Research Article
Open Access

An artificial neural network framework for biomarker development to predict distant metastasis of breast cancer from gene expression data

Abstract

Background: In spite of the low number of overlapping genes between the various published gene signatures for prediction of distant metastasis, the signatures had many pathways in common. Identification of the key biological processes, rather than the assessment of signatures based on individual genes, allows not only to build a biological meaningful and robust gene biomarker from functionally related genes, but also provides insight into the mechanism of the disease development.

Methods: In this work, we develop an artificial neural network framework for biomarker development to predict distant metastasis of breast cancer from gene expression data. Our intention for this structure is to potentially identify the underlying biological process in the design network. We follow the evaluation framework by McShane and Hayes from the National Cancer Institute to evaluate the proposed biomarker herein on three important aspects: analytical validity, clinical validity, and clinical utility.

Results: The Gene Recurrence Risk Score (GRRS) from the artificial neural network model developed herein has demonstrated significant association with the probability of distant metastasis within 10 years after adjustment of other standard clinical or pathological factors. On independent assessment TRANSBIG dataset, with distant metastasis within 10 years as clinical outcome, GRRS produced a sensitivity of 86% and a specificity of 37%, a 4% and 3% improvement over Veridex risk score respectively for ER positive group. GRRS also produced a sensitivity of 91% and a specificity of 32%, a 4% and 5% improvement over Veridex risk score respectively for ER negative group. It also showed on par classification accuracy with Veridex risk score for prediction of distant metastasis within 5 years for both ER subgroups.

Conclusion: The Gene Recurrence Risk Score (GRRS) has demonstrated the prognostic value and high classification accuracy for prediction of distant metastasis within 5 years and within 10 years. The artificial neural network framework for biomarker development for this specific task proves to be robust and effective. The model building itself also reveals the necessity to stratify patients by their ER status for classification.

Keywords: gene expression biomarker, distant metastasis of breast cancer prediction, precision oncology, artificial neural network, personalized medicine

Volume 7 Issue 6 - 2018

Yingying Ma,[1] Xinping Cui[2]

[1]Inga Data, Inc., Los Angeles, USA/Shanghai, China

[2]Department of Statistics, University of California at Riverside, USA

Correspondence: Yingying Ma, Inga Data, Inc., Los Angeles, California Shanghai, P.R. China, Email bm_ai@posteo.org

Received: November 25, 2018 | **Published:** December 28, 2018

Background

Use of biomarker-based tests to guide breast cancer therapy, including tests based on omics data, has steadily increased over the last decade in concert with efforts to personalize treatment strategies to maximize chances patients will receive treatments that most benefit them. In particular, the majority of patients with early breast cancer receive some form of systemic adjuvant therapy (chemotherapy and/or endocrine therapy), which may have important side effects and which puts considerable burden on health care costs. Although guidelines have been developed to assist clinicians in selecting patients who should receive adjuvant therapy, it still remains a challenge to distinguish those patients who would really need adjuvant systemic therapy from those who could be spared such treatment. Two independent groups conducted comprehensive genome- wide assessments of gene expression profiling to identify broadly applicable prognostic markers. The Netherlands Cancer Institute in Amsterdam and Rosetta identified a 70- gene prognostic signature reported by Van't Veer at al.[1,2] Thereafter, Erasmus Medical Center and Veridex identified a 76-gene prognostic signature that could be used to predict the development of distant metastases within 5 years in N- primary breast cancer patients (irrespective of age and tumor size) who did not receive systemic treatment.[3,4] Yu et al.[5,6] showed that in spite of the low number of overlapping genes between the various published gene signatures for breast cancer, the signatures had many pathways in common, implying that different prognostic gene signatures represent common biology. Identification of the key biological processes, rather than the assessment of sig- natures based on individual genes, allows not only to build a biological meaningful and robust gene biomarker from functionally related genes, but also provides in- sight into the mechanism of the disease development. The hallmarks of cancer are represented by pathways rather than individual genes and the crucial aspect of pathways is that their constituting genes are actively interacting with each other (Hanahan and Weinberg).[7] In contrast, biomarkers based on individual genes neglect these completely. In this work, we develop an artificial

neural network framework for biomarker development to predict distant metastasis of breast cancer from gene expression data. Our intention for this structure is to potentially identify and incorporate the underlying biological process in the design network. We follow the evaluation framework by Shane Mc & Hayes[8] from the National Cancer Institute to evaluate the proposed biomarker herein on three important aspects: analytical validity, clinical validity, and clinical utility. There are couple of advantages of this framework:

a. It's simple. We apply a single neural network for prediction and score generation from the gene expression data.

b. It's synthetic. The neural network naturally incorporates the interactions among functionally related genes. The neural network is the most common approach for exact inference of gene regulatory network (GRN).

c. It's robust. Regularization and Cross-Validation can also be easily incorporated into the network model building process to offset "over- fitting".

