Histopathological imaging features- versus molecular measurements-based cancer prognosis modeling

Sanguo Zhang¹, Yu Fan¹,³, Tingyan Zhong²,³ & Shuangge Ma*¹

For lung and many other cancers, prognosis is essentially important, and extensive modeling has been carried out. Cancer is a genetic disease. In the past 2 decades, diverse molecular data (such as gene expressions and DNA mutations) have been analyzed in prognosis modeling. More recently, histopathological imaging data, which is a “byproduct” of biopsy, has been suggested as informative for prognosis. In this article, with the TCGA LUAD and LUSC data, we examine and directly compare modeling lung cancer overall survival using gene expressions versus histopathological imaging features. High-dimensional penalization methods are adopted for estimation and variable selection. Our findings include that gene expressions have slightly better prognostic performance, and that most of the gene expressions are weakly correlated imaging features. This study may provide additional insight into utilizing the two types of important data in cancer prognosis modeling and into lung cancer overall survival.

For most if not all cancers, various prognosis outcomes, such as overall survival, progression free survival, and time to metastasis, are of essential importance. Accordingly, extensive modeling research has been conducted. In “classic” prognosis studies, low-dimensional demographic, clinical, and environmental risk factors are analyzed, and “standard” regression-based techniques (such as Cox model) are usually sufficient. Despite some successes, it has been well recognized that the complexity of cancer prognosis demands additional data and more sophisticated modeling.

Cancer is a genetic disease. In the past 2 decades, with the fast development of high-throughput sequencing techniques, molecular data have been extensively collected in cancer studies. Accordingly, molecular data-based prognosis modeling has been accumulating. For example, an investigation of miRNA expression in 104 pairs of primary lung cancers and corresponding noncancerous lung tissues revealed that high hsa-mir-155 and low hsa-let-7a-2 expressions were correlated with poor survival. The signatures were cross validated using an independent set of adenocarcinomas¹. Since then, hsa-mir-155 over expression has been reported in thyroid carcinoma, breast cancer, colon cancer, and cervical cancer, indicating its potential for serving as a biomarker for tumor detection and evaluation of prognosis outcome². As another example, the study of genome-wide expression of 100 Non-Small-Cell lung cancer (NSCLC) FFPE samples identified a signature composed of 59 genes, which was strongly associated with prognosis for stage I lung cancer patients. This signature was later proven to be robust for clinical usage³. Molecular data are high-dimensional and contain substantial “noises”, that is, the majority of measurements are not associated with prognosis. To effectively remove noises, identify relevant effects, and build reliable models using “signals” only, a myriad of high-dimensional statistical techniques has been developed. A popular family of approaches conducts regularization and applies techniques such as penalization, boosting, Bayesian, and thresholding, which can simultaneously achieve estimation and variable selection. Such techniques have demonstrated statistical, numerical, and empirical successes. We refer to published literature⁴⁺ for reviews and more extensive discussions. With the accumulation of clinical and experimental data, there is increasing knowledge on the functionality of molecular changes. As such, studies have also been conducted using molecular

---

¹School of Mathematics Sciences, University of Chinese Academy of Sciences, Beijing 100049, China. ²SJTU-Yale Joint Center for Biostatistics, Department of Bioinformatics and Biostatistics, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China. ³Department of Biostatistics, Yale School of Public Health, New Haven, CT 06520, USA. ⁴Email: Shuangge.ma@yale.edu
Quality and reproducibility of histopathological images.

Histopathological imaging data.

Whole-slide histopathological images in the svs format are downloaded from the TCGA website (https://portal.gdc.cancer.gov). These tissue slides are formalin-fixed and paraffin-embedded, and the cell morphology is well-preserved and suitable for image feature recognition. They are captured at 20× or 40× magnification by the Aperio medical scanner. In recent studies, we13 and others21,26 have
developed and implemented a pipeline for extracting high dimensional imaging features, which is sketched in Fig. 1. Briefly, it includes the following three main steps. First, whole-slide histopathological images are chopped into small subimages of 500 × 500 pixels, and 20 subimages are randomly selected. Then, imaging features are extracted using CellProfiler27, a publicly available software package that has been adopted in quite a few recent studies11,18,28. In the next step, for each patient, features are averaged. We refer to Zhong et al.13 and Luo et al.26 for more detailed discussions on this imaging processing pipeline as well as alternatives. With this processing pipeline, a total of 299 features can be obtained. We note that this is significantly higher than in studies such as Wang et al.29 and Romo et al.30. As briefly mentioned above, some imaging studies, especially the early ones, utilize low dimensional imaging features. Comparatively, high dimensional features may have less lucid interpretations but can contain information not reflected in the low dimensional features. With their advantages such as cost-effectiveness and reliability, it can be of higher interest to examine their prognosis modeling performance. With the extracted features, we further conduct quality control. In particular, irrelevant features, such as file size and execution information, are removed. We also remove features with severe missingness (> 25%) and no or little variation. A total of 221 features are included in downstream analysis.

