POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Distant supervised learning for cancer subtyping with multiphase imaging data integration

Laurea Magistrale in Mathematical Engineering - Ingegneria Matematica

Author: Matteo Stefano Savino

Advisor: Prof. Francesca Ieva

Co-advisor: Lara Cavinato

Academic year: 2020-2021

1. Introduction

Intrahepatic Cholangiocarcinoma (ICC) is an aggressive disease of the family of cholangiocarcinomas, which are tumors that stem from cholangiocytes of the biliary tree. Curing ICC requires a complete surgical resection with histologically negative margins (ie, an R status equal to R0). In fact, due to its high aggressiveness, long-term survival is only observed in patients with a complete R0 surgical resection. The R status is the major predictor of overall survival after surgical resection for ICC, while other factors, linked to the extent of the disease, are the lymph node involvement, vascular invasion and distant metastases. Because of its increasing incidence and mortality over the past three decades, it is now more than ever arising the urgency to further characterize the disease at early stages, as to modulate therapies and clinical decisions. In fact, detecting at baseline information that might inform the therapeutic pathway would allow to design more efficient lines of treatments. Such perspective has recently grown and developed in a research field, called personalized medicine, that hinges its root in efficiently extracting insights from multi-source patient data to shape clinical practice.

The promise made by personalized medicine in cancer research calls for special efforts for fully exploiting the information of data generated from different sources. A pivotal role in this sense has been played by imaging texture analysis, i.e., radiomics. It has become more and more important thanks to its advantage to non-invasively give access to tumor characterization. Pertinently, radiomics consists in high-throughput quantitative features extracted from regions of interest in medical images such as CT or MRI scans. These features, also known as radiomic or texture features, can be many and are agnostic with respect to the clinical application. They represent a way to describe the information entailed in medical images and transform such information into matrix-shaped data, easier to handle and study [1].

However, radiomics is known to intrinsically possess some limitations, among all instability with respect to segmentation procedures and complexity in exhaustively shape the imaging representation of the lesions. In the context of ICC research, it has been recently proposed to explore a wider area of liver tumor for analysis, including both the very core of the lesion and the margin surrounding it, as to capture also the information of the tumor-tissue interface [2].

On the other hand, for as high-dimensional as they can be, radiomic data need to be properly analyzed to stratify patients basing on their cancer imaging texture characteristics. Ultimately, such analysis would devise subpopulations with different prognosis on which different therapeutic actions could be implemented. Many different techniques exist in literature to perform cancer sub-tying, mainly related to genomics. Recently, a promising distant-supervised approach has been borrowed from genomics and proposed for radiomic data, with the scope of carrying out clinically insightful patient clustering. Such approach was proven to outperform other cancer subtyping methods proposed for genomic-based stratification purposes [3]. The concept of distant supervision comes from the Natural Language Processing field, where it is used to do relation extraction and sentiment analysis. It consists on the training of a model for a task different from the final scope, using labels that are not completely pertinent with the problem to be tackled. It thus brings the possibility to solve tasks with non-retrievable labels in a supervised way. Here, the aim is to cluster patients in groups with different prognosis exploiting their imaging characteristics to predict survival estimates.

In this work, we exploit the Survival Supervised Graph Clustering (S2GC) model [3] as a distant supervision approach for multiphase imaging-based cancer subtyping in ICC. Our aims and contributions are intended to be two-fold: (1) to provide radiomic characterization of groups of patients at different risk of death from ICC, in a risk stratification fashion, and (2) to study the contributions of the three phases of the CT scans, together with the contributions of core cancer information and of the peritumoral tissue information, as to discuss potentialities and limitations of such approach.

