Published in final edited form as:

Clin Radiol. 2018 September ; 73(9): 782–791. doi:10.1016/j.crad.2018.04.009.

Integrative radiomics expression predicts molecular subtypes of primary clear cell renal cell carcinoma

Q. Yin^{a,b,*}, S.-C. Hung^{b,c,d,e,*}, W. K. Rathmell^{f,g,h,i}, L. Shen^a, L. Wang^b, W. Lin^b, J. R. Fielding^j, A. H. Khandani^b, M. E. Woods^{f,k}, M. I. Milowsky^{f,g,k}, S. A. Brooks^f, E. M. Wallen^k, and D. Shen^{b,l,†}

^aInformation Science and Technology College, Dalian Maritime University, Dalian, 116026, China

^bDepartment of Radiology, University of North Carolina, Chapel Hill, NC 27599, USA

^cDepartment of Radiology, Taipei Veterans General Hospital, Taipei, Taiwan 11217

^dSchool of Medicine, National Yang-Ming University, Taipei, Taiwan 11221

^eDepartment of Biomedical Imaging and Radiological Sciences, School of Biomedical Science of Engineering, National Yang-Ming University, Taipei, Taiwan 11221

^fLineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC 27599, USA

[†]Guarantor and correspondent: D. Shen, Information Science and Technology College, Dalian Maritime University, Dalian, 116026, China., address: dgshen@med.unc.edu.

*Co-first authors.

Author Contributions

1. Guarantor of integrity of the entire study
Qingbo Yin, Sheng-Che Hung, Dinggang Shen
2. Study concepts and design
Qingbo Yin, Sheng-Che Hung, W. Kimryn Rathmell, Liran Shen, Weili Lin, Dinggang Shen
3. Literature research
Qingbo Yin, Sheng-Che Hung, W. Kimryn Rathmell
4. Clinical studies
Michael E. Woods, Matthew I Milowsky, Eric. M. Wallen
5. Experimental studies/data analysis
Julia R. Fielding, Amir H. Khandani, Samira A. Brooks, Li Wang, Qingbo Yin
6. Statistical analysis
Qingbo Yin, Sheng-Che Hung, Liran Shen
7. Manuscript preparation
Qingbo Yin, Sheng-Che Hung, Liran Shen, W. Kimryn Rathmell
8. Manuscript editing
Qingbo Yin, Sheng-Che Hung, W. Kimryn Rathmell, Liran Shen, Dinggang Shen

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

^gDepartment of Medicine, Vanderbilt University Medical Center, Nashville, TN 37232, USA

^hDepartment of Cancer Biology, Vanderbilt University, Nashville, TN 37232, USA

ⁱVanderbilt-Ingram Cancer Center, Vanderbilt University, Nashville, TN 37232, USA

^jDepartment of Radiology, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA

^kDepartment of Urology, University of North Carolina, Chapel Hill, NC 27599, USA

^lDepartment of Brain and Cognitive Engineering, Korea University, Seoul 02841, Republic of Korea

Abstract

AIM—To identify combined positron-emission tomography (PET)/magnetic resonance imaging (MRI)-based radiomics as a surrogate biomarker of intratumour disease risk for molecular subtype ccA and ccB in patients with primary clear cell renal cell carcinoma (ccRCC).

MATERIALS AND METHODS—PET/MRI data were analysed retrospectively from eight patients. One hundred and sixty-eight radiomics features for each tumour sampling based on the regionally sampled tumours with 23 specimens were extracted. Sparse partial least squares discriminant analysis (SPLS-DA) was applied to feature screening on high-throughput radiomics features and project the selected features to low-dimensional intrinsic latent components as radiomics signatures. In addition, multilevel omics datasets were leveraged to explore the complementing information and elevate the discriminative ability.

RESULTS—The correct classification rate (CCR) for molecular subtype classification by SPLS-DA using only radiomics features was 86.96% with permutation test $p = 7 \times 10^{-4}$. When multi-omics datasets including mRNA, microvascular density, and clinical parameters from each specimen were combined with radiomics features to refine the model of SPLS-DA, the best CCR was 95.65% with permutation test, $p < 10^{-4}$; however, even in the case of generating the classification based on transcription features, which is the reference standard, there is roughly 10% classification ambiguity. Thus, this classification level (86.96–95.65%) of the proposed method represents the discriminating level that is consistent with reality.

CONCLUSION—Featured with high accuracy, an integrated multi-omics model of PET/MRI-based radiomics could be the first non-invasive investigation for disease risk stratification and guidance of treatment in patients with primary ccRCC.