Molecular data. For molecular data, we analyze gene expressions, which have been considered in many lung cancer prognosis modeling studies31,32. Compared to DNA and epigenetic changes, gene expressions are “closer” to phenotypes. With a lack of high-quality protein data, TCGA gene expression data have been extensively analyzed for prognosis, other phenotypes, and biomarkers. In TCGA, gene expressions were measured using the Illumina HiSeq2000 RNA Sequencing Version 2 analysis platform and processed and normalized using the RSEM software. More detailed information is available in the literature33,34. It is possible to directly conduct whole transcriptome analysis. However, findings may be unreliable when sample sizes are limited. As such, we take a candidate gene approach. In particular, the 61 gene panel developed in Raz et al.7 is adopted. Matching this panel with gene names in the TCGA data leads to 50 genes for analysis. We acknowledge that there is still a lack of definitive consensus on lung cancer prognosis genes and that there are other lung cancer prognosis gene panels. This particular panel is selected as it has been recently examined in authoritative studies. The proposed analysis can be directly applied to other prognosis panels.

Available data. Beyond imaging and gene expression data, clinical characteristics have also been established as associated with prognosis and included in our analysis. Following published studies and considering data availability, we include sex, age, cancer stage, and tumor size. More specifically, tumor size is defined as the longest dimension × shortest dimension, and we combine cancer stages into three levels to avoid small counts. Multiple types of data are combined by matching unique sample IDs. The final LUAD data contains 307 samples. Among them, 106 died, with survival times ranging from 0 to 88.07 months and a median of 20.52 months. There are also 201 censored subjects, with observed times ranging from 0 to 238.11 months and a median of 23.16 months. The final LUSC data contains 334 samples. Among them, 155 died, with survival times ranging from 0.10 to 173.69 months and a median of 18.36 months. There are also 179 censored subjects, with observed times ranging from 0.39 to 156.54 months and a median of 23.55 months. For both LUAD and LUSC, data on 221 histopathological imaging features and 50 gene expressions are available. Summary statistics on the clinical characteristics are presented in Table 1.

Analysis techniques
Denote T and C as the event and censoring times, respectively. With right censoring, we observe \( (U = \min (T, C), \delta = I(T \leq C)) \). Denote \( X \) as the \( p \)-dimensional vector of histopathological imaging features, \( Z \) as the \( q \)-dimensional vector of gene expressions, and \( L \) as the \( r \)-dimensional vector of clinical characteristics. Assume \( n \) iid samples.

Associate histopathological imaging features and gene expressions with survival. Here our goal is to conduct various “standard” survival analysis and associate imaging features and/or gene expressions with overall survival, while properly accounting for the effects of clinical characteristics. We comprehensively consider multiple sets of analysis.

First consider the analysis with \( X^L = (X', L')' \) as input. Consider the Cox model, under which the hazard function:

\[
\lambda(T|X^L) = \lambda_0(T) \exp \left( \beta'X^L \right).
\]

Figure 1. Pipeline for extracting imaging features.
fitting, we compute the correlation coefficient between
where

To accommodate the high data dimensionality, and to remove noises and identify relevant effects, we consider
its Lasso penalized estimate. Note that the baseline hazard functions in this and the above analysis may be dif-
ferent. In this analysis, although the genes have been pre-selected, it is still necessary to apply penalization. In
particular, the number of variables, relative to the sample size, is still large. As such, certain regularization is
needed in estimation. In addition, to be cautious, it may still be sensible to examine whether all genes in the panel
are automatically included, given their established importance in lung cancer prognosis. For a specific imaging
feature, a nonzero estimate suggests its association with survival. Literature review suggests that penalization is
likely the most popular penalization technique. The adopted “Cox model + Lasso estimation” approach has been
one of the most popular techniques for accommodating high-dimensional input and feature selection, and Lasso
feature, a nonzero estimate suggests its association with survival. Literature review suggests that penalization is
an additive modeling strategy, which has been developed in the literature28 and shown as reasonably effective
for data integration. It retains the "structure" of imaging effects and that of gene expressions. It can be more
interpretable compared to some existing approaches, for example the "black-box" deep learning.

