TRIPOD Checklist: Prediction Model Development and Validation

Section/Topic	Item		Checklist Item	Page	Text extracts
Title and abstract					
Title	1	D;V	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	1	Identification and Validation of Tumor Microenvironment-Related Prognostic Biomarkers in Breast Cancer
Abstract	2	D;V	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	2,3	See entire abstract
Introduction					
Background and objectives	3a	D;V	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	3	The “seed and soil” hypothesis postulates that the tumor microenvironment provides fertile soil for the growth of tumor cells (3). Emerging evidence indicates that the cross-talk between tumor cells and tumor microenvironment exerts important effects on initiation, progression and metastasis of tumor (4). For example, tumors, as key drivers, control the differentiation of precursors of cancer-associated fibroblasts by secreting factors; Once present in the developing tumor, cancer-associated fibroblasts shape the tumor microenvironment to support tumor cell survival, dissemination, immune suppression, angiogenesis, and therapy resistance (5). Tumor microenvironment is a complex ecosystem of stromal cells and immune cells (6). Stromal cells and immune cells have been reported to have significant value in the diagnosis and prognosis of various cancers including breast cancer (7).
	3b	D;V	Specify the objectives, including whether the study describes the development or validation of the model or both.	3	we selected and validated prognostic biomarkers from cell scores for breast cancer using the Least absolute shrinkage and selection operator (LASSO) (9).
Methods					
Source of data	4a	D;V	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	4	We downloaded RNA-Seq data and clinical data for 1097 female breast cancer patients from the data portal for TCGA (accessed October 2020) (10). The data of 1904 breast cancer patients were obtained from METABRIC database. Datasets of GSE96058, GSE20194, GSE22358, GSE25066 and GSE32646 were downloaded from GEO.
	4b	D;V	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.		NA
Participants	5a	D;V	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.		NA
	5b	D;V	Describe eligibility criteria for participants.	4	Female breast cancer patients; Only 1063 breast cancer patients with a survival time longer than 0 days were included in present analysis.
	5c	D;V	Give details of treatments received, if relevant.		NA
Outcome	6a	D;V	Clearly define the outcome that is predicted by the prediction model, including how and when	5	The patients were separated into two groups according to expression level of a gene or the risk scores, and the median was used as cut-off. Then, the log-rank test was

			assessed.		used to assess the overall survival (OS) with survival package (R package version 3.1-7). Hazard ratios (HRs) and their 95% confidence intervals (CIs) were calculated using Cox proportional hazards.
	6b	D;V	Report any actions to blind assessment of the outcome to be predicted.		NA
Predictors	7a	D;V	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	4	Cell type enrichment analysis was performed using the xCell (R package version 1.1) (8) with gene expression data. We compared the cell scores of tumors to those of adjacent normal tissues. Violin plot was drew using vioplot package (version 0.3.5) in R software. The difference between two groups in cell scores was assessed using Wilcoxon test, and a P-value less than 0.05 was considered significant.
	7b	D;V	Report any actions to blind assessment of predictors for the outcome and other predictors.		NA
Sample size	8	D;V	Explain how the study size was arrived at.	4	We downloaded RNA-Seq data and clinical data for 1097 female breast cancer patients from the data portal for TCGA (accessed October 2020) (10). The data of 1904 breast cancer patients were obtained from METABRIC database. Datasets of GSE96058, GSE20194, GSE22358, GSE25066 and GSE32646 were downloaded from GEO.
Missing data	9	D;V	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	4	Only 1063 breast cancer patients with a survival time longer than 0 days were included in present analysis.
Statistical analysis methods	10a	D	Describe how predictors were handled in the analyses.	4	Cell type enrichment analysis was performed using the xCell (R package version 1.1) (8) with gene expression data.
	10b	D	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	5	We firstly calculated the cell scores using HTSeq – FPKM data from TCGA with xCell. Only 1063 breast cancer patients with a survival time longer than 0 days were included in present analysis. Then the patients were randomly separated patients into two sets, training set and test set. We performed LASSO Cox regression with cell scores of the training-set patients. Depending on the regulation weight λ, all regression coefficients are shrunken to towards zero in LASSO, and the irrelevant features are set exactly to zero. Risk scores were calculated by as our previously study (2, 11). We used “glmnet” package (R package version 4.0-2) to conduct the LASSO analysis and a P value < 0.05 was considered statistically significance. Receiver operating characteristic (ROC) curve was drawn and the corresponding area under the ROC curve (AUC) was calculated to evaluate the prognostic value of the risk score by using ROCR package (R package version 1.0-11).
	10c	V	For validation, describe how the predictions were calculated.	5	Risk scores were calculated by as our previously study (2, 11); The patients were separated into two groups according to expression level of a gene or the risk scores, and the median was used as cut-off. Then, the log-rank test was used to assess the overall survival (OS) with survival package (R package version 3.1-7). Hazard ratios (HRs) and their 95% confidence intervals (CIs) were calculated using Cox proportional hazards.