Methods

Datasets

We select the Wang[3] data set (GEO series GSE2034) as our training dataset and KJ125 of GEO series GSE2990[9] as our validation dataset for the training purpose. TRANSBIG (GEO series GSE7390)[4] is selected as our independent assessment dataset. All datasets contain gene expression data from patients with lymph-node-negative primary breast cancer having not received adjuvant or neoadjuvant therapy. Microarray analysis of all datasets were performed with Affymetrix U133A Genechips. We avoid datasets with heterogeneous mixture of standard clinical and pathological characteristics when trying to determine the clinical value added by a new test is difficult. The clinical outcome for both GSE2034 and KJ125 is distant metastasis within 5 years. TRANSBIG has longer follow-up time, hence both distant metastasis within 5 years and 10 years data are available. The date of diagnosis of metastasis was defined as that at confirmation of metastasis after symptoms reported by the patient, detection of clinical signs, or at regular follow-up. Even though the clinical outcome of the training datasets is distant metastasis within 5 years, we choose distant metastasis within 10 years as primary clinical outcome and distant metastasis within 5 years as secondary clinical outcome for assessment purpose. The reason is that we would like to see if the prognostic capability of the biomarker holds beyond 5 years and eventually we would like to develop a marker which can provide classification accuracy of distant metastasis within 10 years or beyond, so that patients with longer "dormant" period will not be missed and proper targeted long-term therapy be put into place. The Wang data set contains gene expression and ER status data on 286 women with lymph node negative breast cancer. The tumor samples were selected from the tumor bank at the Erasmus Medical Center from patients who were treated during 1980-95, but who didn't receive systematic neoadjuvant or adjuvant therapy. KJ125 dataset consists of information obtained from a total of 125 patients with lymph-node-negative primary operable invasive breast cancer, whose frozen tumor specimens were archived at the John Radcliffe Hospital (Oxford, UK) and the Uppsala University Hospital (Uppsala, Sweden). TRANSBIG dataset consists of information obtained from a total of 198 patients with lymph-node-negative primary breast cancer, whose frozen tumor specimens were sent to Border Institute to perform the microarray analysis. The median follow-up for the 198 patients included was 14.0 years. And distant metastases were found in 51 (26%) of them, with 35 of them showing progression within 5 years (18%). The patients were assessed to high and low genomic risk using 76-gene prognostic signature as described previously,[3] and to high and low clinical risk, as defined by the Adjuvant! Online software using the pre-defined cutoff. We will name binary genomic risk group classification as Veridex risk score and binary clinical risk classification as AOL risk score. One hundred forty-three (72%) and 55 patients (28%) were classified as high and low genomic risk, whereas 152 (77%) and 46 (23%) patients were considered to be high and low clinical risk, respectively. Table 1 shows ER status and distant metastasis outcome breakdown for all three datasets.

Table 1 ER Status and Distant Metastasis Outcome of Three Data Sets

	WANG			**KJ125**			**TRANSBIG**		
	All patients (n=286)	ER+ (n=209)	ER- (n=77)	All patients (n=119)	ER+ (n=85)	ER- (n=34)	All patients (n=198)	ER+ (n=134)	ER- (n=64)
Metastasis within 5 years									
Yes	107 (37%)	80 (38%)	27 (35%)	34 (29%)	19 (22%)	8 (24%)	35 (21%)	17 (13%)	18 (28%)
No	179 (63%)	129 (62%)	50 (65%)	85 (71%)	66 (78%)	26 (76%)	163 (79%)	117 (87%)	46 (72%)
Metastasis within 10 years									
Yes							51 (26%)	28 (21%)	23 (36%)
No							147 (74%)	106 (79%)	41 (64%)

Gene recurrence risk Score development

We develop an Artificial Neural Network from gene expression data to predict the probability of distant metastasis of breast cancer (recurrence). Our intention for this structure is to potentially model the underlying biological process instead of individual genes in the design network. In order to facilitate the utility of the score, a Gene Recurrence Risk Score (GRRS) — the logarithmic of predicted probability of distant metastasis is created as the biomarker. The prediction accuracy and area under the curve (AUC) of the ROC curve are used to evaluate the model performance and classification capability of the GRRS. The confusion matrix is used to compare

GRRS classification accuracy with the Veridex risk score and AOL risk score. When a large feed forward neural network is trained on a small training set, it typically performs poorly on held-out test data. To address the problem of 'overfitting', we apply Dropout as our regularization technique.[10,11] The key idea of Dropout is to randomly drop units (along with their connections) from the neural network during training. This prevents units from co-adapting too much. This also significantly reduces overfitting and gives major improvements over other regularization methods. We initially use the total samples of the Wang dataset to train the model, but there is a lot of variability in the model performance. We suspect there is heterogeneity due to ER status need to be account for as noted by Mosley et al.[12] and Gruvberger et al.[13,14] that estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. We split the training data stratified by ER status.