|                  | LUAD (n = 307) | LUSC (n = 334) |
|------------------|---------------|---------------|
| **Sex**          |               |               |
| Female           | 170           | 85            |
| Male             | 137           | 249           |
| **Age**          |               |               |
| 65.49 (SD = 9.71)| 67.38 (SD = 8.59) |
| **Cancer stage** |               |               |
| Stage I          | 3             | 1             |
| Stage IA         | 73            | Stage IA      | 60 |
| Stage IB         | 88            | Stage IB      | 115 |
| Stage II         | 0             | Stage II      | 1  |
| Stage II A       | 28            | Stage II A    | 33 |
| Stage II B       | 49            | Stage II B    | 59 |
| Stage III        | 0             | Stage III     | 0  |
| Stage III A      | 40            | Stage III A   | 46 |
| Stage III B      | 7             | Stage III B   | 14 |
| Stage IV         | 19            | Stage IV      | 5  |
| **Tumor size**   | 0.467 (SD = 0.324) | 0.470 (SD = 0.309) |

Table 1. Summary of clinical characteristics.

Here \( \lambda_0(T) \) is the unknown baseline hazard function, and \( \beta \) is the vector of unknown regression coefficients. Consider the log partial likelihood function:

\[
\ell(\beta) = \sum_{i=1}^{n} \delta_i \left( \beta' \mathbf{X}_i^T - \log \left( \sum_{j=1}^{n} \exp \left( \beta' \mathbf{X}_j^T \right) \cdot Y_j(U_i) \right) \right)
\]

where subscripts \( i \) and \( j \) correspond to subjects \( i \) and \( j \), and \( Y_j(U_i) \) is the subject \( j \)'s at risk indicator at time \( U_i \).

To accommodate the high data dimensionality, and to remove noises and identify relevant effects, we consider the Lasso penalized estimate:

\[
\hat{\beta} = \arg \max \left\{ \ell(\beta) - \tau \sum_{l=1}^{p} |\beta_l| \right\},
\]

where \( \tau > 0 \) is the data-dependent tuning parameter and chosen using cross-validation, and \( \beta_l \) is the \( l \)th component of \( \beta \). Here it is noted that penalization is only imposed on the imaging features. As such, the clinical variables are automatically included, given their established importance in lung cancer prognosis. For a specific imaging feature, a nonzero estimate suggests its association with survival. Literature review suggests that penalization is one of the most popular techniques for accommodating high-dimensional input and feature selection, and Lasso is likely the most popular penalization technique. The adopted “Cox model + Lasso estimation” approach has been examined in multiple published studies35,36. In our analysis, it is realized using the R package glmmT. We note that analysis can also be conducted using other penalties and regularization techniques other than penalization, and that analysis results depend on the adopted technique.

Next we consider the analysis with \( \mathbf{Z}^T = (\mathbf{Z}', L') \) as input. Analysis can be conducted in the same manner as for imaging features. Denote \( \gamma \) as the vector of unknown regression coefficients in the Cox model and \( \hat{\gamma} \) as its Lasso penalized estimate. Note that the baseline hazard functions in this and the above analysis may be different. In this analysis, although the genes have been pre-selected, it is still necessary to apply penalization. In particular, the number of variables, relative to the sample size, is still large. As such, certain regularization is needed in estimation. In addition, to be cautious, it may still be sensible to examine whether all genes in the panel are associated with survival for the particular TCGA patient cohort (which may differ from those examined in published studies).

In the next set of analysis, we integrate the imaging features and gene expressions using an additive approach. In particular, we consider a Cox model with input variable \( \left( \hat{\beta}_1, \ldots, \hat{\beta}_p \right) \mathbf{X}, (\hat{\gamma}_1, \ldots, \hat{\gamma}_q) \mathbf{Z}, L' \). Prior to model fitting, we compute the correlation coefficient between \( (\hat{\beta}_1, \ldots, \hat{\beta}_p) \mathbf{X} \) and \( (\hat{\gamma}_1, \ldots, \hat{\gamma}_q) \mathbf{Z} \), which can suggest whether the two types of data have overlapping information in modeling survival (after adjusting for the clinical variables). In model fitting, as the dimensionality is low, we do not impose any penalization. This analysis takes an additive modeling strategy, which has been developed in the literature38 and shown as reasonably effective for data integration. It retains the “structure” of imaging effects and that of gene expressions. It can be more interpretable compared to some existing approaches, for example the “black-box” deep learning.
For the above three sets of survival analysis, we adopt the following random splitting approach to evaluate prediction performance: (a) randomly split all samples into a training and a testing set with sizes roughly 3:1; (b) conduct survival analysis as described above using the training set; (c) for subjects in the testing set, compute the predicted risk scores. For example, for the analysis with imaging features, the risk scores are $\hat{\eta} X^T$. Compute the C-index using the predicted risk scores and testing set (observed time, event indicator). The C-index ranges between 0 and 1, with a larger value indicating better prediction. It is also the time-integrated AUC (Area under the Receiver Operating Characteristic curve). To avoid an extreme split, Steps (a)–(c) are repeated 100 times, and the average C-index is computed to quantify prediction performance. The goal of this analysis is two-fold. The first is to directly compare prognostic performance of the imaging-based model versus that of the gene expression-based. In addition, this analysis also examines whether integrating the two distinct types of measurements using the additive approach can further improve prediction performance.