INTRODUCTION

Translation of gene-based prognostic markers into clinical practice is an important step towards personalised medicine. Renal cell carcinomas (RCC) are a notoriously heterogeneous group of diseases (1, 2), and are known for substantial variability with regard to intratumour mutations and branched evolutionary growth (3, 4). The ClearCode34 transcript signature, which is validated by a 34-gene expression panel, divides clear cell (cc)RCC into two subtypes (ccA/ccB) and has been proven to be a robust prognostic predictor for localised and metastatic RCCs (5, 6). These two groups present with different

biological features. ccA tumours have higher angiogenesis. ccB tumours have higher transforming growth factor β (TGF- β) and epithelial-to-mesenchymal signalling expressions, higher Fuhrman nuclear grade, and larger tumours, and are also more likely to have nodal involvement or metastatic tumours. Importantly, ccB patients have worse overall survival, cancer-specific survival, and recurrence-free survival (7–9). Prognostic models containing this signature have consistently outperformed other pathological features, such as Fuhrman grade, TNM stage, molecular analyses, and nomograms (7–9); however, the applicability of this biomarker is hindered by the high intratumour heterogeneity in ccRCC (3). Gulati *et al.* had demonstrated the intratumour ccA/ccB heterogeneity of primary ccRCC tumours by multiple biopsy samples from each tumour. Eighty percent of the primary tumours consist of a varied percentage of both ccA and ccB subtypes (6). Thus, it has been concluded that multiple samples from every tumour would need to be analysed in order to identify poor prognostic clones (3, 6, 10, 11). The aim of this study was to resolve the issue of heterogeneity by establishing a holistic imaging-based non-invasive ccA/ccB classification model to identify intratumour variation using multiple tumour sampling based on the regionally sampled tumours characterised by ccA/ccB subtypes, and set the stage for future applications of radiometrics in providing a holistic model for disease characterisation.

To link phenotypes with intratumour gene features, radiomics studies, which decode tumour phenotypes into mineable quantitative image features, have successfully interlinked non-invasive imaging features and key cancer genetic mutations in ccRCC (12, 13).

Radiogenomics has remarkable potential to accelerate precision medicine, but the development is still in its early phase with moderate predictive performance (14). The emergence of hybrid positron-emission tomography/magnetic resonance imaging (PET/MRI) systems offers a unique opportunity to obtain combined anatomical details using MRI and metabolic information from PET to better delineate disease functional processes (15–17). Meanwhile, transcriptional outputs of tumours are expressed dynamically and quantitatively relative to physiology, which provides complementary and interchangeable information. To this end, the aim of the study was to leverage the non-invasive high-throughput radiomics information of hybrid PET/MRI to establish a model to discern regional molecular subtype information. Moreover, multi-omics data from regional radiomics, histological information, gene and protein expression, and clinical parameters were involved to refine the model and boost its discriminating ability.

MATERIALS AND METHODS

Study design

The aim of the present study was to resolve the issue of heterogeneity by establishing a holistic imaging-based non-invasive ccA/ccB classification model to identify intratumour variation using multiple tumour sampling based on the regionally sampled tumours characterised by molecular subtypes ccA/ccB. To do this, the whole process of this study involved in three phases as shown in Fig. 1. Phase 1 included a series of successive imaging preprocessing, i.e., imaging, multimodality registration, and marked tumour sampling. In phase 2, high-throughput radiomics features were extracted from PET/MRI. In phase 3, model building was undertaken, as shown in Fig. 1b. Initially, feature screening was used on

these radiomics features, and supervised learning was applied to project the selected features onto the low-dimensional intrinsic latent components as radiomics signatures, which were strongly predictive of ccA/ccB class. Second, multilevel omics datasets were leveraged to explore the complementing information and elevate the discriminative ability. In this step, five omics datasets were used to rectify the signatures for the classification modal. The details of phases 1 and 2 have been described previously (18, 19), and phase 3 is elaborated in the following sections.

Overview of patient cohorts and clinical samples

This study was reviewed and approved by the institutional review board. This work used the same data with reference (19). Between 2012 and 2013, 23 specimens were collected from eight primary ccRCCs. Tumour, node, and metastasis (TNM) stage was classified using the American Joint Committee on Cancer's Cancer Staging 7th edition. All patients underwent PET/MRI using a standard PET protocol and a three-dimensional (3D)-Dixon and dynamic contrast-enhanced (DCE) T1-weighted (W) MRI as previously described (18, 19).