2. Data collection

Our study included two hundred and three patients diagnosed with ICC from six different centers. Per every patient radiomic features, clinical variables and qualitative disease information were collected. Both the segmentation of regions of interest and the feature extraction phases were carried out from all the three phases (Arterial,Portal and Late) of the CT scans by experienced radiologists using the LIFEx software (www.lifexsoft.org). The extracted radiomic data consisted of 50 variables for the core cancer segmentation and 50 variables for the margin segmentation for each one of the three phases of the CT scans. Therefore, the total number of radiomic features available was 300. Pertinently, the margin was computed as the 5-mm region that was semi-automatically generated around the tumor by the software and then manually corrected to ensure that only peritumoral liver tissue had been included. Other personal, i.e. sex and age, and tumor characteristics, i.e., size, number of nodules, ICC pattern and grading, were included along with comorbidities and treatment information.

This study was performed according to the Declaration of Helsinki. The local review board approved the study and informed consent was waived given the observational retrospective design of the study.

3. Methods

The analyses were developed as follows. First, we have performed two supervised analyses employing the logistic regression and the Cox model. In particular, we have built and compared three different logistic regression models, employed to classify whether the death of a patient has occurred within the experiment time. We have always exploited the clinical variables while adding the radiomic variables incrementally, one phase at the time. To assess whether the differences in performances were significant we have employed some McNemar's tests. Instead, for what concerns the Cox model, we have used it as baseline for a survival analysis using all the variables available.

After these analyses, a patient representation has been built from radiomic vectors as extracted from CT regions of interest, i.e., the lesions. Every patient vector carried the information extracted from both the core and the margin of all the three phases. In this sense, three different views of the tumor were assessed and analyzed for stratification, describing the information on the lesion provided by the three different phases. Second, the distant-supervised cancer subtyping has been performed by (1) estimating a patient-to-patient graph basing on their imaging characteristics and survival probabilities and (2) clustering such graph in homogeneous subpopulations of nodes with similar properties. The algorithm's hyperparameters have been optimized. Finally, subpopulations of patients have been clinically characterized with clinical variables, exogenous to the model building, in order to validate the stratification procedure.

3.1. Patient-to-patient graph estimation

According to Supervised Survival Graph Clustering model [3] we performed the abovementioned two steps to perform cancer subtyping and find clinically relevant clusters in ICC patients. The distant-supervised patient-to-patient similarity graph estimation was optimized basing on the following objective function:

$$\begin{aligned}
&\min_{w;S} \sum_{k=1}^{m} \left(- \sum_{i=1}^{n} \delta_i \left(X_i^k w^k - log \sum_{j \in R_i} exp(X_j^k w^k) \right) \right) \\
&+ \lambda \sum_{k \neq j} \|X^k w^k - X^j w^j\|_2^2 + \eta \sum_{k=1}^{m} \|w^k\|_1 \\
&+ \min_{S} \gamma \sum_{i=1}^{n} \sum_{j=1}^{n} (\|X_i - X_j\|^2 + \|X_i w - X_j w\|^2) S_{i,j} + \mu S_{i,j}^2 \\
&s.t. \sum_{j}^{n} S_{I,j} = 1, S_i \succeq 0; i = 1, 2, \ldots, n.
\end{aligned} \tag{1}$$

The loss function in (1) is composed by four terms, each with a specific methodological meaning and a clinical counterpart. The first one represents the estimate of the overall survival risks w^k for each radiomic feature of the $k-th$ view. Estimates were computed by solving the negative partial log-likelihood of the Cox model, where X_i^k is the radiomic vector in $k-th$ view of $i-th$ patient and R_i the set of patients observed alive almost at time T_i. In addition, δ_i is the censoring variable, n is the number of patients and m the number of radiomic views.