To determine the optimal structure of the neural network, we apply double cross- validation, i.e. there is an 'outer loop' of validation with ER positive/negative subgroup of KJ125 as the validation dataset to correct for optimism and an 'inner loop' of 5-fold cross-validation on the training dataset of oversampled ER positive/negative subgroup of the Wang dataset for tuning parameters of the network. The final neural network trained from ER positive subgroup (ERP model) has three layers with two hidden layers. The first hidden layer has 300 nodes and second layer has 70 nodes. We then apply the final model on the independent assessment data set - ER positive subgroup of TRANSBIG dataset. Both distant metastasis within 5 years and within 10 years are used as the clinical outcomes. The classification accuracy is 87% and 79% and the area under the curve of the ROC is 0.77 and 0.69 respectively. Figure 1 shows the diagram of the final neural network. We use the same analysis approach with ER negative group. The final neural network model from ER negative subgroup (ERN model) has three layers with two hidden layers. The first hidden layer has 20o nodes. The second hidden layer has 70 nodes. We apply the final model obtained from training ER negative group on the independent assessment data set - ER negative subgroup of TRANSBIG dataset. Both distant metastasis within 5 years and within 10 years are used as the clinical outcomes. The classification accuracy is 73% and 66% and the area under the curve of the ROC is 0.68 and 0.64 respectively.

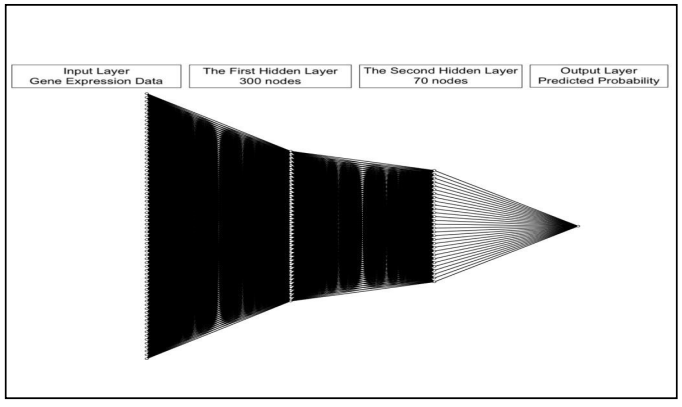

Figure 1 The diagram of final neural network model for ER positive group.

Results

Statistical validity of the predictor

With the independent assessment data TRANSBIG ER positive subgroup, we build a cox proportional hazard model (COX) with the Gene Recurrence Risk Score (GRRS) as one covariate together with other standard clinical factors to assess the effect of potential factors on the hazard of distant metastasis within 10 years simultaneously.

The p-value (p = 0.048) of the likelihood is significant, indicating that the model is significant. Age, surgery type, grade, size, histopathological tumor type, ER status all fail to be significant. The p-value for GRRS score is 0.02, indicating a statistically significant association of GRRS score with the distant metastasis hazard after adjustment for standard clinical or pathological factors. The hazard ratio (HR) is 1.04, i.e. the hazard increases 4% with one unit increase in GRRS.

With the independent assessment data TRANSBIG ER negative subgroup, we build another Cox proportional hazard model with the Gene Recurrence Risk Score (GRRS) as one covariate together with other standard clinical factors. Again, age, surgery type, grade, size, histopathological tumor type, ER status all fail to be significant. The p-value for GRRS score is 0.04, indicating a statistically significant association of GRRS score with the distant metastasis hazard after adjustment for standard clinical or pathological factors. The hazard ratio (HR) is 1.04, i.e. the hazard increases 4% with one unit increase in GRRS. Figure 2 show the ranking of significance of cox regression coefficients and their confidence bands for ER positive and negative subgroup separately.

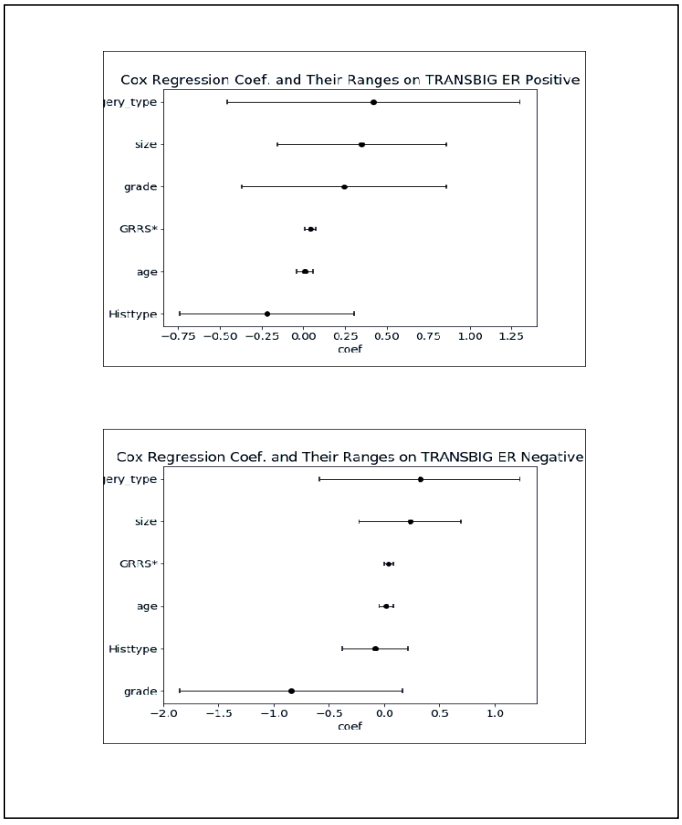

Figure 2 Cox regression coefficients and their ranges on TRANSBIG ER positive & negative subgroup separately.