Preoperative PET/MRI images were reviewed jointly by a board-certified radiologist, a nuclear medicine specialist, and the medical team using a standardised hanging protocol. Several regions of interest (ROIs, ~1cm diameter) were defined in the regions with elevated FDG uptake and exhibiting uniform contrast enhancement. The clinician and radiologist developed a roadmap to guide a surgical strategy to maintain geographic aspect during operation. After total nephrectomy, the surgical specimen was oriented and positioned to match radiological images. Tissue samples at the predefined ROIs were extracted under the supervision of a board-certified radiologist. For each ROI or specimen, there were four corresponding omics datasets available, i.e., mRNA, microvascular density (MVD), clinical parameters, and molecular subtypes (ccA/ccB), which were described previously (8, 19). These data types have been shown to be very valuable and associated with molecular subtypes (19).

Radiomics feature extraction and other omics dataset

This study incorporates five omics datasets obtained from the same 23 specimens: radiomics features, clinical parameter, mRNA, MVD, and molecular subtypes. These omics datasets were defined in Electronic Electronic Supplementary Material Table S1. MVD dataset, $X^3 \in R^{23 \times 4}$, includes four features: total MVD, and MVD of small-sized, middle-sized, and large-sized vessels, respectively. The dataset of 23 specimens with 168 radiomics features, $X_{rf}^1 \in R^{23 \times 168}$, was extracted automatically by an in-house tool implemented in MATLAB 2016b from MRI (including DCE, Dixon_F, and Dixon_W) and PET images. In the mRNA dataset, only eight key selected genes, which had been demonstrated to be obviously associated with molecular subtypes (19), were included in $X^2 \in R^{23 \times 8}$, as listed in the Electronic Supplementary Material Table S1. Clinical parameters, $X^4 \in R^{23 \times 6}$, included TNM stage, tumour Fuhrman grade, and whole tumour size. Molecular subtypes, $X^5 \in R^{23 \times 1}$, represent the ccA/ccB molecular classification of specimens. The subtype classification was based on a previously published 34-gene panel of ClearCode34 (8). Each tumour specimen was assigned to either ccA or ccB according to gene expression level using

Prediction Analysis of Microarrays for R program, a centroid-based classification algorithm (8).

Classification methods

Detailed information on all of classification and statistical methods is in the Electronic Supplementary Material. The classification process involves in the two steps, as summarised in Fig. 1b. To test the accuracy of ccA/ccB classification of primary ccRCC, a fivefold cross-validation was adopted to select parameters and leave-one-out for performance evaluation.

Classification of molecular subtypes of ccRCC using radiomics—Here, sparse partial least squares (SPLS) (20–23) is used to promote variable selection and learn the low-dimensional latent structures of radiomics features X_{rf}^1 which are associated with molecular subtypes Y or $X^5 \in R^{23 \times 1}$. Then, the Fisher's linear discriminant analysis is used to obtain the discriminant model of SPLS-DA, which is constructed in the low dimensional space spanned by the latent variables and can predict molecular subtypes (ccA/ccB) of specimens.

The contribution of combined information of hybrid PET/MRI—To investigate the contribution of radiomics features from different imaging methods to classification rate, SPLS-DA was tested on four groups of imaging features: DCE images alone, Dixon images alone, PET images alone, and PET/MRI.

Classification of molecular subtypes of ccRCC using multi-omics analysis—Different data types provide different snapshots of the underlying biological system. In this part, multi-omics datasets (including mRNA, microvascular density and clinical parameters from each specimen) were combined with radiomics features to refine the model of SPLS. Here, the latent component T from the results of SPLS-DA was used as a radiomics feature to represent the specimens. The generalised PLS (GSPLS) (24, 25) was adopted to identify the latent pairwise relationship among multi-omics datasets. GSPLS maximise the sum of covariance of any two datasets. The details of the multi-omics definition are given in Electronic Supplementary Material Table S1. Then, Fisher's linear discriminant analysis was adopted to train the classification model of GSPLS-DA. In this step, the low-dimensional latent components of SPLS from the above step were as the new dataset of radiomics features, $X^1 \in R^{23 \times 4}$, which was associated with mRNA X^2 , microvascular density X^3 , clinical parameters X^4 and molecular subtypes X^5 .

The purpose of this study was to construct a model to accurately predict the molecular subtype of ccRCC with non-invasive information, using radiomics features. In the multi-omics analysis, for the aspect of discriminating analysis, two strategies were applied to train the discriminating model. The first strategy utilised an integration model, in which the first latent component of two non-invasive factors, radiomics features and clinical parameters, were combined. The second strategy for the discriminating model is a separation model, using only radiomics features as the input during the testing. Based on GSPLS, four groups of separating latent components for multi-omics datasets were obtained, but only one group from radiomics features $X^1 \in R^{23 \times 4}$ was used to discriminate the specimens.

Association of latent components of radiomics features with molecular subtypes—The multidimensional scaling (MDS) was performed to show the ability of the latent components for classification, and heat map analysis was performed to evaluate the associations between latent components of the proposed method and molecular subtypes.

