Combining texture features of whole slide images improves prognostic prediction of recurrence-free survival for cutaneous melanoma patients

Yanbin Peng
Peking University Shenzhen Hospital
Yunfeng Chu
Peking University Shenzhen Hospital
Zhong Chen
Peking University Shenzhen Hospital
Wen Zhou
Peking University Shenzhen Hospital
Shengxiang Wan
Peking University Shenzhen Hospital
Yingfeng Xiao
Peking University Shenzhen Hospital
Youlong Zhang
HuaJia Biomedical Intelligence
Jialu Li (jialu.li@huajiabio.com)
HuaJia Biomedical Intelligence https://orcid.org/0000-0002-6411-6876

Technical innovations

Keywords: cutaneous melanoma, recurrence-free survival, whole slide image, computer-aided image processing.

Posted Date: June 4th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-24723/v2

Version of Record: A version of this preprint was published on June 16th, 2020. See the published version at https://doi.org/10.1186/s12957-020-01909-5.

Abstract

Background: Accurate prediction of recurrence-free survival (RFS) is important for the prognosis of cutaneous melanoma patients. The image-based pathological examination remains as the gold standard for diagnosis. It is of clinical interest to account for computer-aided processing of pathology image when performing prognostic analysis.

Methods: We enrolled in this study a total of 152 patients from TCGA-SKCM (The Cancer Genome Atlas Skin Cutaneous Melanoma project) with complete information in recurrence-related survival time, baseline variables (clinicopathologic variables, mutation status of BRAF and NRAS genes), gene expression data and whole slide image (WSI) features. We preprocessed WSI to segment global or nucleus areas, and extracted 3 types of texture features from each region. We performed cross validation and used multiple evaluation metrics including C-index and time-dependent AUC to determine the best model of predicting recurrence events. We further performed differential gene expression analysis between the higher and lower-risk groups within AJCC pathologic tumor stage III patients to explore the underlying molecular mechanisms driving risk stratification.

Results: The model combining baseline variables and WSI features had the best performance among models with any other types of data integration. The prognostic risk score generated by this model could provide a higher-resolution risk stratification within pathologically-defined subgroups. We found the selected image features captured important immune-related variations, such as the aberration of expression in T cell activation and proliferation gene sets, and therefore contributed to the improved prediction.

Conclusions: Our study provided a prognostic model based on the combination of baseline variables and computer-processed WSI features. This model provided more accurate prediction than models based on other types of data combination in recurrence-free survival analysis.

Trail registration: This study was based on public open data from TCGA and hence the study objects were retrospectively registered.

Background

Melanoma is a type of skin cancer with a high mortality rate. In 2018, 287,732 new cases and 60,712 deaths of melanoma were registered worldwide(1). Cutaneous melanoma, which accounts for over 90% of melanoma cases, remains one of the most aggressive forms of skin cancer and shows an increasing incidence and mortality rate globally(2, 3). Improving prognosis of cutaneous melanoma patients has important implications for a better management of the disease. Routine prognosis method uses clinicopathologic features including Breslow tumor thickness, ulceration, mitotic index, Clark level and AJCC (The American Joint Committee on Cancer) pathologic tumor stage(4, 5). Whether such method can be improved with the addition of WSI or high-throughput sequencing data is under active investigation.

Many previous studies had attempted to develop prognostic models using different types of variables including clinicopathologic, mutation, mRNA, microRNA, and methylation variables. Zhao *et al.*(6) identified a 25-gene signature that can effectively estimate the level of immune cell infiltration in melanoma, providing a robust biomarker significantly related to survival outcome (disease-specific survival, post-recurrence survival, or overall survival). Multidimensional omics data were also utilized to provide more accurate prediction. Jayawardana *et al.*(7) developed models to classify 1-year and 4-year survival status based on clinicopathologic, mutation, mRNA, microRNA, protein information and their different combinations. They identified that models based on the combination of clinicopathologic variables and mRNA expression profile performed the best under a cross-validation framework. Jiang *et al.*(8) used sparse PCA and partial least squares methods to take whole multidimensional omics profiles into consideration. Their methods showed a significant increase of C-index values of overall survival prediction. However, these studies had not extended to including WSI features, while studies that did use image data were not focused on the prediction of survival outcomes. For example, Lu *et al.*(9) proposed a diagnostic model based on epidermis segmentation, keratinocytes segmentation, melanocytes detection and feature construction on whole slide images. This technique achieved a classification accuracy of 90% for skin tissue malignancy. Failmezger *et al.*(10) analyzed the spatial association between different types of nodes in WSI. They identified that two stromal features (stroma clustering and stromal barrier) had significant coefficients in a Cox model. To our knowledge, few studies proposed pathology image-based prediction models for recurrence-free survival analysis.

