75681_Auto_Edite.docx

Retrospective Cohort Study

iCEMIGE: Integration of CEll-morphometrics, MIcrobiome, and GEne biomarker signatures for risk stratification in breast cancers

Xuanyu Mao, Jesus Perez-Losada, Mar Abad, Marta Rodríguez-González, Cesar A. Rodríguez, Jian-Hua Mao, Hang Chang

Abstract

BACKGROUND

The development of precision medicine is essential for personalized treatment and improved clinical outcome, whereas biomarkers are critical for the success of precision therapies.

AIM

We aimed to investigate whether iCEMIGE (**i**ntegration of **CE**ll-morphometrics, **MI**crobiome, and **GE**ne biomarker signatures) improves risk stratification of breast cancer (BC) patients.

METHODS

We used our recently developed machine learning technique (CMS-ML) to identify cellular morphometric biomarkers (CMBs) from the whole histological slide images in The Cancer Genome Atlas (TCGA) breast cancer (TCGA-BRCA) cohort. Multivariate Cox regression was used to assess whether cell-morphometrics prognosis score (CMPS) and our previously reported 12-gene expression prognosis score (GEPS) and 15-microbe abundance prognosis score (MAPS) were independent prognostic factors. iCEMIGE was

built upon the sparse representation learning technique. The iCEMIGE scoring model performance was measured by the area under the ROC curve compared to CMPS, GEPS, or MAPS alone. Nomogram models were created to predict overall survival (OS) and progress-free survival (PFS) rates at 5- and 10-year in the TCGA-BRCA cohort.

RESULTS

We identified 39 CMBs that were used to create a CMPS system in BCs. CMPS, GEPS, and MAPS were found to be significantly independently associated with OS. We then established an iCEMIGE scoring system for risk stratification of BC patients. The iGEMIGE score has a significant prognostic value for OS and PFS independent of clinical factors (age, stage, and estrogen and progesterone receptor status) and PAM50-based molecular subtype. Importantly, the iCEMIGE score significantly increased the power to predict OS and PFS compared to CMPS, GEPS, or MAPS alone.

CONCLUSION

Our study demonstrates a novel and generic AI framework for multimodal data integration toward improving prognosis risk stratification of BC patients, which can be extended to other types of cancer.

INTRODUCTION

Cancer is a complex and heterogeneous disease that displays many morphological, genetic, and epigenetic features [1]. Cancer heterogeneity consistently results in a large variation in clinical outcomes of patients after a certain treatment [2], and therefore the development of precision medicine is essential for personalized treatment and improved clinical outcome [3-6]. The discovery of biomarkers for predicting prognosis, a critical step toward precision medicine, can significantly assist clinical oncologists in making treatment decisions for cancer patients [7-9].

Microscopic examination of the histology, which encompasses the morphological features of cancer cells, is the oldest and most basic way of cancer classification. A complete and accurate pathological cancer classification is still crucial to deciding on the best treatment plan for patients. Recently, we developed a framework powered by artificial intelligence (AI) technique for identifying cellular morphometric biomarkers (CMBs) and cellular morphometric subtypes (CMSs) from the whole slide images (WSI) of Hematoxylin and Eosin (H&E)-stained tissue histology [10, 11]. We demonstrated that CMSs were significantly associated with specific molecular alterations, immune microenvironment, and prognosis in lower-grade gliomas [10].

With the rapid biotechnological development, such as next-generation sequencing, different aspects of genomic heterogeneity have been uncovered in cancers [12], which dramatically speed the discovery of molecular biomarkers for precision diagnosis and therapy. For example, several molecular biomarkers have been developed for clinical practice in breast cancer (BC) [13, 14], including PAM50 (Prosigna, South San Francisco, USA), OncotypeDx (Exact Sciences Corp., Madison, USA), and MammaPrint (Agendia, Amsterdam, Netherlands).

In addition to cancer genomic heterogeneity, a significant number of studies have revealed the diversity of the microbiome in cancer and the roles of the microbiome in cancer development and response to therapies [15-18]. We have recently developed a novel cancer microbiome signature for predicting the prognosis of BC patients [19]. Given the importance of tissue histology, genomics, and microbiome in cancer diagnosis and treatment, efficient and effective integration of these multimodal data is believed to open a new era for precision oncology [20].