Co-regularization between views' contributions on prediction and penalization of covariates are performed by a L2 regularization (second term) and L1 regularization (third term) respectively. Specifically, λ drives the regularization between radiomic views which, in this particular case, refer to the tumor texture of the three different phases of the CT scans. By analyzing the control parameter λ we want to investigate the information provided by the different phases with respect to prognostic risks. On the other hand, the sparsity control parameter η addresses the problem of high-dimensional data, in a feature selection fashion. In addition, importance ranking of features may be deduced according to the penalization coupled with each variable. These terms embody the core of distant supervision. In fact, we predict survival-related risks using a Cox proportional model and intend to exploit such risks, along with the imaging itself, in the definition of similarity between patients. Accordingly, the final term of (1) performs the learning of the graph S structure, i.e., its affinity matrix. It considers both the distance between observations in terms of radiomic views and the survival information of patients estimated in the first term. S is the $R^{n \times n}$ affinity matrix of the patient-to-patient similarity graph where $S_{i,j}$ represents the similarity between patients i and j. γ is the learning rate and μ a trade-off parameter. In this way, two tasks are performed: the survival analysis with the computation of w given S and the similarity graph S estimation given the risks w.

3.2. Hyperparameters optmization

Grid search has been implemented for parameter optimization: optimal values were found for λ (the co-regularization parameter), η (the l_1 penalization parameter) and γ (the learning rate) by maximizing the Harrell's concordance index (c-index) of the estimated survival risks. Values returning the higher c-index were selected as optimal values. The optimal choice was 0.01 for γ, meaning that convergence is almost guaranteed but requires several iterations, whereas regularization was found to be negligible. Indeed, $\lambda = 0$ and $\eta = 0$ were the values that lead to the higher c-index performance.

3.3. Spectral clustering

A spectral clustering algorithm has been implemented for clustering the graph nodes, i.e., the patients, as it is suitable for medical application involving graphs. The number of clusters nc has been chosen by following the eigengap heuristic, which can be applied to the graph Laplacians, either normalized or non-normalized. This consists in choosing nc such that all the eigenvalues up to the $nc-th$ one are small whereas the

$(nc + 1) - th$ one is relatively large. Accordingly, the value of $nc = 5$ was selected. Clusters have been further characterized with exogenous clinical variables, testing differences on survival times and tumor qualitative scores. Radiomic contributions to risk of death from ICC have also been analyzed and discussed. P-values lower than 0.05 were considered significant and Bonferroni correction for multiple testing has been used.

4. Results

Before employing the logistic regression model, due to the highly correlated nature of radiomic data, we have performed a correlation analysis where only the features with a correlation under 0.8 have been kept. Then, we have performed a feature selection thanks to an approach based on the stepwise logistic regression. As result, for all the three models, both clinical and radiomic features have been selected as relevant. The results obtained in this way and applying a 10-fold cross-validation procedure were the following: for the first model, where only the Portal phase have been used, a 0.68 of mean accuracy and a standard deviation of 0.11, for the second model, where also the Arterial phase have been employed, 0.73 with 0.12 as standard deviation while for the third 0.77 and 0.12. As anticipated, thanks to some McNemar's tests, we have found that, with a significance level of $\alpha = 0.05$, the increments between the first model and the second one or the first and the third are statistically significant (p-value = 0.023 and 0.001). Instead, the improvement between the second model and the complete (third) one is not not statistically significant (p-value = 0.227). Also in the case of the Cox model we have performed the a correlation analysis and then, we have applied a principal component analysis in order to reduce the dimensionality. The mean c-index obtained with this approach, again using a 10-fold cross-validation procedure, was 0.67 with a standard deviation of 0.09.

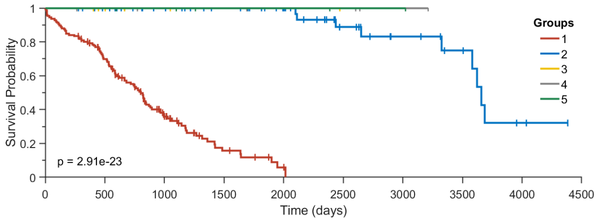

Figure 1: Kaplan-Meier curves estimating clusters' survival probability for the Overall Survival

Then, thanks to the previously described pipeline, we have obtained 5 groups of patients. From Figure 1, which shows the Kaplan-Meier overall survival probability curves, it is possible to notice that there are three small groups: the yellow one, the grey one and the green one. These groups have the same survival curve and the median survival time is not achieved since they are composed only by censored patient. The other two groups are bigger and are clearly characterized by two different survival curves: one with a bad prognosis and the other with a good prognosis. Group 1 (in red), with 126 patients, is the biggest one and has a median survival time of 801 days while Group 2 (in blue), composed by 60 patients, has a median survival time of 3657 days.