Clinical validity and utility of the predictor

We then apply bootstrap resampling on TRANSBIG ER positive subgroup data to estimate the robustness of the predictor and determine the optimal cutoff value of GRRS for classification, i.e. at each iteration, predict on the resampled data to obtain GRRS score,

then classify each sample into high or low risk group with cutoff value at -40, -37, -35 and -34 respectively. At cutoff value -34, we obtain a confidence level of 80% that the sensitivity is greater than 80% and a confidence level of 80% that the specificity is greater than 33%. Figure 3 shows the empirical distribution of sensitivity and specificity of GRRS with cutoff value at -34 from bootstraps samples. Next, we apply bootstrap resampling on TRANSBIG ER negative subgroup data with the same approach as to the ER positive subgroup. At cutoff value -30, we obtain a confidence level of 80% that the sensitivity is greater than 86% and a confidence level of 80% that the specificity is greater than 26%. Figure 4 shows the empirical distribution of sensitivity and specificity of GRRS with cutoff value at-30 from bootstraps samples.

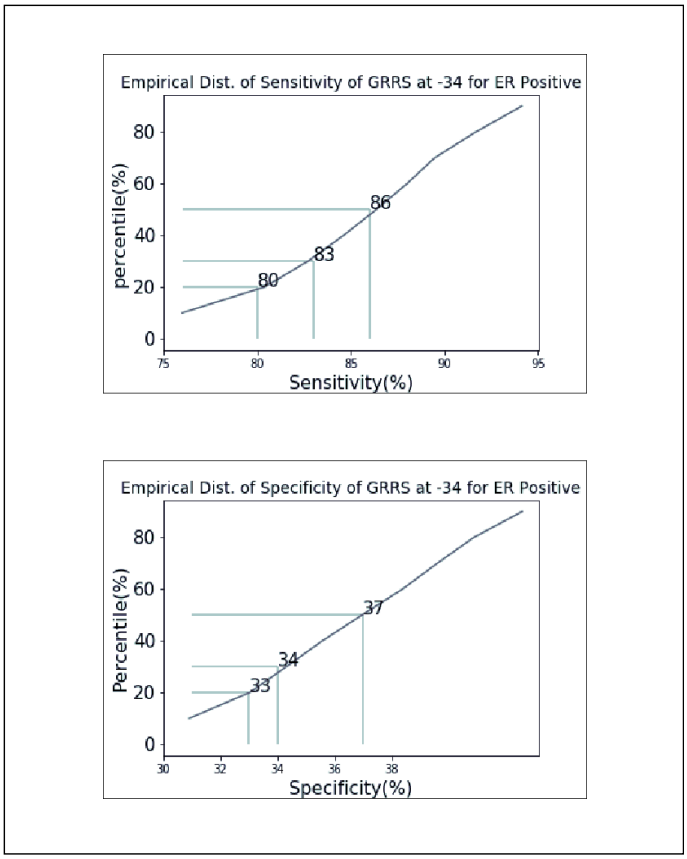

Figure 3 Empirical distribution of sensitivity and specificity of GRRS at cutoff point -34 for ER positive subgroup.

To assess the clinical validity and utility of the GRRS, we compare its classification accuracy with Veridex risk score and with AOL risk score using confusion matrix. We use distant metastasis within 10 years as primary clinical outcome and within 5 years as secondary clinical outcome for comparison and assessment. With distant metastasis within 10 years as clinical outcome, on TRANSBIG ER positive subgroup, we obtain sensitivity at 86% and specificity at 37% with GRRS at cutoff value -34. In comparison, the sensitivity and specificity is 82% and 34% with Veridex risk score, 78% and 37% with AOL risk score. GRRS has a 4% and 3% improvement in sensitivity and specificity over Veridex risk score respectively, and 8% improvement in sensitivity over AOL risk score. With distant metastasis within 10 years as clinical outcome, on TRANSBIG ER negative subgroup, we obtain sensitivity at 91% and specificity at 32% with GRRS at cutoff value at -30. In comparison, the sensitivity and specificity is 87% and 27% with Veridex risk score. GRRS has a 4% and 5% improvement in sensitivity and specificity over Veridex risk score respectively. Even though AOL risk score has assigned all distant metastasis patients to high risk, it also assigns all non- distant metastasis patients to high risk group, with specificity as 0%. With distant metastasis within 5 years as clinical outcome, for TRANSBIG ER positive subgroup, both GRRS and Veridex correctly identify all the patients who had distant metastasis within 5 years to high risk, but GRRS has specificity at 37% vs Veridex at 35%. AOL risk score failed to identify 3 patients with sensitivity at 82%. With distant metastasis within 5 years as clinical outcome, for TRANSBIG ER negative subgroup, both GRRS and Veridex have specificity at 28%, but GRRS failed to assign one distant metastasis patient to high risk. Once again, AOL risk score has assigned all distant metastasis patients to high risk, it also assigns all non- distant metastasis patients to high risk group, with specificity as 0%. Figure 5–8 show the confusion matrixes of GRRS with Veridex and AOL risk score at specified cutoff value/ clinical outcome/ER subgroup as specified above.