Statistical analysis

The MATLAB function “mdscaling” was performed to show the ability of the latent components for classification. The following R packages were used: The spls package was used for SPLS-DA, the mixOmics was used for GSPLS, and the function lda in MASS package was used for Fisher’s linear discriminant analysis. The ggplot2 and pheatmap packages were used for heat maps. The permutation strategy (10,000 times) was used to verify the effectiveness of the classification results, and $p < 0.05$ was considered significant. The cpt package was used for permutation test.

RESULTS

Classification of molecular subtypes of ccRCC using radiomics

In Table 1, the detailed classification results are summarised in a confusion matrix. Note that a confusion matrix displays the number of correct and incorrect predictions made by the classification model, compared with the actual classifications. In the confusion matrix, the rows present the number of actual classifications in the test data, and the columns present the number of predicted classifications made by the model. As shown in Table 1, with four latent components, the overall correct classification rate of SPLS-DA is 86.96%. The association between the number of latent components and the correct classification rate is shown in Electronic Supplementary Material Fig. S1. It is shown that, at first, the correct classification rate (CCR) increases with the increase of latent components; the highest CCR of 86.96% (20/23) with permutation test $p = 7 \times 10^{-4}$ is obtained when the number of latent components K is 4. The CCR has a notable increment from $K = 1$ to 4. When $K > 4$, the CCR decreases gradually.

To make it easier to understand and analyse the characteristics of SPLS-DA, 22 specimens with radiomics features and corresponding molecular subtypes were used to train the model with the corresponding direction vectors W and discriminant coefficient Q , and then all of 23 specimens were projected to their latent components $T = X_{rf}^1 \cdot W$ which were plotted with real category Y of molecular subtypes. Fig. 2 shows that all 23 specimens were projected to four latent components $T = [rf1, rf2, rf3, rf4]$ and displayed in a three-dimensional MDS plot. The MDS plot displayed the position of each specimen in three-dimensional Euclidean space, with the distance between experimental samples reflecting their approximate degree of correlation (26). In Fig. 2, samples belonging to different molecular subtypes (ccA/ccB) are in different regions based on their overall similarity of expression pattern viewed by MDS. It can be inferred that SPLS captures the intrinsic structure information between radiomics features and molecular subtypes, and the latent space spanned by the latent components is a compact and readily separable grouping for classification. The heat map in Electronic Supplementary Material Fig. S2 listed the associations between radiomics feature expression and molecular subtypes, which were derived from the first latent component. The

columns represent specimens and are sorted by molecular subtypes, and the rows represent the radiomics features with their loadings not equal to zero in the first latent component and ordered in the absolute weights values of radiomics features. As demonstrated in Electronic Supplementary Material Fig. S2, there is no single feature among the original radiomics features, which is clearly associated with decisive significance to predict the molecular subtypes.

The contribution of combined information of hybrid PET/MRI

Table 2 present a summary of the classification results of SPLS-DA using a different group of radiomics features. Only the group of features extracted from combined PET and MRI features achieved significant promising CCR to identify the molecular subtype. There were no significant correlations or discriminant ability using the features from DCE, Dixon, or PET separately, which was observed in the correlation analysis between microvascular density and radiomics features(18).

Classification of molecular subtypes of ccRCC using multi-omics analysis

The clustered heat map in Fig. 3a shows the feature expression of multi-omics for each sample. The rows represent specimens, with the specimen number displayed on the right. The columns represent features in multi-omics datasets. A hierarchical clustering algorithm was employed (27, 28) successively on the rows and columns to get the sample clusters. This heat map shows the heterogeneity of feature expression about molecular subtypes of ccA and ccB samples in multi-omics datasets. It can be found again that there is no single feature in multi-omics datasets, which can be utilised as an effective measurement to classify the category of molecular subtypes. Based on these findings from Fig. 3a, the multi-omic integrated analysis of features was utilised to identify the latent component signatures capable of discriminating between molecular subtypes. Fig. 3b is a heat map showing the latent components of multi-omics datasets from GSPLS-DA. The rows represent the specimens with four latent components in each omics dataset except the supervised information of molecular subtypes, and they are sorted in the category of molecular subtypes. The specimen number is displayed on the right. The columns represent the latent components of multi-omics datasets. It can be found that the first latent component RF-LC1 of radiomics features X^1 is the most effective measurement to discriminate the category of molecular subtypes. It is apparent that the latent components of radiomics features have more discriminating power than other features in these datasets. It can be inferred that, when the multi-omics datasets are introduced, some additional information are provided to obtain more precise instinct latent structures. So, the separating model of radiomics features X^1 probably accounts for the better results.