In this study, we developed a strategy to integrate multimodal data (**Fig. 1**) and investigated whether iCEMIGE (**i**ntegration of **ce**ll-morphometrics, **mi**crobiome, and **ge**ne biomarker signatures) improves the risk stratification of BC patients. We first used our recently developed machine learning technique (CMS-ML) to identify the CMBs from the WSIs in The Cancer Genome Atlas (TCGA) breast cancer (TCGA-BRCA) cohort and established a cellular-morphometrics prognosis score (CMPS). We then demonstrated

that CMPS, together with our previously reported 12-gene expression prognosis score (GEPS) [21] and the 15-microbe abundance prognosis score (MAPS) [19] were independent prognostic factors. Finally, we established the iCEMIGE scoring system and assessed its clinical value and prognosis predictive power compared to GEPS, MAPS, and CMPS alone.

MATERIALS AND METHODS

Study design and dataset

The TCGA-BRCA cohort was used in this study. The patient diagnostic tissue histology slides were downloaded from GDCportal (https://portal.gdc.cancer.gov/). TCGA-BRCA microbiome, transcriptome, and clinical data, including PAM50-based molecular subtypes, were downloaded from the cBioPortal (https://www.cbioportal.org/) [22, 23]. No additional modifications were made to the downloaded data during our analyses.

Extraction of cellular morphometric characteristics and stratification of breast cancer patients

Following our previous work [10], we deployed an unsupervised feature learning pipeline, which was based on the stacked predictive sparse decomposition (SPSD) [24, 25], for unsupervised discovery of underlying cellular morphometric characteristics from 15 cellular morphological features that were extracted from the diagnostic slides from the TCGA-BRCA cohort. 256 cellular morphometric biomarkers (CMB) were defined for cellular object representation. Specifically, we used a single network-layer with 256 dictionary elements (i.e., CMBs) and a sparsity constraint of 30 at a fixed random sampling rate of 1000 cellular objects per WSIs from the TCGA-BRCA cohort. The pre-trained SPSD model reconstructed each cellular region (represented as a vector of 15 morphometric properties) as a sparse combination of pre-defined 256 CMBs and thereafter represents each patient as an aggregation of all delineated cellular objects belonging to the same patient.

The prognostic effect of high or low levels of each CMB on overall survival (OS) was assessed by Kaplan-Meier analysis (survminer package in R, Version 0.4.8) and log-rank test (survival package in R, Version 3.2-3), where the TCGA-BRCA cohort was divided into two groups (i.e., CMB-high and CMB-low groups) based on each CMB (survminer package in R, Version 0.4.8). The set of CMBs as a prognostic signature were selected *via* a multivariate CoxPH regression model including these CMBs with a significant effect on OS.

Finally, we calculated the cellular morphometric prognosis score (CMPS) using the formula below, where the coefficients of the final CMBs as categorical variables were obtained from multivariate CoxPH regression analysis:

$$\text{CMPS} = \sum_{i=1}^{N} \left(coefficient\ of\ CMB\_Category^i\right) * \left(CMB\_Category^i\right)$$

Where N is the number of final CMBs that were independently and significantly associated with OS, and $CMB\_Category^i$ is the category of the ith CMB (i.e., CMB-high: 1; CMB-low: 0).

Mining of multi-modal iCEMIGE biomarker signature

We extended the unsupervised feature learning pipeline (SPSD) [24, 25] to achieve efficient and effective mining of multi-modal biomarker signatures from prebuilt cellular-morphometrics, microbiome, and gene biomarkers. Given $\mathbf{X} = [\mathbf{x_1}, \dots, \mathbf{x_N}] \in \mathbf{R}^{m \times N}$ as a set of patients (N) with a combination of biomarkers from different modalities (i.e., cellular-morphometrics, microbiome, and gene biomarkers), the formulation of the iCEMIGE multi-modal biomarker mining model was defined as follows.