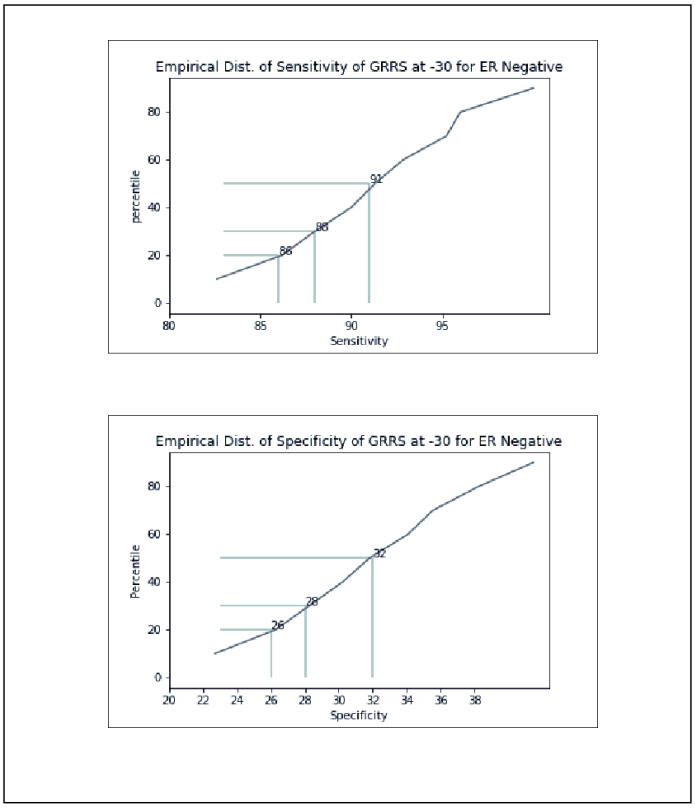

Figure 4 Empirical distribution of sensitivity and specificity of GRRS at cutoff point -30 for ER negative subgroup.

Discussion

The Gene Recurrence Risk Score (GRRS) from the artificial neural network model developed herein has demonstrated the significant association and the additional prognostic value for prediction of distant metastasis within 10 years from other standardly measured clinical factors. On the independent assessment data set TRANSBIG, GRRS has also produced better classification accuracy with higher sensitivity and specificity than Veridex risk score and AOL risk

score for prediction of distant metastasis within 10 years, and on par with Veridex risk score for prediction of distant metastasis within 5 years, demonstrating a strong clinical utility potential. And the model building process itself is insightful. There are couple of insights worthwhile to be mentioned here:

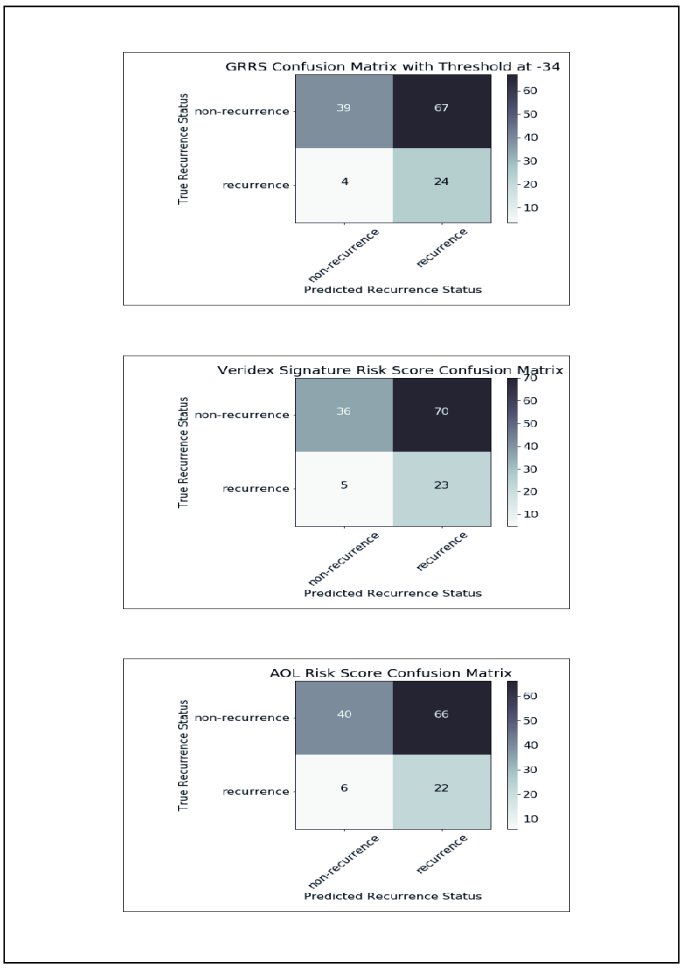

Figure 5 GRRS classification accuracy comparison with multiple risk scores on TRANSBIG ER positive subgroup (distant metastasis within 10 years).