Regarding GSPLS-DA, two strategies were applied for the discriminating model. The first strategy utilised an integration model, in which clinical parameters and the first latent component of radiomics features were combined. According to this strategy, the classification results were summarised in the part GSPLS-DA (integration model) of Table 1. It can be found that the total CCR of integration model with 91.3% is better than that of the SPLS-DA.

Fig. 4a shows the matrix scatterplot, in which all specimens were projected to a two-dimensional space spanned by the multi-omics latent components of two datasets. In the matrix plot, the dataset names are marked in the diagonal in order, and the correlation coefficients of the corresponding multi-omics latent components from two datasets are in the lower triangle matrix. The colours and ellipses related to the molecular subtypes of ccA and ccB, which were regarded as the identification of the discriminative power of each component to separate different molecular subtypes. Fig. 4a shows the associations between the first multi-omics latent components of the two omics datasets.

The second strategy for the discriminating model is a separation model, using only radiomics features as the input during the testing. The association between the number of latent components and the correct classification rates are summarised in the GSPLS-DA (separation model) part of Table 1. All CCRs with a different number of latent components of X^1 are equal to or greater than that obtained by SPLS-DA. Obviously, other datasets in multi-omics provide additional information for radiomics features. Compared to the result of SPLS-DA, the separating model of radiomics features X^1 with just one latent component from multi-omics datasets achieves better results than the best result of SPLS-DA. The highest accuracy can be achieved when the first three latent components of X^1 are adopted with the CCR of 95.65% (permutation test, $p < 10^{-4}$). All ccB samples, with which patients have worse prognoses, were correctly identified and only one ccA sample was misclassified.

In Fig. 4b, all specimens are displayed in the latent variable space, which are spanned by the first two multi-omics latent components of each dataset. In the same way with SPLS, 22 specimens were used to train the multi-omics model, and then all of 23 specimens were projected to their latent components, respectively. From the multi-omics latent component space, it was observed that (1) the specimens of ccA and ccB were located explicitly in the separated part of radiomics feature spaces; (2) the specimens cannot be divided clearly to two classes of ccA and ccB by a line using the mRNA dataset and clinical dataset information; (3) the specimens of ccA and ccB are in different parts about microvascular density, but cannot be separated by a simple line. Only using the first latent component of radiomics features X^1 , the category of most specimens can be identified, except for the 19th sample, which is probably classified incorrectly because it seems to get a closer distance to the samples of ccA relevant to the first latent component of radiomics features.

DISCUSSION

The major advantages of this study were three-fold. First, the region-wise approach links regional radiomics information and local ccA/ccB classification using histological tissue specimens, which were all carefully extracted from pre-defined ROIs. Second, the SPLS-DA method was applied, which has been used to deal with high-dimensional data in other fields for coping with high-dimensional radiomics dataset, and then used the post-processed low-dimensional latent components to train the classification model with high accuracy of classification. Third, the multi-omics analysis of GSPLS-DA, with either integration model or separation model, is demonstrated to be able to improve the prediction accuracy of SPLS-DA.

In previous reports of classification models based on radiomics features, the predictive performance usually ranged between 0.5 and 0.68 (14). To improve the predictive performance of classification, several processing runs were performed using SPLS. First, it is important that high-throughput radiomics features were extracted for each ROI because there were no prior knowledge about the relationships between radiomics features and molecular subtypes (ccA/ccB) of ccRCC. Second, by the variable selection and projecting these selected features to low-dimensional latent components associated with molecular subtypes, the implicit relations between radiomics features and molecular subtypes can be learned. Third, in multi-omic analysis, each feature in these datasets must be demonstrated to be associated with the objective of molecular subtypes. Otherwise, GPLS-DA does not work well because there is a great deal of noise information. In this study, the datasets except the radiomics had been demonstrated to very valuable and associated with molecular subtypes (ccA/ccB) described previously (19). The worst classification result was obtained when the original radiomics features $X_{rf}^1 \in R^{23 \times 168}$ were used in multi-omic analysis, which were not elaborated in this paper. So, in multi-omic analysis, using the low-dimensional latent components of SPLS as the new dataset of radiomics features, $X^1 \in R^{23 \times 4}$, the best CCR of 95.65% was obtained (permutation test, $p < 10^{-4}$). Although there was some overlap between the activities of these two types, previous evaluation of this dataset has shown that the ccB subtyped samples of ccRCC display significantly higher maximal standardised uptake value than the ccA in FDG PET, and suggested that FDG PET was superior in distinguishing ccA and ccB than conventional MRI parameters(29); however, in the present study, the PET information alone does not have the independent resolving capacity needed to supplant the molecular tool. Images were decoded into 168 quantitative features and it was observed that the Dixon sequence, which can assess the fat and water components of tumours, had higher classification accuracy (21.74%) than PET (0%). In addition, the integration of PET/MRI radiomics information could achieve much higher classification accuracy (86.95%) than using the FDG PET information alone (Table 2). This suggests that PET/MRI, which combines both structural and metabolic images, can be of value in future applications of extracting radiogenomic biomarkers.