### Results

#### Comparison of modeling using histopathological imaging features with gene expressions.

The first set of analysis regresses survival on the imaging features and clinical characteristics. For the variables included in the final models, their estimated regression coefficients are shown in Tables 2 (LUAD) and 3 (LUSC), respectively. It is noted that Level C is chosen as the reference level for stage, thus having an "NA" estimate. Beyond the clinical characteristics, 7 and 9 imaging features are identified, representing AreaShape, Texture, Granularity, and other characteristics. It has been noted in the literature that, unlike omics and some other types of data, high-dimensional imaging features extracted using automated algorithms/software do not have lucid functional interpretations. As such, we do not further pursue bioinformatics interpretations. In the next set of analysis, we regress survival on gene expressions. The identified gene expressions and clinical characteristics as well as their estimated coefficients are shown in Tables 4 (LUAD) and 5 (LUSC), respectively.

#### Associate gene expressions with histopathological imaging features.

As briefly discussed in “Introduction”, histopathological imaging features can be affected by molecular changes, and such a relationship has been studied in some recent publications. We note that this analysis is unsupervised in the sense that it does not involve survival. As such, the most direct goal is not to improve prognosis modeling but rather to understand, in a broad sense, overlapping information contained in the two distinct types of data.

With normalization to zero means, consider the model:

$$X = \eta Z + \varepsilon,$$

where $\eta$ is the $p \times q$ matrix of regression coefficients, and $\varepsilon$ is the $p$-dimensional vector of random errors. Here we model the “downstream” imaging features using the “upstream” gene expressions. Linear regression is adopted with the consideration that more complex modeling may not be reliable with the limited sample size and high dimensionality of both sides of modeling. For estimating $\eta$, consider:

$$\hat{\eta} = \arg \min \left\{ \sum_{i=1,\ldots,n} X_i - \eta Z_i^2 + \tau \sum_{j=1,\ldots,q} \eta_{j,2} \right\},$$

where subscript $i$ corresponds to subject $i$, $\tau > 0$ is a data-dependent tuning parameter and chosen using cross-validation, $\eta_j$ is the $j$th row of $\eta$, and $\| \cdot \|_2$ is the $L_2$ norm. Here to accommodate the high data dimensionality and select gene expressions that are relevant for imaging features, we apply the group Lasso penalization.

Similar to above, to more objectively evaluate the relationship, we consider the following approach: (a) randomly split data into a training and a testing set in the same way as above; (b) conduct the group Lasso estimation using the training set; (c) for the testing set subjects, predict imaging feature values using gene expressions and the training set estimate. For each imaging feature, compute the correlation coefficient between the predicted feature values and testing set (observed time, event indicator). The goal of this analysis is two-fold. The first is to directly compare prognostic performance of the imaging-based model versus that of the gene expression-based. In addition, this analysis also examines whether integrating the two distinct types of measurements using the additive approach can further improve prediction performance.

### Table 2.

Analysis of LUAD data: identified imaging features and clinical characteristics associated with overall survival and their estimated coefficients.

| Imaging feature                                      | Coef   | Clinical characteristic | Coef   |
|------------------------------------------------------|--------|-------------------------|--------|
| AreaShape_Zernike_6_4                                | 0.3697 | Sex                     | −0.0245|
| AreaShape_Zernike_8_6                                | 0.0426 | Age                     | 0.0095 |
| AreaShape_Zernike_9_7                                | 0.1409 | Tumor_Size              | 0.1154 |
| Count_identifytissuergion                            | 0.1759 | Stage_Level_A           | −1.2100|
| Neighbors_AngleBetweenNeighbors_Adjacent             | −0.1033| Stage_Level_R           | −0.2976|
| Neighbors_FirstClosestObjectNumber_Adjacent          | −0.2527| Stage_Level_C           | NA     |
| Threshold_WeightedVariance_identifyhemaprimarynuclei | −4.04E-05|                         |        |
Among the identified genes, there are “familiar” discoveries such as PIK3CG and RND3. In addition, there are also genes that have not yet been well examined in the literature, such as DNMT2 and UQCRC2.