As the treatment of cutaneous melanoma has improved over years, especially since the advent of targeted immunotherapy(11, 12), the RFS measurement has gained increasing importance for post-surgery management of melanoma patients. An accurate RFS analysis can decrease the cancer-related death rate by not only personalizing treatment options, but also prompting active cancer surveillance at an early stage for specific risk groups. On the other hand, as recurrence events have already been used as an effective endpoint for cancer clinical trials, a higher accuracy of RFS measurements allows precise enrollment of patients that are more likely responsive to new therapies.

Our work hence presented a RFS model based on baseline variables and texture features extracted from WSI. This model provided a higher accuracy in predicting recurrence-free survival than baseline variables-based model or models with other types of data integration. The extracted WSI features contributed to the improved prediction by capturing variations in immune-related gene expressions.

Methods

Patients and samples

This study was performed using 152 patients from the TCGA-SKCM(13) (The Cancer Genome Atlas Skin Cutaneous Melanoma) project. We enrolled patients that have complete information in AJCC stage, the dominant clinicopathologic variable used for RFS analysis (Table S4). The enrolled patients should also have non-missing values in tumor location (metastatic or locoregional), WSI images, high-throughput

gene expression data and tumor recurrence-related follow-up. The study work flowchart was shown in Figure 1. For each patient, the recurrence status was defined as 1 for those who had experienced recurrence and as 0 for censoring. The censoring time was set as the death time if one had the record of death or as last follow-up time if otherwise.

We collected 4 types of data for RFS analysis: clinicopathologic variables, high-throughput gene mutational profile, high-throughput gene expressional profile and WSI features. For clinicopathologic variables, 4 covariates with complete records were included: age at diagnosis, gender, primary location and AJCC pathologic tumor stage. We selected the mutation status of NRAS and BRAF genes to represent the gene mutation data as previously described(7). These clinicopathologic variables and gene mutations were used as the baseline variables for following model development. The distributions of these variables were summarized in Table 1. For gene expression, we used the FPKM (fragments per kilobase of exon model per million reads mapped) value to represent gene expression levels. A total of 5277 genes were selected for further analysis.

Table 1: Summary of distributions of baseline variables.

Baseline variables	Summary
Age at diagnosis	mean=60.06, std=14.05
Gender	92 males
Primary location	60 metastatic, 92 locoregional
AJCC pathologic tumor stage	67 stage III&IV, 85 stage <III
BRAF	82 mutated
NRAS	32 mutated

Whole slide image processing and feature extraction

All the melanoma tissue slides were stained by hematoxylin and eosin (H&E) and scanned by Aperio Digital Pathology Slide Scanner. Ten slides were magnified 20 times (20X) and the other 142 slides were magnified 40 times (40X). We extracted 3 types of texture features in 2 regions of interest (ROI): global and nucleus ROI, respectively. This process consisted of two steps: region segmentation and feature extraction. An illustration of these procedures was shown in Figure S11. More details on the region segmentation can be found in the Supplementary methods.

For feature extraction, we extracted 24 GLCM (gray level cooccurrence matrix) features, 16 GLRLM (gray level run length matrix) features and 16 GLSZM (gray level size zone matrix) features(14) from each ROI. For nucleus ROIs from the same WSI, we calculated their mean, standard deviation, range and disorder as the summary statistics(15). This resulted in a total of 224 nucleus features and 56 global features. Image processing and feature extraction were performed by Python 3.7 and packages including “Pyradiomics”(14).

Modeling and evaluation

We performed QR decomposition-based method(16) to reduce linear dependencies among the gene expressional profile before model fitting. In addition to the three types of features mentioned above, we

also combined different types of features as new feature sets for model development. The number of features in each set was summarized in Table S1.

We performed 3-fold cross validation for models developed based on each feature set. In each fold, a lasso Cox model was trained on the training set and was tested on the validation set. Due to the limited sample size, we used C-index computed from validation data to evaluate model performance. We also computed the time-dependent AUCs(17) from day 31 (5% percentile) to day 4,631 (95% percentile) to evaluate the performance at each time point. The model development and comparison were performed by R version 4.1.

Differential gene expression analysis

The differential gene expression analysis was performed using edgeR(18) package, and the gene ontology enrichment analysis was performed by GOseq(19) or clusterProfiler(20). More details on the data processing and parameter setting can be found in the Supplementary methods.

Results

Workflow and patient characteristic

The study workflow was shown in Figure 1. We performed RFS modeling based on three types of variables: baseline variables (including clinicopathologic variables, and mutation status of NRAS and BRAF genes), gene expressional profile and WSI features. We developed 7 sets of features based on these three types of data (Table S1). The model performance was evaluated by C-index and time-dependent AUC on 3-fold cross validation. We further presented the potential application of the best model under clinical setting. We also performed differential gene expression analysis between the higher and lower-risk subgroups stratified by our model for AJCC stage III patients.