In the multi-omics analysis, other datasets provide additional information for radiomics features and improve the predictive performance of the model. Compared to CCR of SPLS-DA, the separating model of radiomics features X^1 with just one latent component from multi-omics datasets achieves better result than the best result of SPLS-DA. The highest accuracy was achieved when the first three latent components of X^1 are adopted with the CCR of 95.65% (permutation test, $p < 10^{-4}$). All ccB samples, with conferring the worse prognosis, were correctly identified and only one ccA sample was misclassified. This suggests that this highly sensitive classification model could be used to localise ccB tumour parts for guiding biopsies, or to identify patients who require systemic or aggressive treatment in clinical practice. Integrated radiomics strategies are particularly intriguing as systemic adjuvant therapies have recently been approved for use, but currently lack biomarkers to guide their implementation (30–32); however, further investigation with an independent population is needed to validate the consistency of the classification model.

Referring to Fig. 4a for the integration model of GSPLS-DA with a combination of radiomics features X^1 and clinical parameters X^4 , it is noticeable that radiomics features were the most crucial dataset to distinguish the two types of samples accurately. No matter which dataset is combined with the radiomics features, effective parameters can be obtained to identify two types of samples according to the first multi-omics latent component of each omics dataset; however, as two kinds of ellipses are overlapped more or less, there are a certain risk of false identification, but as shown in Electronic Supplementary Material Fig. S3, valuable information for identifying the types of samples based on the second multi-omics latent component of each omics dataset was not obtained.

In Fig. 4a, two or three samples overlapped for ccA and ccB, for which there is high probability leading to classification error. The same situation was observed in Fig. 4b for the separation modal of GSPLS-DA; however, even in the case of generating the classification based on transcription features, which is a reference standard (7), there is roughly 10% classification ambiguity. Thus, this classification level (91.3–95.65%) of the proposed method represents a discriminating level that is consistent with reality. In summary, significant efforts have been taken to develop prognostic models for stratifying patients in risk groups, including pathological features, such as nuclear grade, TNM stage, molecular analyses, and nomograms. The ccA/ccB signature is the first independent prognostic biomarker to be validated; however, molecular subtyping is seriously hindered by regional heterogeneity and the potential for dynamic change over the course of time. Radiomics imaging strategies provide a unique opportunity to simultaneously evaluate disease throughout the body, as well as to repeat the analysis at a future time point. Previous evaluation of this dataset has shown that the ccB subtyped samples of ccRCC display significantly higher maximal standardised uptake value than the ccA in FDG PET(29); however, the PET information alone does not have independent resolving capacity needed to supplant the molecular tool. Although the present conclusions are limited by the small sample size, and further investigation is needed to validate the findings using a larger cohort, the present data demonstrate that no single feature is sufficient, but a four-feature latent component radiomics signature, combining both PET and MRI, allowed the non-invasive ccA/ccB classification feasible with high accuracy. Integrative multi-omics analysis to explore functions of candidate phenotype, regional expression of cancer genes, and proteins, and examinations of heterogeneous radiogenomic signal is a potentially useful model to identify holistic patient-centred approaches to monitoring and defining disease states.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported in part by NIH grants (EB006733, EB008374, EB009634, MH100217, MH108914, AG041721, AG049371, AG042599, DE022676), and by grants from the National Cancer Institute (NCI; 3P30CA016086) and the UNC University Cancer Research Fund (UCRF); National Research Service Award T32ES007126 from the National Institute of Environmental Health Sciences to S.B.; the Howard Hughes Medical Cancer Institute of the NIH Award 5T32CA160001 and an un-restricted grant from Siemens Medical Solutions; the Fundamental Research Funds for the Central University of China (no. 3132014091).