When integrating the combined imaging effect with the combined gene expression effect in one Cox model, for the LUAD data, we obtain regression coefficients 0.9842 (imaging feature, p value = 2.12e−6) and 0.4726 (gene expression, p value = 5.36e−9). For the LUSC data, we obtain regression coefficient 0.9709 (imaging feature, p value = 5.55e−9) and 0.8769 (gene expression, p value = 2.04e−3).

In the random-splitting based prediction evaluation, for the LUAD data, the median prediction C-index values are 0.6202 (imaging features), 0.6864 (gene expressions), and 0.6823 (combined). For the LUSC data, the median prediction C-index values are 0.5466 (imaging features), 0.5606 (gene expressions), and 0.5511 (combined).

More detailed information, for example on the prediction C-index of each split, is available from the authors.

**Remarks**

In the separate survival analysis with imaging features and gene expressions, relevant effects have been identified. For imaging features, extensive additional research will be needed to annotate and fully comprehend the identified variables. We note that this issue has been noted in the literature. In the analysis of gene expression data, the “familiarity” of findings may provide support to the validity of analysis to a certain extent. However, it is noted that more definitive validation will be needed to confirm the findings. The survival analysis with both imaging and gene expression signatures as covariates seems to suggest that the two types of measurements have independent effects. In the random splitting-based evaluation, it is observed that for LUAD, gene expression has moderate predictive performance, and imaging data has moderate/weak predictive performance. For LUSC, both types of measurements have weak predictive performance. For both datasets, gene expression has better performance, which is sensible considering the genetic nature of lung cancer (and other cancers too).

**Table 3.** Analysis of LUSC data: identified imaging features and clinical characteristics associated with overall survival and their estimated coefficients.

| Imaging feature                  | Coef  | Clinical characteristic | Coef  |
|---------------------------------|-------|-------------------------|-------|
| AreaShape_EulerNumber           | −0.1575| Sex                     | 0.5259|
| ObjectNumber                    | −0.2416| Age                     | 0.0231|
| Granularity_12_ImageAfterMath   | 0.2382 | Tumor_Size              | −0.0369|
| Threshold_SumOfEntropies_identifytissueregion | 0.1466 | Stage_Level_A           | −0.7496|
| Location_Center_X.1             | −0.0812 | Stage_Level_B           | −0.4852|
| AreaShape_Center_X              | −0.0903 | Stage_Level_C           | NA    |
| AreaShape_Orientation           | −0.0985 |                         |       |
| Neighbors_AngleBetweenNeighbors_Adjacent | 0.1414 |                         |       |
| Granularity_9_ImageAfterMath    | 0.1395 |                         |       |

**Table 4.** Analysis of LUAD data: identified gene expressions and clinical characteristics associated with overall survival and their estimated coefficients.

| Gene expression | Coef  | Clinical characteristic | Coef  |
|-----------------|-------|-------------------------|-------|
| CCNB1           | 0.0033| Sex                     | 0.0011|
| CTSL            | 0.3694| Age                     | 0.0173|
| GLI2            | 0.2555| Tumor_Size              | 0.0640|
| MFHAS1          | −0.2228| Stage_Level_A           | −1.2460|
| PIK3CG          | −0.3782| Stage_Level_B           | −0.4012|
| RND3            | 0.1841 | Stage_Level_C           | NA    |

**Table 5.** Analysis of LUSC data: identified gene expressions and clinical characteristics associated with overall survival and their estimated coefficients.