We enrolled a total of 152 patients with complete information in recurrence-related survival time, baseline variables, gene expression data and WSI features. A total of 82 patients had experienced recurrence, while 65 had last follow-up time and 5 died without recurrence. We performed Kaplan-Meier estimation for all patients and the RFS probability curve was shown in Figure S2. The median survival time was 1,757 days.

Model comparisons

We compared the performance of models developed based on each single type of data to assess the prognostic power of single-type feature sets. The C-index of models based on baseline variables (mean/std = 0.654/0.014) was the highest (Table S2, Figure S3). The models based on gene expression or WSI features had a slight difference (mean/std of C-index: expr=0.639/0.039, im=0.635/0.033). To evaluate the prediction accuracy at each time point, we computed the time-dependent AUC on the validation results. As shown in Figure 2A, the models based on baseline variables showed obvious superiority until about day 2,500. In contrast, the model based on WSI features had increasing prediction

accuracy since about day 1,500. This motivated the combinatorial modeling, as combing baseline and WSI feature in survival prediction might utilize the prediction advantage of single-type data-based model within specific time intervals.

We then developed models based on combinations of different types of data, and compared their prediction performance. For WSI image analysis, we extracted texture features from global regions or segmented nucleus regions (Figure S1). As summarized in Tables S2, the best C-index (mean/std = 0.772/0.029) was achieved by the model combining baseline variables and WSI features (Figure S3). As shown in Figure 2B, such model also had the best performance at almost every time point as measured by time-dependent AUC (mean/std of time-dependent AUC = 0.785/0.038), indicating a clear benefit of data integration. We therefore used this feature set and the optimal penalty (Figure S3) to develop a lasso Cox model on all patients. This model was used as the final prognostic model and the coefficients of selected features were showed in Table S3.

The image-based prognostic model

The proposed model included 20 WSI features and 5 baseline variables. For the WSI features, 14 of them were extracted from nucleus ROIs and 6 were from global ROIs. The computational formula of each feature was shown in Supplementary Methods. A positive value of coefficient represented that the hazard of recurrence would increase with the feature values. As shown in Table S3, the 3 largest absolute values of the coefficients were from GLRLM (*RunEntropy_std* and *ShortRunEmphasis_range*) and GLCM (*Idn_range*) in nucleus ROIs. Both the *RunEntropy* and *ShortRunEmphasis* were the measurement of the distribution of run lengths. The *Idn* (Inverse Difference Normalized) quantified the local homogeneity within nucleus ROI, which could be low value if there was necrosis or dissolution of nucleus. Since the standard deviation or range of these features was significant predictor in the model, we inferred that the variance of nuclei shape, surrounding textures and homogeneity contributed to an effective prognostic analysis. For the features extracted from global ROIs, the *glcm_Idmn* and *glszm_LargeAreaHighGrayLevelEmphasis* had the largest absolute coefficients. The *Idmn* (Inverse Difference Moment Normalized) also measured the local homogeneity, while the *LargeAreaHighGrayLevelEmphasis* computed the emphasis of regions with large area and high gray level. Both of the features were indicators of the tumor region size.

Risk stratification

To illustrate potential applications of the prognostic model, we computed a risk score by summing over the product of model features and their coefficients. To compare the risk score with traditional prognostic variables, we also computed risk scores for models based on AJCC pathologic tumor stage or baseline variables. We used each risk score to fit a univariate Cox PH model on 5 subgroups of patients (all, AJCC stage<III, AJCC stage≥III, metastatic or locoregional) and applied likelihood ratio test to assess the significance of the score. As summarized in Table S4, both the likelihood ratio test statistic and its p value showed that adding WSI features to baseline variables greatly improved the prediction accuracy.

To illustrate the independent prognostic value of image-based risk score, we set the median of risk score as the threshold and stratified all the patients into a higher or lower-risk group. We then performed Kaplan-Meier estimation for each group. As shown in Figure 3A, the survival distributions of the two risk groups characterized among all patients were significantly different (Log-rank p value < 0.0001). The median survival time of the two risk groups were 678 days and 3,716 days, respectively. We also estimated the survival probability within 4 specific pathologically-defined subgroups of patients (Figure 3). The survival distributions were all significantly different for each subgroup. The median survival time of each subgroup was shown in Table S5. Of note, the lower-risk group in patients with severe stage (AJCC stage≥III) had a median RFS time of 5,354 days, which is 1.54 times longer than that of with mild or moderate stage (AJCC stage<III), who had a median RFS time of 3,488 days. As shown in Figure S4, this risk score could significantly stratify the higher and lower-risk group for overall survival as well.

