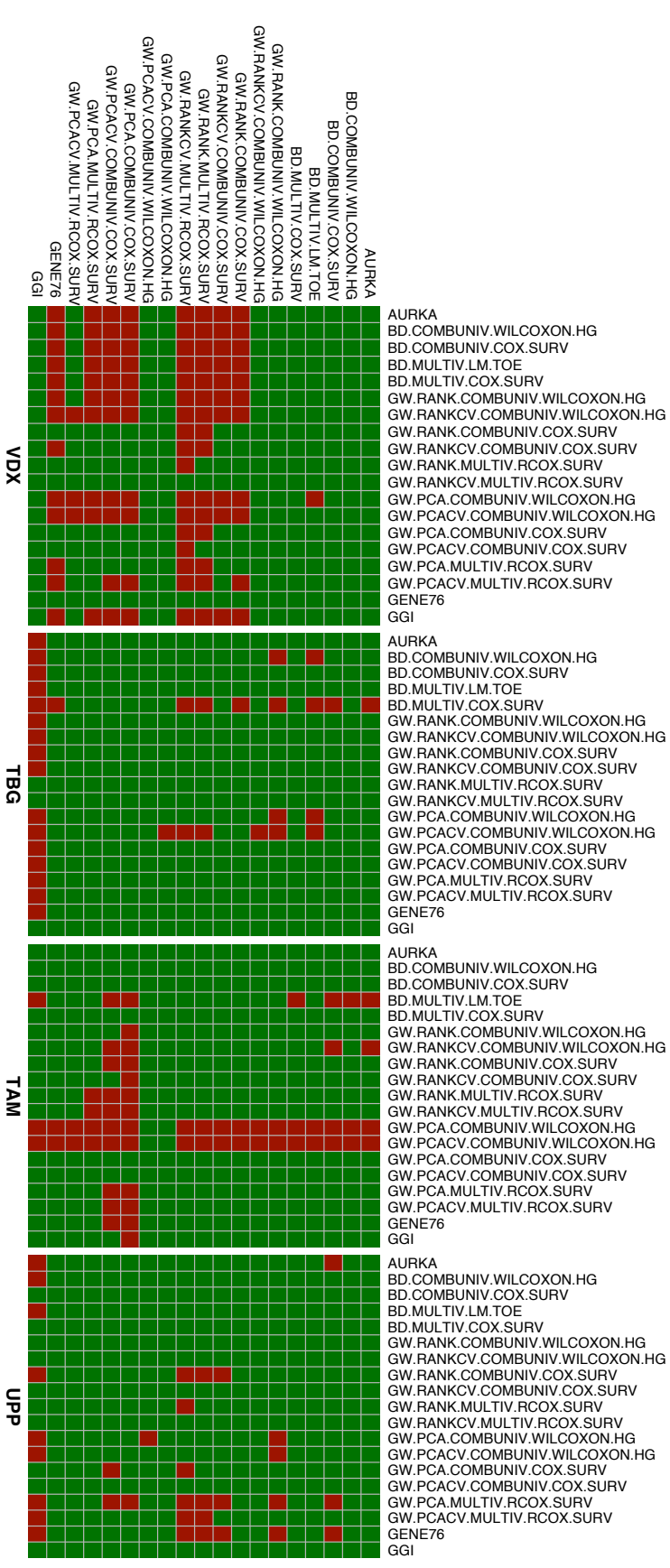
Supplementary Figure 21: Heatmap for all the pairwise comparisons between the AUCs for the risk scores predicted by all the risk prediction methods in the training set (VDX) and the three validation sets (TBG, TAM, and UPP). AURKA and GGI models were not fitted on VDX which can be considered as a validation set for these models.


Supplementary Figure 22: Heatmap for all the pairwise comparisons between the IBSCs for the risk scores predicted by all the risk prediction methods in the training set (VDX) and the three validation sets (TBG, TAM, and UPP). AURKA and GGI models were not fitted on VDX which can be considered as a validation set for these models.

## 9.2 Risk Group Prediction




Supplementary Figure 23: Heatmap for all the pairwise comparisons between the concordance indices for the risk groups predicted by all the risk prediction methods in the training set (VDX) and the three validation sets (TBG, TAM, and UPP). AURKA and GGI models were not fitted on VDX which can be considered as a validation set for these models.




Supplementary Figure 24: Heatmap for all the pairwise comparisons between the hazard ratios for the risk groups predicted by all the risk prediction methods in the training set (VDX) and the three validation sets (TBG, TAM, and UPP). AURKA and GGI models were not fitted on VDX which can be considered as a validation set for these models.