$$min_{\mathbf{B,Z,W,G}}\ \|\mathbf{X} - \mathbf{BZ}\|_F^2\ + \|\mathbf{Z} - \mathbf{G\sigma(WX)}\|_F^2\ + \lambda_1 \|\mathbf{Z}\|_1$$

$$s.t.\ \|b_i\|_2^2 = 1, \forall i = 1, \dots, h$$

where $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_h] \in \mathbf{R}^{m \times h}$ was a set of multi-modal biomarkers to be mined. Each multi-modal biomarker $(\mathbf{b})$ was composed of m individual biomarker (e.g., m=66 in our study); $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_N] \in \mathbf{R}^{h \times N}$ was the sparse multi-modal biomarker expression matrix, where $\mathbf{z}_i$ was the sparse multi-modal biomarker expression profile of the original patient biomarkers $(\mathbf{x}_i)$, consisting of relative abundances of all (h) multi-modal biomarkers that contributed to the reconstruction of $\mathbf{x}_i$; $\mathbf{W} \in \mathbf{R}^{h \times m}$ was the auto-encoder for efficient and effective extraction of sparse multi-modal biomarker expression matrix $(\mathbf{Z})$ from original patient biomarker data $(\mathbf{X})$; $\mathbf{G} = diag(g_1, \ldots, g_h) \in \mathbf{R}^{h \times h}$ was a scaling matrix with $diag$ being an operator aligning vector $[g_1, \ldots, g_h]$, along the diagonal; $\sigma(\cdot)$ was an element-wise sigmoid function; λ_1 was the regularization constant to ensure the sparsity of $\mathbf{Z}$, such that only a subset of multi-modal biomarkers was utilized during the reconstruction of original patient biomarker data.

The first constraint: $\|\mathbf{X} - \mathbf{BZ}\|_F^2$, penalized the reconstruction error of original patient biomarker data $(\mathbf{X})$ with multi-modal biomarker $(\mathbf{B})$ and the corresponding sparse multi-modal biomarker expression matrix $(\mathbf{Z})$, which helped minimize the loss of individual biomarker information; the second constraint: $\|\mathbf{Z} - \mathbf{G}\sigma(\mathbf{WX})\|_F^2$, penalized the approximation error of sparse multi-modal biomarker expression matrix $(\mathbf{Z})$ with the auto-encoder, which helped improve the accuracy of multi-modal biomarker extraction for new patients; the third constraint: $\|\mathbf{Z}\|_1$, penalized the sparsity of the multi-modal biomarker expression matrix, which helped ensure the utilization/activation of dominant multi-modal biomarkers during the learning process.

Construction of the iCEMIGE score

After multi-modal biomarker mining (i.e., 256 multi-modal biomarkers mined in this study), a multivariate Cox regression was performed on 256 multi-modal biomarker signatures, defined as 256 covariates using the TCGA-BRCA dataset. The iCEMIGE score of each patient was calculated by the following formula:

$$iCEMIGE\ score = \sum_{i=1}^{256} (\ covariate\ i\ coefficient) * (\ covariate\ i\ expression\ level)$$

Nomogram, ROC and C-index

A nomogram model (rms package in R, Version 6.0-1) was constructed to predict 5- and 10-year OS probability of BC patients. The time-dependent receiver operating characteristic (ROC) curve (survivalROC package in R, Version 1.0.3) and concordance index (C-index) were used to evaluate the performance of the nomogram model, where the C-index was repeated with 1000 bootstrapping iterations and an 80% sampling rate per iteration. Mann-Whitney non-parametric test was used for the comparison across models.

Statistical analysis

The cohort of patients were divided into three groups (Poor: top third; Intermediate: middle third; and Good: bottom third) based on CMPS or iCEMIGE score. The independent prognostic impact of different scores (CMPS and iCEMIGE) was assessed by multivariate CoxPH regression including the clinical factors (age, stage, ER, and PR status) and PAM50-based molecular subtype. All statistical analyses were performed through either SPSS 24.0 (IBM, NY, USA) or R (version 4.0.2, https://www.r-project.org/). Graphic visualizations were generated by R (ggpubr package, Version 0.4.0; ggplot2 package, Version 3.3.3) or SPSS. The statistical significance was defined as p<0.05 (two-tails).