a. The variability of model performance training ER positive and negative together indicates the heterogeneity in the data need to be account for. When we use the modeled trained on ER positive to apply on ER negative subgroup, we end up with an AUC at 0.5 indicating that the classification mechanism holds for ER positive doesn't apply for ER negative and vice versa. All these indicate that the mechanisms for disease progression could differ for these two ER-based subgroups of breast cancer patients. And it's necessary to stratify patients by their ER status for classification.

b. For all three datasets, Veridex risk score shows strong classification capability for ER positive subgroup. But it shows weaker prognostic capability for ER negative subgroup.

c. GRRS has better classification accuracy than Veridex risk score for prediction of distant metastasis within 10 years and on par with Veridex risk score for prediction distant metastasis within 5 years. It indicates that we might have identified a biological process with the network capturing the interactions between genes vs ones based on individual genes.

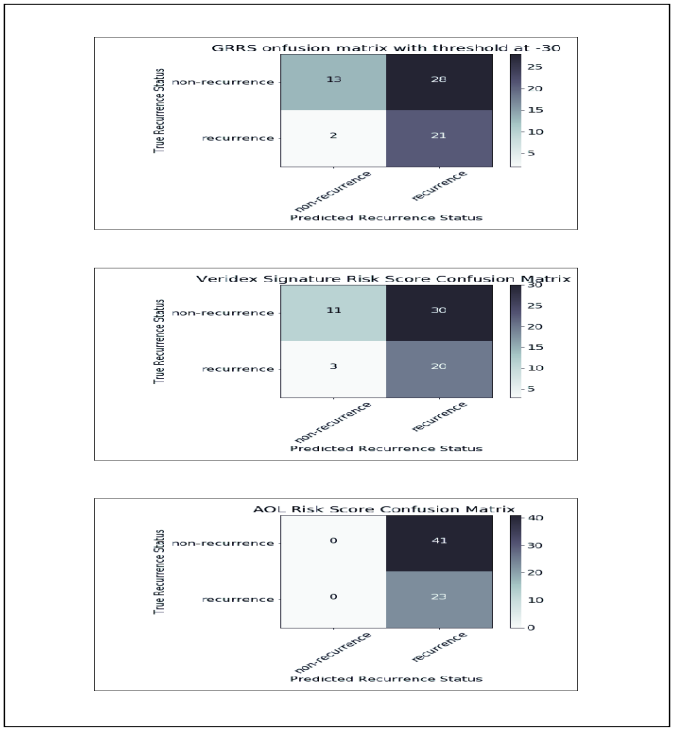

Figure 6 GRRS classification accuracy comparison with multiple risk scores on TRANSBIG ER negative subgroup (distant metastasis with 10 years).

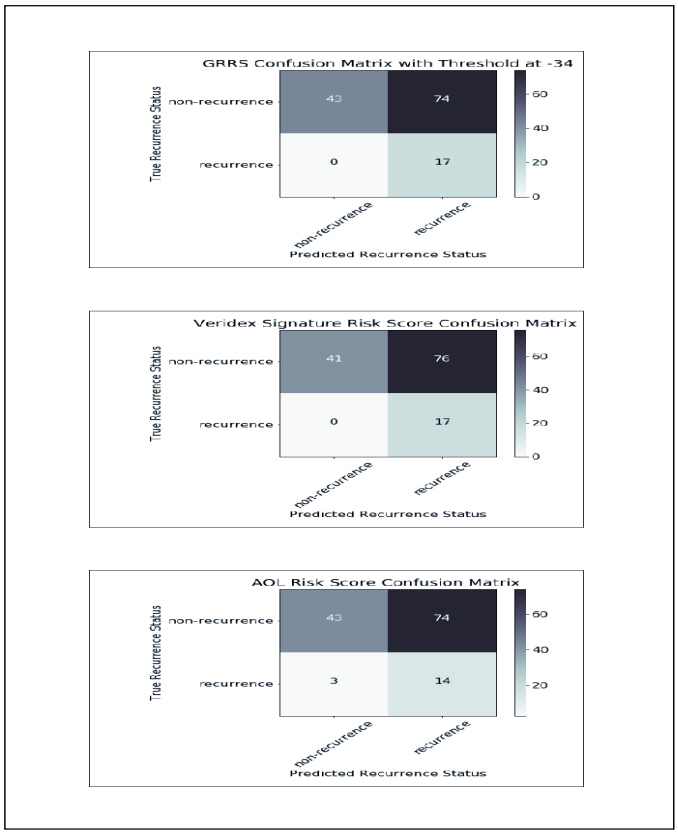

Figure 7 GRRS classification accuracy comparison with multiple risk scores on TRANSBIG ER positive subgroup (Distant Metastasis within 5 years).