Differential risk-related enrichment of gene sets

To explore the molecular mechanisms underlying the superiority of image-based prognostic model, we performed differential gene expression analysis between the higher (43 patients) and the lower-risk (15 patients) group characterized by our model for patients within AJCC stage III. To reduce the biological variation within each risk group, we calculated the spearman correlation coefficients between any pairs of samples and filtered out those with a correlation coefficient smaller than 0.85 with more than 20 other samples among the higher-risk group, or with more than 7 other samples among the lower-risk group. This resulted in 9 lower-risk patients and 21 higher-risk patients for further analyses.

A total of 188 down-regulated and 28 up-regulated genes were identified as significantly differentially expressed. We found 226 enriched BP (biological process), 18 enriched CC (cellular component) and 12 MF (molecular function) GO terms. As shown in Figure S5, the most majority of top 20 enriched BP terms was involved in immune response, immune cell activation and proliferation. In particular, the T cell activation and proliferation-related pathways were significantly enriched. These enriched BP terms were also identified by the clusterProfiler package as shown in Figure S8, which suggested that the variation in the regulation of T cell activation was the potential key driver for differential risk. For CC terms, the T cell receptor-related GO terms were significantly enriched as shown in Figure S6 & S9. For MF terms, MHC protein binding and cytokine activity-related GO were associated with the risk stratification (As shown in Figure S7 & S10). These enriched GO terms suggested that the selected image features accounted for the variations of T cell activities and hence provided a more accurate assessment of disease progression.

Discussion

In this study, we performed recurrence-free survival analysis, and developed lasso Cox prediction models based on different types of data (baseline variables, gene expression and whole slide image features). Accurate RFS measurements have become increasingly important as the treatment of cutaneous melanoma has significantly improved over years. A reliable RFS prediction could thus assist in more precision treatment selection for melanoma patients. Our evaluation criteria included C-index and time-

dependent AUC. We identified that combining baseline variables and WSI features achieved the best prediction performance (cross validation C-index: 0.772/0.029, time-dependent AUC: 0.785/0.038). We showed that models trained on single-type data could have varied prediction accuracy within specific time intervals. We were then motivated to combine different types of data to develop a model that achieved uniformly best prediction accuracy at most majority of time after initial diagnosis. We also showed that this combinatorial model provided significant risk stratifications within specific subgroups defined by metastatic status or AJCC pathological tumor stages. We found from gene expressional profiles that T cell activities were significantly associated with the differential risk determined by the image-based prognostic model.

T-cell activation is closely related to the adoptive or targeted immunotherapies. Such immunotherapies either provide co-stimulatory signals to trigger T-cell proliferation, or block inhibitory molecules to unleash the anti-tumor T-cell activities(21). Our findings indicate that, even within specific risk groups (e.g.: AJCC stage III) as defined by routine clinicopathological variables, it is still possible to further identify subgroups characterized with the variation in T-cell activation using WSI features. Therefore, whether such subgroups have differential responses to immunotherapy, such as PD-1 or CTLA4-based immune checkpoint therapies, warrants further investigation.

Our study provided a cost-effective way to evaluate the prognosis by only taking baseline variables and automatically generated WSI features as the input. We showed that many top enriched GO terms were related to immune response, T cell activation and proliferation. Genes related to T cell development and function were also identified by Jiang *et al.*(8) as significantly associated with overall survival by analyses performed based on clinicopathologic variables, methylation, CNA (copy number alteration), gene expression and mutational data. Pastorfide *et al.*(21) and others(22, 23) also proved that lymphocytic infiltrates were associated with metastasis and survival outcome by artificial analyses of histologic images. These findings were consistent with our study, but we used computer-aided image processing instead to automatically quantify such information from routinely-made WSI.

Our study has limitations. First, due to a large number of missing values, we only enrolled 152 patients with complete information in our study. The best model was determined by 3-fold cross validation without being tested on an independent dataset. Second, the quality of pathology images varied. The variation in staining intensity and marker-pen pollution were the two most frequent problems. The former problem could affect nucleus segmentation, and we mitigated this by segmenting ROIs centering around the localized nucleus region. We also manually checked each ROI to remove polluted ones. Third, we did not evaluate the effects of interaction between different types of data when performing data integration. Fourth, we did not include some other routinely used prognostic variables, such as Breslow tumor thickness, ulceration, mitotic rate and Clark levels, due to a large proportion of missing values (Table S8). These variables might improve the prognostic prediction performance of our models.

For the future extension of our study, we will perform external validation of the prediction model using multi-center retrospective cohort data. Another key point that needs to be addressed in future studies is

the assessment of effect of clinicopathological variables and treatment on RFS prediction using more complete data. Notwithstanding these limitations, we believe our method has potential for clinical translation to reduce the level of heterogeneity in RFS for cutaneous melanoma patients. This could influence the treatment selection for patients, as well as the patient enrollment for related clinical trials. Moreover, considering the capacity of WSI texture features in capturing immune cell activity in this study, we believe that WSI texture features could also have prognostic values in other types of cancer, which warrants exploration in future studies.