RESULTS

Identifying Cellular Morphometric Biomarkers for prognosis of BC Patients

Over 300 million cellular objects from 1085 diagnostic slides of 1017 TCGA-BRCA patients were recognized and delineated by an unsupervised feature learning pipeline based on SPSD [24]. Each cellular object was represented with 15 morphometric properties as described in our previous work [10].

Next, we optimized and trained our SPSD model based on pre-quantified cellular objects randomly selected from the TCGA-BRCA cohort to discover the underlying cellular morphometric biomarkers (CMBs). After training, the prebuilt SPSD model reconstructed each cellular object as a sparse combination of the pre-identified 256 cellular morphometric biomarkers, which led to the novel representation of every single cellular object as 256 sparse code (reconstruction coefficient); and thereafter, the corresponding 256-dimensional cellular morphometric context representation of each patient as an aggregation of all delineated cellular objects belonging to the same patient (Supplementary Table S1). The final patient-level cellular morphometric context representation consisted of 256 CMBs.

We next evaluated the association of 256 CMBs with OS in the TCGA-BRCA cohort. Survival analysis revealed that 148 of 256 CMBs had a significant prognostic impact ($p < 0.05$, Supplementary Table S2). Among these 148 CMBs, 39 CMBs demonstrated independent and significant association with OS by multivariate CoxPH regression analysis (**Fig. 2A;** Supplementary Figure S1; Supplementary Table S3), which were defined as a 39-CMB signature.

Assessing Prognostic Value of the 39-CMB Signature

To further evaluate the prognostic value of the 39-CMB signature, we constructed the cellular morphometric prognosis score (CMPS) (see Methods) and divided TCGA-BRCA cohort into three groups (Poor: top third; Intermediate: middle third; and Good: bottom third) based on CMPS (Supplementary Table S4). Patients with good scores had significantly longer OS than those with poor scores. The OS of patients with intermediate scores was between these two groups ($P = 1.61E\text{-}23$, **Fig. 2B**). Moreover, CMPS provided additional prognostic value to clinical factors (age, ER, PR, and stage) and PAM50-based molecular subtypes (**Fig. 2C**).

Establishing the iCEMIGE prognostic model

Omics analyses of cancers have further revealed their genomic heterogeneity. FDA has approved many genomic biomarkers for clinical use, such as PAM50. Based on the omics data, we have previously identified 12-gene [21] and 15-microbe signatures [19] for the prognosis of BC patients (Supplementary Table S3). We conducted a multivariate Cox regression analysis to address whether GMPS, MAPS, and GEPS are independent prognostic factors. Indeed, CMPS, MAPS, and GEPS were significantly and independently associated with OS (**Fig. 2D**). We then integrated 39 CMBs, 15 microbes, and 12 genes in an unsupervised representation framework ("iCEMIGE") and mined 256 multi-modal biomarkers (Supplementary Table S3) with experimentally optimized parameters for C-index for OS (Supplementary Figure S3). The optimal iCEMIGE score was then constructed to assess a patient's risk for death and disease progression (Supplementary Table S4, details see Materials and Methods).

Evaluating the prognostic value of the iCEMIGE score

A total of 919 BC patients in the TCGA-BRCA cohort with full signature (iCEMIGE) data were included in this evaluation (Supplementary Table S5). 919 BC patients were stratified into different prognostic groups (Poor: top third; Intermediate: middle third; and Good: bottom third) according to the iCEMIGE score. Patients within the poor prognosis group had significantly shorter OS compared to those within the intermediate and good prognosis groups ($P = 4.02E-58$, **Fig. 3A**). Importantly, we showed that the iCEMIGE score was more effective in predicting OS of BC patients than CMPS, MAPS, and GEPS alone (**Fig. 3B** and **C**; Supplementary Figure S2A and B). Moreover, we found that the iCEMIGE score was also significantly associated with PFS ($P = 2.40E-19$, **Fig. 3D**) and had more effective in predicting PFS (**Fig. 3E** and **F**; Supplementary Figure S2C and D).