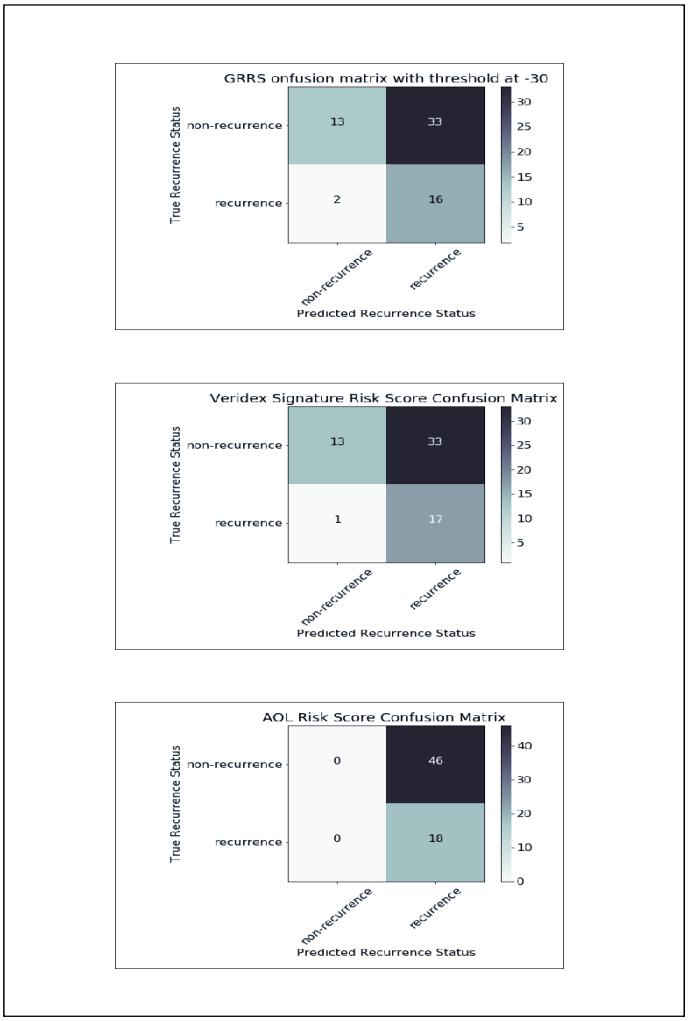

Figure 8 GRRS classification accuracy comparison with multiple risk scores on TRANSBIG ER negative (distant metastasis within 5 years).

The artificial neural network we developed herein is a GRN following the definition by Emmert-Streib & Dehmer[15] that "a network that has been inferred from gene expression data a gene regulatory network" which potentially provide information about regulatory interactions between regulators and their potential targets and gene-gene interactions. The exact inference of real world and synthetic large scale GRN remains to be a difficult task to be achieved.[16] Over the past few years, there has been a growing interest in approaches that integrate large scale inference information on molecular interactions into biomarker discovery.[17,18] The main disadvantage of this approach is that the set of known interactions might be quite large, many of them might not be relevant to the biological conditions under investigation. However, the targeted gene expression prediction using neural network is comparatively easier task than network inferences.

The follow-up we would like to do next is to work with physicians to understand the biological representation of the nodes and interactions identified by the artificial neural network developed herein and provide deeper theoretical, biological, and casual understanding of the identified artificial neural network and the underlying biological process and explore more applications of the proposed development and evaluation framework, such as therapy-guiding biomarker developments.

Conclusion

The Gene Recurrence Risk Score (GRRS) has demonstrated the prognostic value and high classification accuracy for prediction of distant metastasis within 5 years and within 10 years. The artificial neural network framework for biomarker development for this specific task proves to be robust and effective. The model building itself also reveals the necessity to stratify patients by their ER status for classification.[19–30]

List of abbreviations

GRRS, Gene Recurrence Risk Score is the logarithmic of the predicted probability of distant metastasis obtained from the artificial neural network developed herein.

Veridex Risk Score, High or low risk group classification according to 76-gen signature developed be Veridex using the pre-defined cutoff.

AOL risk score: High or low risk group classification according to Adjuvant online software using the pre-defined cutoff.

ER, Estrogen receptors.

GRN, A network that has been inferred from gene expression data a gene regulatory network, briefly denoted as GRN.

ROC curve, Receiver operating characteristic curve.

AUC, Area under the curve.

ERP model, The final neural network trained from ER positive subgroup. ERN model: The final neural network trained from ER negative subgroup. COX model: The cox proportional hazard model

Availability of data and material

The Wang dataset supporting the conclusions of this article is available in the Gene Expression Omnibus repository (GEO series GSE2034).

The KJ125 dataset supporting the conclusions of this article is available in the Gene Expression Omnibus repository (GEO series GSE2990).

The TRANSBIG dataset supporting the conclusions of this article is available in the Gene Expression Omnibus repository (GEO series GSE7390).

Acknowledgments

None.

Conflicts of interest

The authors declares that there is no conflict of interest.

References

1. Van't Veer LJ, Dai H, Hart AA, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature.* 2002;415(6871):530–536.
2. Cardoso F, Van't Veer LJ, Bogaerts J, et al. 70-Gene Signature as an Aid to Treatment Decisions in Early- Stage Breast Cancer. *N Engl J Med.* 2016;375(8):717–729.