Conclusions

In summary, we developed an image-based combinatorial model and demonstrated its prediction ability for recurrence-free survival. The model includes 20 automatically generated WSI features, 3 clinicopathologic variables and mutation status of 2 genes. We hence provided a cost-effective prognostic model as a substitute for gene expression profiling-based prognosis methods.

List Of Abbreviations

RFS, recurrence-free survival; WSI, whole slide image; AJCC, the American Joint Committee on Cancer; TCGA-SKCM, The Cancer Genome Atlas Skin Cutaneous Melanoma; FPKM, fragments per kilobase of exon model per million reads mapped; H&E, hematoxylin and eosin; ROI, region of interest; GLCM, gray level cooccurrence matrix; GLRLM, gray level run length matrix; GLSZM, gray level size zone matrix; TMM, the trimmed mean of M values; AUC, area under the curve; FDR , false discovery rate; GO, gene ontology; BP, biological process; CC, cellular component; MF, molecular function; CNA, copy number alteration; Idn, Inverse Difference Normalized; Idmn, Inverse Difference Moment Normalized; cln, clinicopathologic variables, mutation status of BRAF and NRAS genes; expr, the selected gene expression data; im, whole slide image features; cln_expr, the combination of cln and expr; cln_im, the combination of cln and im; expr_im, the combination of expr and im; cln_expr_im, the combination of cln, expr and im.

Declarations

Ethics approval and consent to participate

No ethics approval was required for this work. All data in this study are publicly available.

Consent for publication

Not applicable.

Availability of data and materials

The datasets analyzed during the current study are available in the TCGA-SKCM repository, https://portal.gdc.cancer.gov/projects/TCGA-SKCM.

Competing interests

The authors have no competing interests to declare.

Funding

This work is supported by the Shenzhen Science and Technology Project under Grant JCYJ20180228175315535.

Author's contributions

(I) Conception and design: Yanbin Peng, Jialu Li; (II) Administrative support: Jialu Li; (III) Collection and assembly of data: Yanbin Peng, Youlong Zhang; (IV) Data analysis and interpretation: Yanbin Peng, Youlong Zhang, Jialu Li; (V) Manuscript writing: All authors; (VI) Final approval of manuscript: All authors.

Acknowledgements

Not applicable.

Author's information

Yanbin Peng, p1y1b1@163.com;

Yunfeng Chu, fengerhekuangren@163.com;

Zhong Chen, 925197672@qq.com;

Wen Zhou, 798380277@qq.com;

Shengxiang Wan, sxwan0328@sina.com;

Yingfeng Xiao, yfxiao@163.com;

Youlong Zhang, youlong.zhang@huajiabio.com;

Jialu Li, jialu.li@huajiabio.com.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68(6):394-424.
2. Ali Z, Yousaf N, Larkin J. Melanoma epidemiology, biology and prognosis. EJC Suppl. 2013;11(2):81-91.

3. Chang AE, Karnell LH, Menck HR. The National Cancer Data Base report on cutaneous and noncutaneous melanoma: a summary of 84,836 cases from the past decade. The American College of Surgeons Commission on Cancer and the American Cancer Society. Cancer. 1998;83(8):1664-78.
4. Dickson PV, Gershenwald JE. Staging and prognosis of cutaneous melanoma. Surg Oncol Clin N Am. 2011;20(1):1-17.
5. Hyams DM, Cook RW, Buzaid AC. Identification of risk in cutaneous melanoma patients: Prognostic and predictive markers. J Surg Oncol. 2019;119(2):175-86.
6. Zhao Y, Schaafsma E, Gorlov IP, Hernando E, Thomas NE, Shen R, et al. A Leukocyte Infiltration Score Defined by a Gene Signature Predicts Melanoma Patient Prognosis. Mol Cancer Res. 2019;17(1):109-19.
7. Jayawardana K, Schramm SJ, Haydu L, Thompson JF, Scolyer RA, Mann GJ, et al. Determination of prognosis in metastatic melanoma through integration of clinico-pathologic, mutation, mRNA, microRNA, and protein information. Int J Cancer. 2015;136(4):863-74.
8. Jiang Y, Shi X, Zhao Q, Krauthammer M, Rothberg BE, Ma S. Integrated analysis of multidimensional omics data on cutaneous melanoma prognosis. Genomics. 2016;107(6):223-30.
9. Lu C, and Mrinal Mandal. . Automated analysis and diagnosis of skin melanoma on whole slide histopathological images. Pattern Recognition. 2015;48.8:2738-50.
10. Failmezger H, Muralidhar S, Rullan A, de Andrea CE, Sahai E, Yuan Y. Topological Tumor Graphs: A Graph-Based Spatial Model to Infer Stromal Recruitment for Immunosuppression in Melanoma Histology. Cancer Res. 2020;80(5):1199-209.
11. Robert C, Thomas L, Bondarenko I, O'Day S, Weber J, Garbe C, et al. Ipilimumab plus dacarbazine for previously untreated metastatic melanoma. N Engl J Med. 2011;364(26):2517-26.
12. Robert C, Schachter J, Long GV, Arance A, Grob JJ, Mortier L, et al. Pembrolizumab versus Ipilimumab in Advanced Melanoma. N Engl J Med. 2015;372(26):2521-32.
13. Cancer Genome Atlas N. Genomic Classification of Cutaneous Melanoma. Cell. 2015;161(7):1681-96.
14. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res. 2017;77(21):e104-e7.
15. Doyle S, Feldman MD, Shih N, Tomaszewski J, Madabhushi A. Cascaded discrimination of normal, abnormal, and confounder classes in histopathology: Gleason grading of prostate cancer. BMC Bioinformatics. 2012;13:282.
16. Jones S YZ, Xie Z, et al. A Proposed Data Analytics Workflow and Example Using the R Caret Package.
17. Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics. 2000;56(2):337-44.
18. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139-40.

19. Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol. 2010;11(2):R14.
20. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16(5):284-7.
21. Pastorfide GC, Kibbi AG, de Roa AL, Barnhill RL, Sober AJ, Mihm MC, Jr., et al. Image analysis of stage 1 melanoma (1.00-2.50 mm): lymphocytic infiltrates related to metastasis and survival. J Cutan Pathol. 1992;19(5):390-7.
22. Kornstein MJ, Brooks JS, Elder DE. Immunoperoxidase localization of lymphocyte subsets in the host response to melanoma and nevi. Cancer Res. 1983;43(6):2749-53.
23. Ralfkiaer E, Hou-Jensen K, Gatter KC, Drzewiecki KT, Mason DY. Immunohistological analysis of the lymphoid infiltrate in cutaneous malignant melanomas. Virchows Arch A Pathol Anat Histopathol. 1987;410(4):355-61.

Additional Files

Additional file 1

Table S1: The composition and number of features in each feature set.

Table S2: Summary of C-index and time-dependent AUC.

Table S3: The name and coefficient of features selected in the final image-based model.

Table S4: The likelihood ratio (LR) and its p value of models. “AJCC stage”, “cln” and "cln_im” represent models based on AJCC tumor pathologic stage, baseline variables and the combination of baseline variables and WSI features, respectively. Abbreviations: “All”, all the patients; “AJCC stage<III”, patients within AJCC tumor pathologic stage<III group; “AJCC stage≥III”, patients within AJCC tumor pathologic stage≥III group; “Metastatic”, the group of patients with metastatic tumors; “Locoregional”, the group of patients with locoregional tumors.

Table S5: The median survival time of higher and lower-risk subgroups in each pathologically-defined groups of patients.

Table S6: Summary of treatment information and their RFS associations of the study cohort.

Table S7: Summary of therapeutics type among the 50 patients with pharmaceutical treatment information available.

Table S8: Summary of some omitted clinicopathologic variables routinely used for prognostic analysis in the study cohort.

Figure S1: Three examples of nucleus segmentation results. Figure A shows an image block with a small number of nuclei; B is a block with a higher number of nuclei; C is a block almost all filled with nuclei.

Figure S2: The RFS probability curve of the 152 patients enrolled in this study.

Figure S3: Analysis of variation of cross-validation C-index along with the penalty (log-transformed). Figure A was for models developed based on baseline variables, while figure B for that of based on both baseline variables and WSI features.

Figure S4: The overall survival probability of subgroups stratified by the risk score. A represents all the patients; B represents the patients in AJCC stage<III; C represents the patients in AJCC stage≥III; D represents the patients with metastatic tumors; E represents the patients with locoregional tumors.

Figure S5: The dot plot of the top 20 GO in BP identified by GOseq package. The DE Ratio is the ratio of differentially expressed genes among all the genes in a specific GO category. The GO Description displays the ID and brief information of each GO. The color of dot shows the adjusted p value of the GO term. The size of dot represents the number of differentially expressed genes.

Figure S6: The dot plot of the top 20 gene ontologies in CC identified by GOseq package. The DE Ratio is the ratio of differentially expressed genes among all the genes in a specific GO category. The GO Description displays the ID and brief information of each GO. The color of dot shows the adjusted p value of the GO term. The size of dot represents the number of differentially expressed genes.

Figure S7: The dot plot of the top 20 gene ontologies in MF identified by GOseq package. The DE Ratio is the ratio of differentially expressed genes among all the genes in a specific GO category. The GO Description displays the ID and brief information of each GO. The color of dot shows the adjusted p value of the GO term. The size of dot represents the number of differentially expressed genes.

Figure S8: The directed acyclic graph of the enriched GO terms in biological process category identified by clusterProfiler package. The color represents the significance of GO terms (more significant from yellow to red). The arrow represents the hierarchical relationship between two terms. The shape of each term represents the top 10 significant GO terms (rectangle) and others (ellipse). In each term the GO ID, brief description, FDR, the number of differentially expressed genes and all genes were displayed.