Beside life expectancy, we were interested in understanding whether the five groups correspond also to a different clinical characterization. Therefore, some tests on the clinical variables have been performed. The variables that have resulted significant were the following: *Severe complications*, *R status*, *Microscopic vascular invasion*, *Grading* and *Metastatic disease*. We have performed these tests both considering all the five groups and only the major two and the variable resulted as significantly different were the same. In Table 1 we have reported the characterization of the groups according to the clinical variables that resulted significantly different together with the p-values of the respective tests. Regarding *Severe complications*, *R status* and *Microscopic vascular invasion* it is possible to notice that they were more present in the group with the worst prognosis. This is coherent with what seen in the literature. Indeed,

Table 1: Group characterization according to the exogenous categorical clinical variables resulted significant to the tests performed on the proportion of these variables in the different groups.

Variables (% in the group)	Group 1	Group 2	Group 3	Group 4	Group 5	P-value
Metastatic disease	42 (35%)	5 (10%)	0 (0%)	0 (0%)	0 (0%)	0.0011
Severe complications	38 (30%)	3 (5%)	0 (0%)	0 (0%)	0 (0%)	0.0003
R status	52 (41%)	13 (22%)	0 (0%)	1 (17%)	1 (20%)	0.0220
Microscopic vascular invasion	70 (56%)	22 (37%)	3 (50%)	1 (17%)	1 (20%)	0.0375
Grading = 2	60 (48%)	44 (73%)	3 (50%)	5 (83%)	4 (80%)	0.0074
Grading = 3	48 (38%)	8 (13%)	3 (50%)	1 (17%)	1 (20%)	0.0081

Table 2: Weights of the ten most relevant radiomic features

	Arterial phase		Portal phase		Late phase	
Variable	Core Risk	Margin Risk	Core Risk	Margin Risk	Core Risk	Margin Risk
GLZLM_ZLNU	-1.99	75.97	56.58	89.93	28.82	79.44
HUmin	35.35	-100.78	12.29	60.69	4.18	13.41
HUKurtosis	-78.69	-24.62	92.51	-41.92	2.23	44.18
HUQ1	-40.24	7.94	117.96	-212.94	-189.61	-23.78
GLCM_Contrast	15.27	-88.96	-9.91	75.73	-41.51	151.30
NGLDM_Busyness	-64.96	78.37	-175.51	-72.57	224.81	21.23
GLZLM_LZLGE	-66.39	-56.02	17.32	-24.55	7.60	-73.46
HUSkewness	92.05	28.83	-30.19	-36.22	-73.99	100.35
NGLDM_Contrast	-56.96	-36.31	0.90	-59.06	3.62	-34.43

they are well known risk factors especially the *R status*. For what concerns the *Grading*, focusing on the two biggest groups, we have that the Group 2 is characterized by a bigger percentage of Grading 2 while Group 1 by Grading 3 and this is in line with the fact that Group 1 is composed of patients with a more severe situation. In general, we can see that Group 1 is characterized by the more difficult disease to operate and this reflects to the high percentage of patients presenting an R1 status and Severe complications. Finally, we can notice that all the groups with the best prognosis are characterized by patients without a metastatic disease and this too is in line with what is known in literature.

After having clinically characterized these groups, we wanted to analyse their radiomic characterization. A ranking has been made on the radiomic features and their associated risks w. In Table 2 the nine most relevant features have been reported, each with its own counterpart in the other ROI, and highlighted in blue. In this table we have also reported the weights of these variables in the other two views. In most of the cases, especially between the highlighted variables, features provide opposite contributions in the two ROIs to the cumulative risk of death, supporting both the importance and the difference in the two regions of interest. Looking, instead, at all the table it is possible to notice that for the majority of the variables the weights relative to the three phases are different for both sign and magnitude.