Supplementary Figure 25: Heatmap for all the pairwise comparisons between the IBSCs for the risk groups predicted by all the risk prediction methods in the training set (VDX) and the three validation sets (TBG, TAM, and UPP). AURKA and GGI models were not fitted on VDX which can be considered as a validation set for these models.

## 10 TBG AS TRAINING SET

In this work, we used VDX as training set and TBG, TAM and UPP as validation sets, this choice being guided by the original publications in BC prognostication (see Section 3.1). We also performed all the analyses using TBG as training set to ensure that our results were not driven by the choice of the training set. We obtained very similar results for risk score prediction (see Supplementary Table 8) and risk group prediction (see Supplementary Table 9). It is worth to mention that the accuracy measures for AURKA, GENE76 and GGI did not change due to the fact that these models were fully defined in previous publications. Moreover, VDX cannot be considered as a validation set for GENE76 since this dataset was used to fit this model.

Since the results were very similar, the conclusions of this study remained unchanged in using TBG as training set instead of VDX. We did not use TAM or UPP as training set because the selection of patients was different in terms of treatment and estrogen receptor status (see Section 3.1).

### 10.1 Signature Size

We estimated the performance using a 5-fold cross-validation for the signature size values from 1 to 45, and values from 50 to 200 spaced by 10. Due to the complexity of the learning algorithm RCOX, we did not estimate the performance for signature size values larger than 45.

| Model | Signature size |
|---|---|
| GW.RANKCV.COMBUNIV.WILCOXON.HG | 16 |
| GW.RANKCV.COMBUNIV.COX.SURV | 13 |
| GW.RANKCV.MULTIV.RCOX.SURV | 28 |
| GW.PCACV.COMBUNIV.WILCOXON.HG | 41 |
| GW.PCACV.COMBUNIV.COX.SURV | 17 |
| GW.PCACV.MULTIV.RCOX.SURV | 35 |

Supplementary Table 7: Signature size selected using cross-validated dimension reduction strategy for each method.

## 10.2 Performance for Risk Score Prediction

| | *C*-index | | | | IAUC | | | | IBSC | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | TBG | VDX | TAM | UPP | TBG | VDX | TAM | UPP | TBG | VDX | TAM | UPP |
| KM | | | | | | | | | 0.126 | 0.211 | 0.121 | 0.138 |
| AURKA | 0.609* | 0.636 | 0.683 | 0.637 | 0.601* | 0.636 | 0.674 | 0.63 | **0.123*** | **0.204** | **0.117** | **0.138** |
| BD.COMBUNIV.WILCOXON.HG | 0.625 | 0.59 | 0.671 | 0.622 | **0.651** | 0.592 | 0.666 | 0.602 | 0.123 | 0.212 | 0.117 | 0.138 |
| BD.COMBUNIV.COX.SURV | **0.676** | 0.571 | 0.63 | 0.587 | **0.711** | 0.592 | 0.62 | 0.585 | **0.12** | 0.206 | 0.116 | **0.133** |
| BD.MULTIV.LM.TOE | 0.649 | 0.585 | 0.67 | 0.653 | **0.676** | 0.587 | **0.684** | **0.687** | 0.123 | 0.21 | 0.116 | **0.133** |
| BD.MULTIV.COX.SURV | 0.668 | 0.584 | 0.617 | 0.589 | **0.698** | 0.604 | 0.611 | 0.593 | **0.118** | 0.212 | 0.119 | **0.131** |
| GW.RANK.COMBUNIV.WILCOXON.HG | **0.646** | 0.611 | 0.71 | **0.682** | **0.665** | 0.609 | **0.711** | **0.691** | 0.123 | 0.21 | **0.115** | 0.137 |
| GW.RANKCV.COMBUNIV.WILCOXON.HG | 0.645 | 0.604 | 0.708 | **0.678** | **0.662** | 0.602 | **0.71** | **0.687** | 0.123 | 0.21 | **0.115** | 0.138 |
| GW.RANK.COMBUNIV.COX.SURV | **0.833** | 0.578 | 0.708 | 0.625 | **0.857** | 0.587 | **0.697** | 0.631 | **0.0833** | 0.233 | 0.128 | 0.165 |
| GW.RANKCV.COMBUNIV.COX.SURV | **0.825** | 0.569 | 0.708 | 0.621 | **0.851** | 0.577 | **0.693** | 0.612 | **0.0823** | 0.237 | 0.126 | 0.159 |
| GW.RANK.MULTIV.RCOX.SURV | **0.88** | 0.556 | 0.684 | 0.591 | **0.914** | 0.552 | **0.683** | 0.578 | **0.0722** | 0.242 | 0.158 | 0.192 |
| GW.RANKCV.MULTIV.RCOX.SURV | **0.882** | 0.557 | 0.698 | 0.604 | **0.911** | 0.553 | **0.699** | 0.593 | **0.0694** | 0.245 | 0.158 | 0.188 |
| GW.PCA.COMBUNIV.WILCOXON.HG | 0.621 | 0.578 | 0.686 | 0.655 | **0.648** | 0.584 | **0.688** | **0.653** | 0.125 | 0.212 | 0.118 | 0.139 |
| GW.PCACV.COMBUNIV.WILCOXON.HG | 0.622 | 0.577 | 0.685 | 0.652 | **0.645** | 0.583 | **0.686** | **0.649** | 0.124 | 0.212 | 0.118 | 0.139 |
| GW.PCA.COMBUNIV.COX.SURV | **0.752** | 0.623 | 0.662 | 0.601 | **0.783** | 0.652 | 0.65 | 0.591 | **0.115** | 0.209 | 0.12 | **0.134** |
| GW.PCACV.COMBUNIV.COX.SURV | **0.724** | 0.616 | 0.662 | 0.596 | **0.748** | 0.644 | 0.651 | 0.586 | **0.118** | 0.21 | 0.119 | **0.134** |
| GW.PCA.MULTIV.RCOX.SURV | **0.769** | 0.657 | 0.708 | 0.647 | **0.809** | **0.682** | **0.681** | **0.648** | **0.102** | 0.212 | 0.121 | 0.142 |
| GW.PCACV.MULTIV.RCOX.SURV | **0.79** | 0.663 | 0.696 | 0.641 | **0.822** | **0.682** | 0.666 | **0.639** | **0.0967** | 0.204 | 0.126 | 0.145 |
| GENE76 | 0.64 | **0.754**# | 0.667 | 0.557 | **0.632** | **0.794**# | 0.633 | 0.558 | 0.122 | **0.198**# | 0.117 | 0.143 |
| GGI | **0.652*** | 0.613 | **0.718** | 0.67 | **0.671*** | 0.611 | **0.717** | **0.686** | 0.123* | 0.207 | **0.113** | **0.137** |