| Gene expression | Coef  | Clinical characteristic | Coef  |
|-----------------|-------|-------------------------|-------|
| IL11            | 0.0526| Sex                     | 0.4661|
| MUC1            | 0.0977| Age                     | 0.0309|
| PIK3CG          | 0.0702| Tumor_Size              | −0.4890|
| PRKCA           | 0.1295| Stage_Level_A           | −0.7719|
| WDHD1           | −0.1404| Stage_Level_B           | −0.6034|
|                 |       | Stage_Level_C           | NA    |
Although both LUAD and LUSC are lung cancer subtypes, we observe significantly different results, which can be attributable to the complexity of cancer and suggest that there may not be a definitive conclusion applicable to all cancers. The random splitting evaluation further suggests that integrating the two types of signatures in an additive manner may not further improve prediction, which seems to “contradict” the analysis above. There can be multiple interpretations for this finding. First, the distinction between estimation and prediction should be made – a “good” estimation result may not directly translate into a good prediction. Second, the estimation analysis is repeatedly based on the same data, and there is a risk of over fitting. Third, in the random splitting evaluation, both the training and evaluation are based on fewer observations. An improvement that can be potentially observed with a larger dataset may not be observable with a smaller dataset. It is also noted that penalization and some other sparse approaches have been designed for estimation and may not be ideal for prediction, which may explain the less satisfactory prediction performance observed here.

Association of gene expressions and histopathological imaging features. We first regress imaging features on gene expressions. Detailed information on the identified gene expressions and their estimated coefficients are provided in the Supplementary Materials 1 and 2. In Fig. 2, we show the heatmaps of the estimated coefficients. Briefly, for the LUAD data, in the 50 x 221 coefficient matrix, a total of 7,735 elements are nonzero. A total of 35 genes, including MKI67, ACSL6, NFX1, and WIF1, are identified as associated with the 221 imaging features. For the LUSC data, a total of 6,618 elements are nonzero. A total of 28 genes, including ARAF, BCL7A, NXF1, and TP53, are identified as associated with the 221 imaging features.

The random-splitting based prediction evaluation results are summarized in Fig. 3, where we sort performance, from the worst to the best, across imaging features. More detailed numerical results are provided in the Supplementary Materials.

Remarks The regression analysis suggests that certain gene expressions are connected to imaging features. This observation is sensible considering, as described in “Introduction”, that properties reflected in imaging features are regulated by molecular changes to a certain extent. On the other hand, the prediction results, as shown in Fig. 3, suggest that such associations are mostly weak to moderate. The majority of information in imaging
features cannot be readily explained by gene expressions, and this finding differs from that in some other studies. It is unclear whether such a difference is attributable to the complexity of cancer, difference in analysis approach, or other factors. More exploration, especially a direct comparison, will be needed.

Additional analysis. To complement the above analysis, we conduct additional exploration and present the findings in the Supplementary Materials. In particular, (1) in some cancer studies with high dimensional variables, marginal screening is conducted prior to modeling to reduce dimensionality to a more manageable level. In the above analysis, as the dimensionalities are not as high, screening is not conducted. Results presented in the Supplementary Materials suggest that, for our particular data and analysis, screening can change estimation and identification results, but has no substantial impact on prediction performance. (2) The above penalized estimations involve a tuning parameter, which is selected using cross validation. In the literature, there are many tuning parameter selection methods, and cross validation has been among the most extensively used. In the Supplementary Materials, we show that varying the tuning parameter values near the cross-validation-selected optimal has some moderate impact on estimation. But the findings on prediction are not strongly impacted.

Conclusions

Accurately modeling prognosis and other cancer outcomes has been and will remain an important problem for a long time to come. Molecular and histopathological imaging data have played important roles in cancer prognosis modeling. In particular, with unique advantages including broad availability and high cost-effectiveness, it will be of interest to develop more histopathological imaging-based prognosis modeling. In this study, we have analyzed and integrated molecular and imaging data on the same ground using regularization techniques. More analysis of this kind will be needed to better understand the relative roles that molecular and imaging data play for other cancer types. Some of our findings are “negative”: for example, we have found that integrating data using the additive approach cannot improve prediction. More sophisticated methodological development will be needed to conclude whether this lack of improvement should be attributable to data/cancer type or analysis approach. The revealed interconnections between imaging and molecular features warrants additional investigation.

Received: 1 June 2020; Accepted: 27 August 2020
Published online: 14 September 2020