We then evaluated whether the prognostic value of the iCEMIGE score was independent of ER status, stage, and molecular subtypes. As shown in **Fig. 4A**, patients with poor iCEMIGE scores had significantly shorter OS and PFS compared to those with good iCEMIGE scores in both ER+ and ER- groups. Moreover, the iCEMIGE score was

significantly associated with OS and PFS in all different stages (**Fig. 4B**) and subtypes (**Fig. 5**).

Finally, using multivariate Cox regression analyses (including pathological stage, age, PR status, ER status, molecular subtype, iCEMIGE), we demonstrated that iCEMIGE was an independent prognostic factor for both OS (**Fig. 6A**) and PFS (Supplementary Figure S4A). These findings indicate that the iCEMIGE score has an independent prognostic value in BCs.

To further assess the clinical value of the iCEMIGE score, we established a nomogram model, a valuable clinical tool for prognosis prediction, where we integrated iCEMIGE with clinical factors (age, stage, ER, and PR), PAM50-based molecular subtypes to predict the 5- and 10-year OS probability of BC patient (**Fig. 6B**). The iCEMIGE score significantly improved the predictive power of prognosis (**Fig. 6C**). Similar results were found for PFS (Supplementary Figure S4B and S4C).

DISCUSSION

High BC heterogeneity brings up a significant challenge for predicting a patient's response to treatment or prognosis. In this study, we established a new strategy for tackling this challenge by integrating multimodal signatures and demonstrated that such approach significantly improved the power for prognostic prediction compared to the single modal biomarker. In addition, we showed that iCEMIGE is significantly superior in predicting OS and PFS compared to the PAM50-based molecular subtype in the TCGA-BRCA cohort, although additional validation is required, as stated later in the limitations of this study.

The majority of biomarker developments are limited to a single modal data [20]. In the past, we followed the same path to define the 12-gene expression prognosis score (GEPS) [21] and the 15-microbe abundance prognosis score (MAPS) [19] in BC. Here, we developed the 39-CMB prognosis score (CMPS) using an AI-driven CMB detection technique [10]. We found that CMPS, MAPS, and GEPS had an independent prognostic value. This suggests that different modal data provide unique clinical value for prognosis

prediction and raises the possibility that integrating multimodal biomarkers can advance precision oncology by more accurately predicting the risk of treatment failure, relapse *etc.*

Integrating multimodal data to yield improved performance compared with each modality alone remains challenging. In this study, we presented a multi-step approach to integrate cellular morphometric, molecular, and microbiome landscapes into a multimodal prognostic system for BC. Firstly, we identified the biomarker signature and systematically assessed its prognostic value in each type of modal data. Secondly, we investigated whether these modal-specific biomarker signatures are independent prognostic factors. Thirdly, we established the final predictive model incorporating all modal biomarker signatures with significantly improved prognostic risk stratification compared with each modality alone. Finally, we systematically evaluated the clinical value of the final predictive model. Such a strategy can extend to other types of cancers.

Modern clinical instruments are generating massive amounts of multimodal data, including radiology, histology, and molecular data, where each of them provides unique value for cancer diagnosis and treatment. Therefore, the efficient and effective integration of multimodal data becomes critical and, however, remains challenging in terms of robustness, interpretability, and translational impact, even with the current advancesin artificial intelligence techniques [26-28]. Two major trends in multimodal integration in cancer research are modal-specific raw data integration (MDI) [29, 30] and modal-specific representation integration (MRI) [31, 32]. The MDI strategy handles each modality (e.g., histology and genomics) using different neural network structures and then combines the corresponding output of each neural network branch in subsequent network layers to predict the health outcome. Trained in an end-to-end fashion (i.e., black-box fashion), this strategy delivers a convenient and powerful utilization of information and interaction across modalities; however, in general, it lacks biomedical interpretability. In addition, such a strategy does not guarantee the learning of clinically significant and independent information per each modality, and thus the alternative deployment of an individual modality or a subset of modalities is nearly impossible.