*As AURKA and GGI models were not fitted on TBG, this dataset can be considered as a validation set.
# GENE76 was fitted on VDX which cannot be considered as a validation set.

Supplementary Table 8: Performance for risk score prediction in the training set (TBG) and the three validation sets (VDX, TAM, and UPP). The accuracy measures in **bold** are significantly better than the accuracy of AURKA model. In case of IBSC, the accuracy measures of AURKA are in **bold** if they are significantly better than KM, the benchmark model, , whatever the performance improvement.

### 10.3 Performance for Risk Group Prediction

| | $C$-index | | | | IAUC | | | | IBSC | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | TBG | VDX | TAM | UPP | TBG | VDX | TAM | UPP | TBG | VDX | TAM | UPP |
| KM | | | | | | | | | 0.126 | 0.211 | 0.121 | 0.138 |
| AURKA | 0.729* | 0.685 | 0.834 | 0.673 | 2.43* | 2.04 | 4.64 | 1.84 | **0.122*** | **0.206** | **0.114** | **0.133** |
| BD.COMBUNIV.WILCOXON.HG | 0.722 | 0.709 | 0.805 | 0.669 | 2.38 | 2.12 | 4.28 | 1.98 | 0.122 | 0.205 | 0.115 | 0.134 |
| BD.COMBUNIV.COX.SURV | 0.741 | 0.621 | 0.674 | 0.571 | 2.91 | 1.58 | 1.95 | 1.33 | 0.122 | 0.208 | 0.118 | 0.136 |
| BD.MULTIV.LM.TOE | 0.815 | 0.676 | 0.656 | 0.726 | 3.59 | 1.93 | 1.82 | 2.48 | **0.118** | 0.205 | 0.121 | **0.131** |
| BD.MULTIV.COX.SURV | 0.791 | 0.615 | 0.645 | 0.584 | 4.22 | 1.57 | 1.83 | 1.4 | **0.12** | 0.209 | 0.12 | 0.136 |
| GW.RANK.COMBUNIV.WILCOXON.HG | **0.905** | 0.682 | 0.824 | **0.846** | **7.08** | 1.89 | 4.22 | **4.97** | **0.114** | 0.206 | 0.115 | **0.126** |
| GW.RANKCV.COMBUNIV.WILCOXON.HG | **0.906** | 0.686 | 0.847 | **0.873** | **7.18** | 1.97 | 4.71 | **5.2** | **0.114** | 0.205 | 0.114 | **0.14** |
| GW.RANK.COMBUNIV.COX.SURV | **0.949** | 0.648 | 0.77 | 0.648 | **16.1** | 1.71 | 3.31 | 1.51 | **0.113** | 0.209 | 0.117 | 0.154 |
| GW.RANKCV.COMBUNIV.COX.SURV | **0.95** | 0.647 | 0.806 | 0.662 | **15.8** | 1.77 | 4.27 | 1.66 | **0.113** | 0.208 | 0.115 | 0.151 |
| GW.RANK.MULTIV.RCOX.SURV | **0.943** | 0.58 | 0.786 | 0.571 | **20.6** | 1.42 | 2.94 | 1.3 | **0.114** | 0.213 | 0.116 | 0.158 |
| GW.RANKCV.MULTIV.RCOX.SURV | **0.972** | 0.577 | 0.813 | 0.606 | **37.9** | 1.41 | 3.88 | 1.45 | **0.113** | 0.214 | 0.116 | 0.158 |