Figure S9: The directed acyclic graph of the enriched GO terms in cellular component category identified by clusterProfiler package. The color represents the significance of GO terms (more significant from yellow to red). The arrow represents the hierarchical relationship between two terms. The shape of each term represents the top 10 significant GO terms (rectangle) and others (ellipse). In each term the GO ID, brief description, FDR, the number of differentially expressed genes and all genes were displayed.

Figure S10: The directed acyclic graph of the enriched GO terms in molecular function category identified by clusterProfiler package. The color represents the significance of GO terms (more significant from yellow to red). The arrow represents the hierarchical relationship between two terms. The shape of each term represents the top 10 significant GO terms (rectangle) and others (ellipse). In each term the GO ID, brief description, FDR, the number of differentially expressed genes and all genes were displayed.

Figure S11: An illustration of WSI processing and feature extraction. A, image foreground segmentation; B, cropping global ROI; C, zooming in the selected global ROI; D, blocks sampling; E, nucleus segmentation; F, sampling nucleus and cropping its ROI; G, extracting texture features from each nucleus ROI; H, extracting texture features from global ROI.

Whole slide image processing and feature extraction.

Differential gene expression analysis.

Computational formulas of texture features included in the final model.

Figures

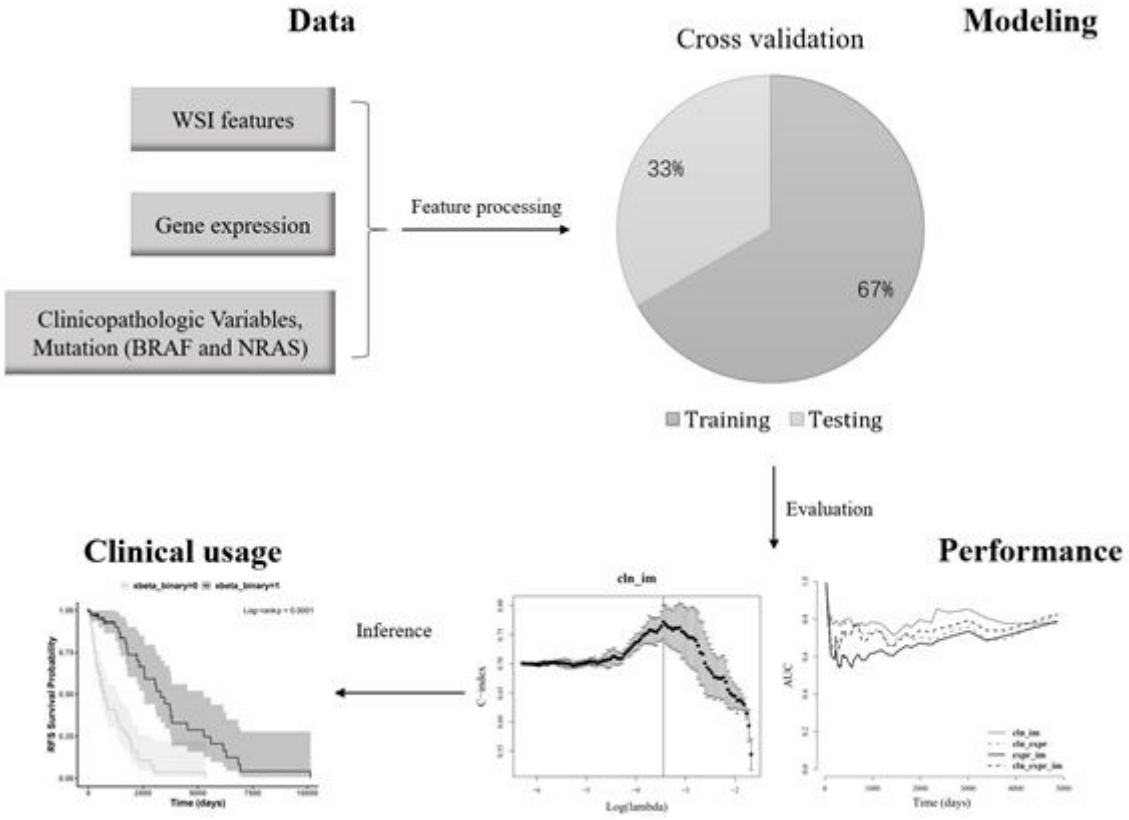

Figure 1

The workflow used in this study.