3. Wang Y, Zhang Y, Sieuwerts AM, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*. 2005;365(9460):671–679.
4. Desmedt C, Piette F, Wang Y, et al. Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clin Cancer Res*. 2007;13(11):3207–3214.
5. Yu J, Sieuwerts AM, Zhang Y, et al. Pathway analysis of gene signatures predicting metastasis of node-negative primary breast cancer. *BMC Cancer.* 2007;7:182.
6. Goeman JJ, Oosting J, Cleton Jansen AM, et al. Testing association of a pathway with survival using gene expression data. *Bioinformatics* 2005;21(9):1950–1957.
7. Hanahan D, Robert A Weinberg. Hallmarks of cancer: the next generation. *Cell*. 2011;144(5):646–674.
8. Mcshane LM, Polley MY C. Development of omics-based clinical tests for prognosis and therapy selection: the challenge of achieving statistical robustness and clinical utility. *Clin Trial.* 2013;10(5):653–665.
9. Sotiriou C, Wirapati P, Loi S, et al. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J Natl Cancer Inst.* 2006;98(4):262–272.
10. Srivastava Nitish, Hinton Geoffrey, Alex Krizhevsky, et al. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*. 2014;15:1929–1958.
11. Hinton Geoffrey, Srivastava Nitish, A. Krizhevsky, et al. Improving neural networks by preventing co-adaptation of feature detectors. *Neural and Evolutionary Computing* 2012;7.
12. Mosley J, Keri R. Intrinsic bias in breast cancer gene expression data sets. *BMC Cancer*. 2004;9:214.
13. Gruvberger SK, Ringner M, Marc Jvan de Vijver, et al. Expression profiling to predict outcome in breast cancer: the influence of sample selection. *Breast Cancer Res*. 2003;5(1):23–26.
14. Gruvberger S, Ringner M, Chen Y, et al: Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Res*. 2001;61(16):5979–5984.
15. Emmert Streib F, Dehmer M, Haibe Kains B. Gene regulatory network and their applications: understanding biological and medical problems in terms of networks. Frontiers in Cell and Development Biology. 2014;19(2):38.
16. Mandal S, Saha G, Rajat K Pal. A survey on recurrent neural network based modelling on gene regulatory network. *MOJ Proteomics Bioinformatics.* 2016;4(3):244–254.
17. Ahmad FK, Deris S, Othman NH. The inference of breast cancer metastasis through gene regulatory networks. *Journal of Biomedical Informatics. 2012;45:350–362.*
18. Zaminghalam K, Enayetallah A, Reddy P, et al. Robust clinical outcome prediction based on Bayesian analysis of transcriptional profiles and causal networks. *Bioinformatics*. 2014;30(12):69–77.
19. Khan A, Mandal S, Pal RK et al. Construction of gene regulatory network using recurrent neural network and swarm intelligence. *Scientifica(Cairo)*. 2016:1060843.
20. Mandal S, Khan A, Saha G, et al. Reverse engineering of gene regulatory networks based on S-systems and bat algorithm. *Journal Bioinform Comput Biol*. 2016;14(3):1650010.
21. Frohlich H, Balling R, Beerenwinkel N, et al. From hype to reality: data science enabling personalized medicine. *BMC Medicine.* 2018;16(1):150.
22. Jan RR Lewis, Kerridge I, Lipworth W. Use of real-world data for the research, development and evaluation of oncology precision medicine. *JCO Precis Onco*l. 2017:1-11.
23. Grossman RL, Heath AP, Ferretti V, et al. Toward a shared vision for cancer genomic data. *N Engl Med*. 2016;375:1109–1112.
24. Fakoor R, Ladhak F, Nazi A, et al. Using deep learning to enhance cancer diagnosis and classification. Atlanta, Georgia: Proceeding of the ICML workshop on the role of machine learning in transforming health care; 2013.
25. Djuric U, Zadeh G, Kenneth Aldape, et al. Precision histology: how deep learning is poised to revitalize histomorphology for personalized cancer care. *Npj Precision Oncology.* 2017;1(22).
26. Gupta A, Wang H, Madhavi Ganapathiraju, et al. Learning structure in gene expression data using deep architectures, with an application to gene clustering. *Bioinformatics*. 2015.
27. Hoffman MA, Williams MS. Electronic medical records and personalized medicine. *Hum Genet*. 2011;130(1):33–39.
28. Choi E, Bahadori MT, Schuetz A, et al. Doctor AI: predicting clinical events via recurrent neural networks. *JMLR Workshop Conf Proc.* 2016;56:301–318.
29. Sobradilo P, Pozo F, Agusti A. P4 medicine: the future around the corner. *Arch Bronconeumol.* 2011;47(1):35–40.
30. Mathur S, Sutton J. Personalized medicine could transform healthcare. *Biomed Rep.* 2017;7(1):3–5.

Citation: Yingying M, Xinping C. An artificial neural network framework for biomarker development to predict distant metastasis of breast cancer from gene expression data. *MOJ Proteomics Bioinformatics.* 2018;7(6):192–198. DOI: 10.15406/mojpb.2018.07.00257