References

1. Yanaihara, N. et al. Unique microRNA molecular profiles in lung cancer diagnosis and prognosis. Cancer Cell 9, 189–198. https://doi.org/10.1016/j.ccr.2006.01.025 (2006).
2. Faraoni, I., Antonetti, F. R., Cardone, J. & Bonmassar, E. miR-155 gene: A typical multifunctional microRNA. Biochim. Biophys. Acta 1792, 497–505. https://doi.org/10.1016/j.bbapap.2009.02.013 (2009).
3. Xie, Y. et al. Robust gene expression signature from formalin-fixed paraffin-embedded samples predicts prognosis of non-small-cell lung cancer patients. Clin. Cancer Res. 17, 5705–5714. https://doi.org/10.1158/1078-0432.CCR-11-0196 (2011).
4. Ma, S. & Huang, J. Penalized feature selection and classification in bioinformatics. Brief Bioinform. 9, 392–403. https://doi.org/10.1093/bib/bbn027 (2008).
5. Ding, C. & Peng, H. Minimum redundancy feature selection from microarray gene expression data. J. Bioinform. Comput. Biol. 3, 185–205. https://doi.org/10.1142/S0219720005001044 (2005).
6. Saehs, Y., Inza, I. & Larrañaga, P. A review of feature selection techniques in bioinformatics. Bioinformatics 23, 2507–2517. https://doi.org/10.1093/bioinformatics/btm344 (2007).
7. Raz, D. J. et al. A multigene assay is prognostic of survival in patients with early-stage lung adenocarcinoma. Clin. Cancer Res. 14, 5565–5570. https://doi.org/10.1158/1078-0432.CCR-08-0544 (2008).
8. Kothari, S., Phan, J. H., Stokes, T. H. & Wang, M. D. Pathology imaging informatics for quantitative analysis of whole-slide images. J. Am. Med. Inform. Assoc. 20, 1099–1108. https://doi.org/10.1136/amiajnl-2012-001540 (2013).
9. Runn, D. L. et al. An international multicenter study to evaluate reproducibility of automated scoring for assessment of Ki67 in breast cancer. Mod. Pathol. https://doi.org/10.1093/signedpaper/bbb027 (2018).
10. Zha, X. L., Yao, J. W. & Huang, J. Z. Deep convolutional neural network for survival analysis with pathological images. IEEE Int. Conf. Bioinformatics 20, 544–547 (2016).
11. Yu, K. H. et al. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. Nat. Commun. 7, 12474. https://doi.org/10.1038/ncomms8124 (2016).
12. Xu, Y., Zhong, T., Wu, M. & Ma, S. Histopathological imaging-environment interactions in cancer modeling. Cancers 11, 579 (2019).
13. Zhong, T., Wu, M. & Ma, S. Examination of independent prognostic power of gene expressions and histopathological imaging features in cancer. Cancers 11, 361 (2019).
14. Kather, J. N. et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nat. Med. 25, 1054 (2019).
15. Shia, J. et al. Morphological characterization of colorectal cancers in The Cancer Genome Atlas reveals distinct morphology-molecular associations: Clinical and biological implications. Modern. Pathol. 30, 599–609 (2017).
16. Kather, J. N. et al. Pan-cancer image-based detection of clinically actionable genetic alterations. Nat. Cancer https://doi.org/10.1038/s40838-020-00887-6 (2020).
17. Hao, J., Kosaraju, S. C., Tsaku, N. Z., Song, D. H. & Kang, M. PAGE-net: Interpretable and integrative deep learning for survival analysis using histopathological images and genomic data. Proc. Symp. Biocomput. 25, 355–366 (2020).
18. Sun, D., Li, A., Tang, B. & Wang, M. Integrating genomic data and histopathological images to effectively predict breast cancer clinical outcome. Comput. Methods Programs Biomed. 161, 45–53. https://doi.org/10.1016/j.cmpb.2018.04.008 (2018).
19. He, B. J. et al. Integrating spatial gene expression and breast tumour morphology via deep learning. Nat. Biomed. Eng. 4, 827–834. https://doi.org/10.1038/s41551-020-0577-8 (2020).
20. Zhang, Y., Li, A., He, J. & Wang, M. A novel MKL method for GBM prognosis prediction by integrating histopathological image and multi-omics data. IEEE J. Biomed. Health Inform. 24, 171–179. https://doi.org/10.1109/JBHI.2019.2898471 (2020).
21. Yu, K. H. et al. Association of omics features with histopathology patterns in lung adenocarcinoma. Cell Syst. 5, 620–627. https://doi.org/10.1016/j.cels.2017.10.014 (2017).

https://doi.org/10.1038/s41598-020-72201-5
22. Shoag, J. E., Tossoian, J. I., Salami, S. S. & Barbieri, C. E. Unraveling prostate cancer genomics, pathology, and magnetic resonance imaging viability. *Eur. Urol.* 76, 24–26 (2019).

23. Hutter, C. & Zenklusen, J. C. The cancer genome atlas: Creating lasting value beyond its data. *Cell* 173, 283–285. https://doi.org/10.1016/j.cell.2018.03.042 (2018).