In contrast, the MRI provides a stepwise strategy, where the first step consists of outcome-driven representation mining per modality, and the second step integrates modal-specific representation towards the outcome. Obviously, MRI is more likely (without guarantee) to mine model-specific representation with independent clinical value *via* a stepwise mechanism and consequently provides more flexibility in individual/subset modality deployment. This flexibility is important in clinical practice, especially when all modalities are not available. Extended from the MRI strategy, our work realizes the modal-specific knowledge integration (MKI) by enforcing the mining and utilization of biomedically interpretable, clinically significant and independent, and double-blindly validated knowledge (i.e., cellular morphometric biomarkers, microbiome biomarkers, and genomic biomarkers) through an AI-powered systems biology workflow for maximized clinical implications and translation impact.

Our study established a new promising strategy for integrating multimodal data to enhance prognostic prediction. A significant limitation was that we did not have independent cohorts to validate our findings. In addition, due to the limited clinical information in the TCGA-BRCA cohort, we were unable to comprehensively explore the potential confounding clinical factors, including tumor size, different cancer treatments, *etc*. The clinical utility of iCEMIGE needs to be further validated in retrospective and prospective cohort studies to determine whether the iCEMIGE score can provide sufficient predictive information to stratify patients by risk and guide treatment. If so, the iCEMIGE score could assist clinicians in decision-making about cancer treatment and enable more personalized cancer therapy.

CONCLUSION

Our study demonstrates a novel and generic AI framework for multimodal data integration toward improving prognosis risk stratification of BC patients, which can be extended to other types of cancer.

ARTICLE HIGHLIGHTS

Research objectives

To develop a strategy to integrate multimodal data and to investigate whether iCEMIGE (**i**ntegration of **ce**ll-morphometrics, **mi**crobiome, and **ge**ne biomarker signatures) improves the risk stratification of breast cancer patients.

Research motivation

Modern clinical instruments are generating massive amounts of multimodal data, including radiology, histology, and molecular data, where each of them provides unique value for cancer diagnosis and treatment. Efficient and effective integration of these multimodal data is believed to open a new era for precision oncology.

Research background

Cancer heterogeneity consistently results in a large variation in clinical outcomes of patients after treatment. The discovery of biomarkers for tailoring cancer treatments is a critical step toward personalized medicine.

Research perspectives

The iCEMIGE score could assist clinicians in decision-making about cancer treatment and enable more personalized cancer therapy.

Research conclusions

Our study indicates that multimodal integration (iCEMIGE) can more accurately predict the prognostic risk of breast cancer patients.

Research results

iCEMIGE is significantly superior in predicting overall and progression-free survival of breast cancer patients compared to single modal biomarker and the PAM50-based molecular subtype, which is one of FDA approved biomarkers and is currently used in clinical practice.

Research methods

The artificial intelligence pipeline powered is used to identify cellular morphometric biomarkers. Single modal biomarker signatures are integrated using the sparse representation learning technique to establish iCEMIGE. Clinical value of iCEMIGE is evaluated using different statistical methods.

75681_Auto_Edite.docx

PRIMARY SOURCES

1. Xiao-Ping Liu, Xiaoqing Jin, Saman Ahmadian, Xu Yang et al. "Clinical Significance and Molecular Annotation of Cellular Morphometric Subtypes in Lower Grade Gliomas discovered by Machine Learning: a retrospective multicentric study", Research Square Platform LLC, 2021
Crossref Posted Content
113 words — 3%

2. assets.researchsquare.com
Internet
108 words — 3%

3. link.springer.com
Internet
97 words — 3%

4. www.ncbi.nlm.nih.gov
Internet
57 words — 2%

5. Ju Han, Gerald V Fontenay, Yunfu Wang, Jian-Hua Mao, Hang Chang. "Phenotypic characterization of breast invasive carcinoma via transferable tissue morphometric patterns learned from glioblastoma multiforme", 'Institute of Electrical and Electronics Engineers (IEEE)', 2016
Internet
22 words — 1%

6. A. W. Mao, H. Barck, J. Young, A. Paley, J. -H. Mao, H. Chang. "Identification of a novel cancer microbiome signature for predicting prognosis of human breast cancer patients", Clinical and Translational Oncology, 2021
Crossref
13 words — < 1%