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2016. *CA Cancer J Clin*. 2016; 66(1):7–30. [PubMed: 26742998]
2. Znaor A, van den Hurk C, Primic-Zakelj M, et al. Cancer incidence and mortality patterns in South Eastern Europe in the last decade: gaps persist compared with the rest of Europe. *Eur J Cancer*. 2013; 49(7):1683–1691. [PubMed: 23265703]
3. Gerlinger M, Rowan AJ, Horswell S, et al. Intratumour heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med*. 2012; 366(10):883–892. [PubMed: 22397650]
4. Janzen NK, Kim HL, Figlin RA, Belldgrun AS. Surveillance after radical or partial nephrectomy for localized renal cell carcinoma and management of recurrent disease. *Urol Clin North Am*. 2003; 30(4):843–852. [PubMed: 14680319]
5. de Velasco G, Culhane AC, Fay AP, et al. Molecular subtypes improve prognostic value of International Metastatic Renal Cell Carcinoma Database Consortium prognostic model. *Oncologist*. 2017; 22(3):286–292. [PubMed: 28220024]
6. Gulati S, Martinez P, Joshi T, et al. Systematic evaluation of the prognostic impact and intratumour heterogeneity of clear cell renal cell carcinoma biomarkers. *Eur Urol*. 2014; 66(5):936–948. [PubMed: 25047176]
7. Brannon AR, Reddy A, Seiler M, et al. Molecular stratification of clear cell renal cell carcinoma by consensus clustering reveals distinct subtypes and survival patterns. *Genes Cancer*. 2010; 1(2):152–163. [PubMed: 20871783]
8. Brooks SA, Brannon AR, Parker JS, et al. ClearCode34: a prognostic risk predictor for localized clear cell renal cell carcinoma. *Eur Urol*. 2014; 66(1):77–84. [PubMed: 24613583]
9. Haake SM, Brooks SA, Welsh E, et al. Patients with ClearCode34-identified molecular subtypes of clear cell renal cell carcinoma represent unique populations with distinct comorbidities. *Urol Oncol*. 2016; 34(3):122.e121–127.
10. Gerlinger M, Catto JW, Orntoft TF, Real FX, Zwarthoff EC, Swanton C. Intratumour heterogeneity in urologic cancers: from molecular evidence to clinical implications. *Eur Urol*. 2015; 67(4):729–737. [PubMed: 24836153]
11. Gerlinger M, Horswell S, Larkin J, et al. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat Genet*. 2014; 46(3):225–233. [PubMed: 24487277]
12. Karlo CA, Di Paolo PL, Chaim J, et al. Radiogenomics of clear cell renal cell carcinoma: associations between CT imaging features and mutations. *Radiology*. 2014; 270(2):464–471. [PubMed: 24029645]
13. Shinagare AB, Vikram R, Jaffe C, et al. Radiogenomics of clear cell renal cell carcinoma: preliminary findings of The Cancer Genome Atlas–Renal Cell Carcinoma (TCGA–RCC) Imaging Research Group. *Abdom Imaging*. 2015; 40(6):1684–1692. [PubMed: 25753955]
14. Parmar C, Grossmann P, Bussink J, Lambin P, Aerts HJ. Machine learning methods for quantitative radiomic biomarkers. *Sci Rep*. 2015; 5:13087. [PubMed: 26278466]
15. Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer*. 2012; 48(4):441–446. [PubMed: 22257792]
16. Aerts HJWL, Velazquez ER, Leijenaar RTH, et al. Decoding tumour phenotype by non-invasive imaging using a quantitative radiomics approach. *Nat Commun*. 2014; 5:4006. [PubMed: 24892406]
17. Kristensen VN, Lingjaerde OC, Russnes HG, Vollen HK, Frigessi A, Borresen-Dale AL. Principles and methods of integrative genomic analyses in cancer. *Nat Rev Cancer*. 2014; 14(5):299–313. [PubMed: 24759209]
18. Yin Q, Hung S-C, Wang L, et al. Associations between tumour vascularity, vascular endothelial growth factor expression and PET/MRI radiomic signatures in primary clear-cell–renal-cell–carcinoma: proof-of-concept study. *Sci Rep-Uk*. 2017; 7:43356.