| GW.PCA.COMBUNIV.WILCOXON.HG | 0.777 | 0.704 | 0.851 | **0.864** | 3.1 | 2.03 | 5.41 | **4.11** | **0.12** | 0.205 | 0.114 | **0.138** |
| GW.PCACV.COMBUNIV.WILCOXON.HG | 0.777 | 0.694 | 0.852 | **0.828** | 3.04 | 2 | 5.43 | **3.53** | **0.12** | 0.205 | 0.114 | **0.138** |
| GW.PCA.COMBUNIV.COX.SURV | **0.95** | 0.676 | 0.751 | 0.508 | **15.2** | 2.01 | 3.43 | 0.987 | **0.114** | 0.206 | 0.118 | 0.144 |
| GW.PCACV.COMBUNIV.COX.SURV | **0.976** | 0.635 | 0.732 | 0.515 | **32.8** | 1.67 | 3.03 | 1.03 | **0.113** | 0.211 | 0.12 | 0.145 |
| GW.PCA.MULTIV.RCOX.SURV | **0.898** | 0.724 | 0.861 | 0.665 | **6.64** | 2.38 | 6.91 | 1.83 | **0.115** | **0.202** | **0.111** | 0.134 |
| GW.PCACV.MULTIV.RCOX.SURV | **0.898** | 0.75 | 0.857 | 0.717 | **6.69** | 2.85 | 6.8 | 2.37 | **0.115** | **0.2** | **0.112** | 0.131 |
| GENE76 | 0.756 | **0.903**# | 0.754 | 0.548 | 2.79 | **8.37**# | 3.52 | 1.16 | **0.12** | **0.195**# | 0.116 | 0.138 |
| GGI | **0.906*** | 0.706 | 0.824 | **0.769** | **7.24*** | 2.12 | 4.03 | **2.88** | **0.114*** | 0.204 | 0.115 | **0.13** |

*As AURKA and GGI models were not fitted on TBG, this dataset can be considered as a validation set.
#GENE76 was fitted on VDX which cannot be considered as a validation set.

Supplementary Table 9: Performance for risk score prediction in the training set (TBG) and the three validation sets (VDX, TAM, and UPP). The accuracy measures in **bold** are significantly better than the accuracy of AURKA model. In case of IBSC, the accuracy measures of AURKA are in **bold** if they are significantly better than KM, the benchmark model, whatever the performance improvement.

## 11 LIMITED COMPARISON WITH CLINICAL RISK PREDICTION METHODS

| | Specificity | | $C$-index | | IAUC | | IBSC | |
|---|---|---|---|---|---|---|---|---|
| Model | VDX | TBG | VDX | TBG | VDX | TBG | VDX | TBG |
| KM | | | | | | | 0.189 | 0.145 |
| AURKA | 0.253 | 0.348 | 0.636 | 0.609 | 0.636 | 0.601 | 0.178 | 0.144 |
| GGI | 0.258 | 0.522 | 0.613 | 0.652 | 0.611 | 0.671 | 0.183 | 0.14 |
| AOL | 0.142 | 0.199 | 0.495 | 0.637 | 0.495 | 0.672 | 0.188 | 0.146 |
| NPI | 0.209 | 0.399 | 0.627 | 0.634 | 0.616 | 0.665 | 0.179 | 0.126 |

Supplementary Table 10: Performance of GGI, AURKA, AOL and NPI models for risk score prediction in VDX and TBG datasets. The columns refers to the specicity for a sensitivity of 90%, the Concordance index, the IAUC and the IBSC.