24. Radzikowska, E., Glaz, P. & Roszkowski, K. Lung cancer in women: Age, smoking, histology, performance status, stage, initial treatment and survival. Population-based study of 20,561 cases. *Ann. Oncol.* 13, 1087–1093. https://doi.org/10.1093/annonc/mdf187 (2002).

25. Collins, L. G., Haines, C., Perkel, R. & Enck, R. E. Lung cancer: Diagnosis and management. *Am. Fam. Physician* 75, 56–63 (2007).

26. Luo, X. et al. Comprehensive computational pathological image analysis predicts lung cancer prognosis. *J. Thorac. Oncol.* 12, 501–509. https://doi.org/10.1016/j.jtho.2016.10.017 (2017).

27. Carpenter, A. E. et al. Cell profiler: Image analysis software for identifying and quantifying cell phenotypes. * Genome Biol.* 7, R100. https://doi.org/10.1186/gb-2006-7-10-r100 (2006).

28. Zhu, X. L. et al. In *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on.*, 1173–1176 (IEEE).

29. Wang, S. D. et al. ConvPath: A software tool for lung adenocarcinoma digital pathological image analysis aided by a convolutional neural network. *Ebiomedicine* 50, 103–110. https://doi.org/10.1016/j.ebiom.2019.10.033 (2019).

30. Romo-Bucheli, D., Janowczyk, A., Gilmore, H., Romero, E. & Madabhushi, A. Automated tubulinuclei quantification and correlation with Oncotype DX risk categories in ER+ breast cancer whole slide Images. *Sci. Rep.* 6, 32706. https://doi.org/10.1038/srep32706 (2016).

31. Shedden, K. et al. Gene expression-based survival prediction in lung adenocarcinoma: A multi-site, blinded validation study. *Nat. Med.* 14, 822–827. https://doi.org/10.1038/nm.1790 (2008).

32. Navab, R. et al. Prognostic gene-expression signature of carcinoma-associated fibroblasts in non-small cell lung cancer. *Proc. Natl. Acad. Sci. USA* 108, 7160–7165. https://doi.org/10.1073/pnas.1014596108 (2011).

33. Cerami, E. et al. The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* 2, 401–404. https://doi.org/10.1158/2159-8290.CD-12-0095 (2012).

34. Gao, J. et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* 6, l1. https://doi.org/10.1126/scisignal.2004088 (2013).

35. Tibshirani, R. The lasso method for variable selection in the cox model. *Stat. Med.* 16, 385–395. https://doi.org/10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380>3.0.CO;2-3 (1997).

36. Fan, J. Q. & Li, R. Z. Variable selection for Cox’s proportional hazards model and frailty model. *Ann. Stat.* 30, 74–99 (2002).

37. Belloni, E. et al. Whole exome sequencing identifies driver mutations in asymptomatic computed tomography-detected lung cancers with normal karyotype. *Cancer Genet. Ny* 208, 152–155. https://doi.org/10.1016/j.cancergen.2015.02.004 (2015).

38. Tang, Y. et al. Rnd3 regulates lung cancer cell proliferation through notch signaling. *PLoS One* 9, 20. https://doi.org/10.1371/journal.pone.0111897 (2014).

39. Calabro, A. et al. Effects of infiltrating lymphocytes and estrogen receptor on gene expression and prognosis in breast cancer. *Breast Cancer Res. Treat.* 116, 69–77. https://doi.org/10.1007/s10549-008-0105-3 (2009).

40. Newman, A. M. et al. Robust enumeration of cell subsets from tissue expression profiles. * Nat. Methods* 12, 453–457. https://doi.org/10.1038/nmeth.3337 (2015).

41. Chen, C. H. & Lu, T. P. Utilizing gene expression profiles to characterize tumor infiltrating lymphocytes in cancers. *Ann. Transl. Med.* 7, S289. https://doi.org/10.21037/atm.2019.11.59 (2019).

Acknowledgements

We thank the editor and reviewers for their careful review and insightful comments. This work is supported by the University of Chinese Academy of Sciences (Grant no. Y95401TXX2), Beijing Natural Science Foundation (Grant no. Z190004), Key Program of Joint Funds of the National Natural Science Foundation of China (Grant no. U19B2040), NIH (CA241699, CA204120, and CA196530), and a Yale Cancer Center Pilot Award.

Author contributions

Conceptualization, S.M. and S.Z.; methodology, S.M. and S.Z.; software, Y.F.; formal analysis, Y.F.; data curation, T.Z.; writing—original draft preparation, S.M. and Y.F.; writing—review and editing, all authors. All authors have read and agreed to the latest version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-72201-5.

Correspondence and requests for materials should be addressed to S.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020