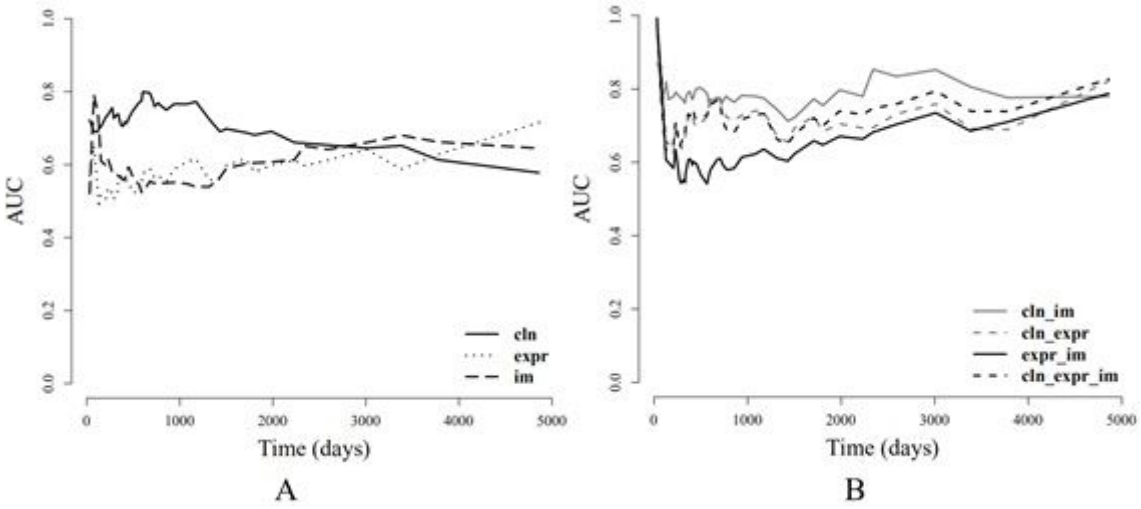

Figure 2

The time-dependent AUC of the models developed based on single-type or multi-type data. A represents the time-dependent AUC of the models based on single-type data; B represents the time-dependent AUC of the models based on multi-type data. Abbreviations: “cln”, baseline variables-based model; expr, models based on gene expression; “im”, models based on WSI features; “cln_im”, models based on baseline variables and WSI features; “cln_expr”, models based on baseline variables and gene expression; “expr_im”, models based on gene expression and WSI features; “cln_expr_im”, models based on baseline variables, gene expression and WSI features.

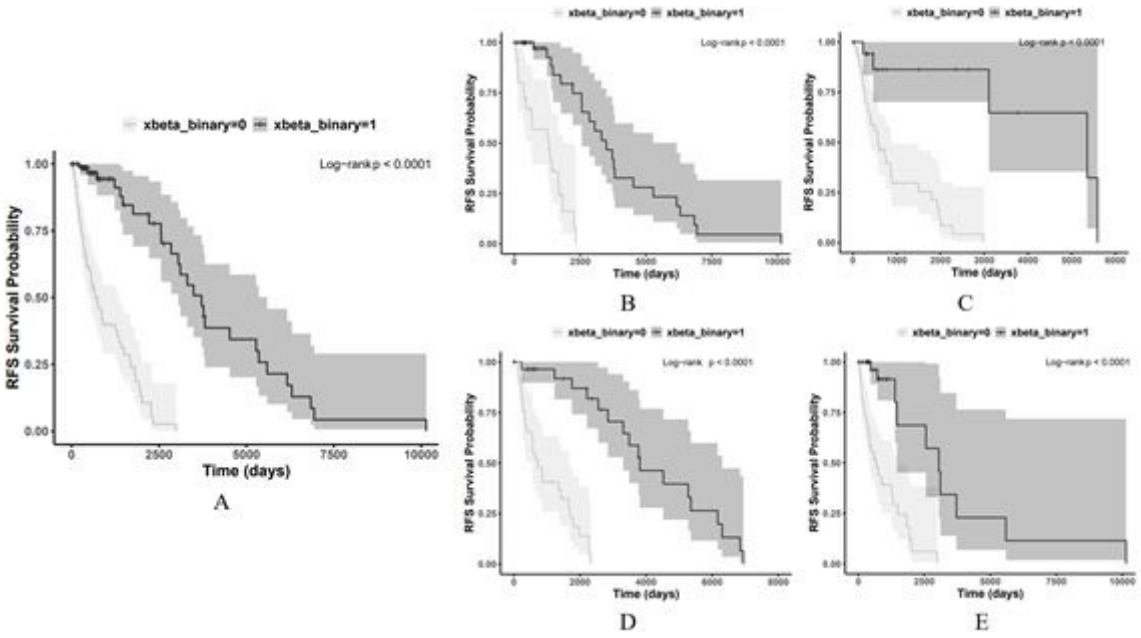

Figure 3

The recurrence-free survival probability of subgroups stratified by the risk score. A represents all the patients; B for those in AJCC stage<III; C for those in AJCC stage≥III; D for those with metastatic tumors; E for those with locoregional tumors.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Equation1.pdf
- Additionalfile1.pdf