	10d	D;V	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	5	Receiver operating characteristic (ROC) curve was drawn and the corresponding area under the ROC curve (AUC) was calculated to evaluate the prognostic value of the risk score by using ROCR package (R package version 1.0-11). The patients were separated into two groups according to expression level of a gene or the risk scores, and the median was used as cut-off. Then, the log-rank test was used to assess the overall survival (OS) with survival package (R package version 3.1-7). Hazard ratios (HRs) and their 95% confidence intervals (CIs) were calculated using Cox proportional hazards.
	10e	V	Describe any model updating (e.g., recalibration) arising from the validation, if done.		NA
Risk groups	11	D;V	Provide details on how risk groups were created, if done.	5	The patients were separated into two groups according to expression level of a gene or the risk scores, and the median was used as cut-off.
Development vs. validation	12	V	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	5	Then the patients were randomly separated patients into two sets, training set and test set.
Participants	13a	D;V	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	4	Only 1063 breast cancer patients with a survival time longer than 0 days were included in present analysis.
	13b	D;V	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	4	Female breast cancer patients; Only 1063 breast cancer patients with a survival time longer than 0 days were included in present analysis.
	13c	V	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).		NA
Model development	14a	D	Specify the number of participants and outcome events in each analysis.	4	Only 1063 breast cancer patients with a survival time longer than 0 days were included in present analysis.
	14b	D	If done, report the unadjusted association between each candidate predictor and outcome.		NA
Model specification	15a	D	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	6	We obtained four biomarkers including myocytes, natural killer T cell (NKT), conventional dendritic cell (cDC) and sebocytes, and their coefficients were 0.098, -0.131, -0.021 and 0.012.
	15b	D	Explain how to the use the prediction model.	5	The patients were separated into two groups according to expression level of a gene or the risk scores, and the median was used as cut-off. Then, the log-rank test was used to assess the overall survival (OS) with survival package (R package version 3.1-7). Hazard ratios (HRs) and their 95% confidence intervals (CIs) were calculated

					using Cox proportional hazards.
Model performance	16	D;V	Report performance measures (with CIs) for the prediction model.	6	The constructed risk score model divided patients into low-and high-risk groups according to their risk score. Low risk was associated with longer survival times and less deaths (Figure 2a). Survival analysis indicated that high risk patients had a HR of 1.75 (95% CI: 1.08 − 2.83; P=0.022) (Figure 2b). ROC curve analysis indicates the AUC was 0.90, 0.63 and 0.58 for 3, 5 and 10 year survival (Figure 2c). Analysis of the test cohort corroborated the findings in training cohort (Figure 2d). The HR of high-risk patients was 1.68 (95% CI: 1.07 − 2.66, P = 0.024; Figure 2e) compared to low-risk score patients; The AUC was 0.71, 0.66 and 0.65 for 3, 5 and 10 year survival (Figure 2f).
Model-updating	17	V	If done, report the results from any model updating (i.e., model specification, model performance).		NA
Limitations	18	D;V	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	10	When interpreting the results, the limitations of the present study need to be considered. We did not validate our results using experimental method. The mechanisms underlying cDCs are needed to further confirmed. We raised cDCs as a key signature, however, myocytes, NKT and sebocytes do indeed have their effects on breast cancer paitents.
Interpretation	19a	V	For validation, discuss the results with reference to performance in the development data, and any other validation data.	8-10	See discussion
	19b	D;V	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	10	In conclusion, cDC score was a key signature predicting prognosis for breast cancer and high cDC score was associated with elevated immune activity and better prognosis of breast cancer. cDCs may exert antitumor effects by upregulating IL-2.
Implications	20	D;V	Discuss the potential clinical use of the model and implications for future research.	10	cDC score was a key signature predicting prognosis for breast cancer and high cDC score was associated with elevated immune activity and better prognosis of breast cancer; The mechanisms underlying cDCs are needed to further confirmed.
Supplementary information	21	D;V	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	4	We downloaded RNA-Seq data and clinical data for 1097 female breast cancer patients from the data portal for TCGA (accessed October 2020) (10). The data of 1904 breast cancer patients were obtained from METABRIC database. Datasets of GSE96058, GSE20194, GSE22358, GSE25066 and GSE32646 were downloaded from GEO.
Funding	22	D;V	Give the source of funding and the role of the funders for the present study.	10	Funding: This work was supported by the National Natural Science Foundation of China [grant number 81602551, 81872485], the High-level Innovative and Entrepreneurial Talent Introduction Plan of Jiangsu Province [303073540ER21] and the young talents program of Jiangsu Cancer Hospital [QL201810].

TRIPOD Checklist: Prediction Model Development and Validation

*Items relevant only to the development of a prediction model are denoted by D, items relating solely to a validation of a prediction model are denoted by V, and items relating to both are denoted D;V. We recommend using the TRIPOD Checklist in conjunction with the TRIPOD Explanation and Elaboration document.

Article information: https://dx.doi.org/10.21037/tcr-21-1248