19. Brooks SA, Khandani AH, Fielding JR, et al. Alternate metabolic programs define regional variation of relevant biological features in renal cell carcinoma progression. *Clin Cancer Res*. 2016; 22(12):2950–2959. [PubMed: 26787754]
20. Lee D, Lee Y, Pawitan Y, Lee W. Sparse partial least-squares regression for high-throughput survival data analysis. *Stat Med*. 2013; 32(30):5340–5352. [PubMed: 24105836]
21. Chun H, Keles S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J Roy Stat Soc B*. 2010; 72:3–25.
22. Hoffmann I, Filzmoser P, Serneels S, Varmuza K. Sparse and robust PLS for binary classification. *J Chemometr*. 2016; 30(4):153–162.
23. Huang CC, Tu SH, Huang CS, Lien HH, Lai LC, Chuang EY. Multiclass prediction with partial least square regression for gene expression data: applications in breast cancer intrinsic taxonomy. *Biomed Res Int*. 2013; 2013(1 article 6):248648. [PubMed: 24490149]
24. Tenenhaus A, Tenenhaus M. Regularized generalized canonical correlation analysis for multiblock or multigroup data analysis. *Eur J Oper Res*. 2014; 238(2):391–403.
25. Günther OP, Shin H, Ng RT, et al. Novel multivariate methods for integration of genomics and proteomics data: applications in a kidney transplant rejection study. *Omics*. 2014; 18(11):682. [PubMed: 25387159]
26. Everitt BS, Dunn G. *Applied Multivariate Data Analysis*. 2. 2013.
27. Massaro JM. *Clustering, Complete Linkage*. John Wiley & Sons, Ltd; 2005.
28. Hyatt D, Locascio PF, Hauser LJ, Uberbacher EC. Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics*. 2012; 28(17):2223–2230. [PubMed: 22796954]
29. Manara R, Carlier RY, Righetto S, et al. Basilar artery changes in Fabry disease. *AJNR Am J Neuroradiol*. 2017; 38(3):531–536. [PubMed: 28126747]
30. Ravaud A, Motzer RJ, Pandha HS, et al. Adjuvant sunitinib in high-risk renal-cell carcinoma after nephrectomy. *N Engl J Med*. 2017; 375(23):2246–2254.
31. Buti S, Puligandla M, Bersanelli M, et al. Validation of a new prognostic model to easily predict outcome in renal cell carcinoma: the GRANT score applied to the ASSURE trial population. *Ann Oncol*. 2017; 28(11):2747–2753. [PubMed: 28945839]
32. Bandini M, Smith A, Marchioni M, et al. Adjuvant therapies in nonmetastatic renal-cell carcinoma: a review of the literature. *Clin Genitourin Canc*. 2018

Highlights

1. Radiomics is noninvasive and used to predict the molecular subtypes of clear cell renal cell carcinoma (ccRCC).
2. Sparse partial least squares discriminant analysis (SPLS-DA) can learn on small sample size and accurately predict the molecular subtype
3. Multilevel -omics datasets are integrated to elevate the discriminative ability.
4. PET/MR supply the imaging differentiator of ccA and ccB subtypes.

Figure 1. The flowchart of the overall design of the study. (a) The three phases of the process. (b) The proposed method to obtain the classification model.

Figure 2. MDS three-dimensional plot of all 23 specimens. The MDS plot displays the position of each specimen in three-dimensional Euclidean space, with the distance between experimental samples reflecting their approximate degree of correlation.

Figure 3. Heat maps of associations between the multi-omics signature in relation with the samples. (a) Clustered heat map representing the multi-omics datasets in relation with the samples. The rows represent the specimens, with the specimen number displayed on the right. The columns represent features of all datasets. (b) Heat map showing the latent components of multi-omics datasets from GSPLS-DA. The rows represent the specimens with four latent

components in each omics dataset except the supervised information of molecular subtypes, and they are sorted in the category of molecular subtypes. The specimen number is displayed on the right. The columns represent the latent components of multi-omics datasets.

Figure 4.

The specimen scatter plots depicting the clustering of molecular subtypes in multi-omics latent component space. (a) In the matrix plot, the correlation and cluster between the first multi-omics latent components of two datasets, in which the dataset names are marked in the diagonal in order, and the correlation coefficients of the corresponding multi-omics latent component from two datasets are in the lower triangle matrix. It was observed that, no matter which dataset was combined with radiomics features, (or distinguish) the two types of samples could be identified accurately according to the first multi-omics latent component of

each dataset. (b) The specimen scatter plot in the space spanned by the first two multi-omics latent components of each dataset.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1

Confusion matrix of classification results

Method	No. of latent components	Genome-based classification	Radiomics-based prediction results		CCR (%)	Total CCR (%)
			ccA	ccB		
SPLS-DA	4	ccA	7	2	77.78	86.96
		ccB	1	13	92.86	
GSPLS-DA (integration model)	1	ccA	8	1	88.89	91.30
		ccB	1	13	92.95	
		ccA	8	1	88.89	
		ccB	1	13	92.95	
GSPLS-DA (separating model)	2	ccA	7	2	77.78	86.96
		ccB	1	13	92.86	
	3	ccA	8	1	88.89	95.65
		ccB	0	14	100	
	4	ccA	7	2	77.78	86.96
		ccB	1	13	92.86	

SPLS-DA, sparse partial least squares discriminant analysis; GSPLS-DA generalised partial least squares discriminant analysis.

Table 2

Correct classification rate (CCR) using different techniques with four latent components.

Dataset	Hybrid PET/MRI	DCE alone	Dixon alone	PET alone
CCR	20/23	0/23	5/23	0/23
(%)	(86.95)	(0)	(21.74)	(0)

PET/MRI, positron-emission tomography/magnetic resonance imaging; DCE, dynamic contrast-enhanced.