5. Discussion

First of all, we can notice that with the supervised analyses we have obtained some poor results. Indeed, we have highlighted some limitation that this approach has in case of radiomic data, such as the need of a high number of samples and the need of a reduced dimensionality that leads to a potential loss of information. However, the comparison of the performances of the three logistic model provides us a first evidence that the three phases contribute with complementary information. Indeed, by adding more views, the prediction performances improve.

Then, looking at the cancer subtyping, we can see that the optimal values of the regularization terms, η for L1 and λ for L2, have been set to zero. On one hand, the null L1 sparsity penalization implies the importance of all the radiomic features in the prognosis estimating process. On the other hand, the null L2 consistency radiomic view regularization suggests that the three phases provide complementary information. Indeed, in the prediction of clinically relevant cancer subtyping, the prognostic information carried by the three views is both mandatory to consider and valuable to access.

As variable-dependent risk coefficients w can be studied according to the penalization factor η, features that are more likely to survive at different levels of η are to be considered robust and important with respect to the task. We recall that these variables have been highlighted in blue in Table 2. Among them, we have noticed how several radiomic variables provided negative, i.e., subtractive, quantities to the patients' cumulative hazard. Interestingly, the very same variables provided a different contribution when coming from the margin of the tumor. For instance, *HU Skewness*, which represents the asymmetry of the Hounsfield distribution, diminishes the risk of death when high in the core area. It however enforces this risk when high in the margin. This means that the more accentuated this difference in the tumor-tissue interface, the more aggressive the disease, thus the poorer the prognosis of the outcome. Similar considerations can be made for *GLCM contrast*, which is the variability of the grey level co-occurrence matrix, and for *GLZLM ZLNU*, which represents the length of the homogeneous zones. Opposite yet analogue conclusions can be drawn for *HU Kurtosis*, that reflects the shape of the Hounsfield distribution relative to a normal distribution, *GLZLM LZLGE*, which is the distribution of the long homogeneous zones with low grey-levels and *NGLDM Contrast*, that measures the difference of intensity between neighbouring regions. Additionally, also when the risk coefficient w brings the same sign in the two ROIs, the absolute value is never equal, leading to a milder yet similar discussion.

It is, also, interesting to notice that an analogous discussion can be made for what concerns the three views. Indeed, looking at the contributions of these variables in the different phases, it is possible to see that they can be different in both sign and magnitude. For example, for the *HU Q1*, which represents the first quartile of the cancer CT Hounsfield values, the weights of the core and the margin in the Arterial phase are respectively negative and positive, while in the Portal phase are the opposite with the core one being positive.This highlights the fact that the different phases provide complementary information all useful to obtain a more complete representation of the tumor in analysis.

A similar discussion can be made also for *GLCM contrast*. Additionally, also when it doesn't happen that the two ROIs exchange their role in the different phases the three phases have weights with different behaviours. Indeed, the weights can be both, i.e. the one of the core and the one of the margin, positive in one phase and both negative in another one or they can be of different magnitude. By looking at Table 2, in particular to the highlighted ones, it is also possible to notice that all the three phases are represented and this is in support to the importance of all the views. Focusing on the highlighted weights and in particular on the ones relative to the margin, it is interesting to notice that variables that measures the heterogeneity such as *GLCM_ Contrast* are linked to a positive weight, while variables that measures the homogeneity as *HUKurtosis* to a negative one. This means that a big heterogeneity in the margin produces an higher risk, while homogeneity is a protective factor. This can be due to the fact that if the peritumoral tissue is more homogeneous it means that it is still predominantly composed by healthy tissues, while heterogeneity means that the disease is penetrated also in the margin.

Instead, from a clinical point of view, it is interesting to highlight that we have found a confirmation of what seen in literature about: the R status, the presence of a metastatic disease and of vascular invasion. Indeed, as expected we have found that the groups with a worst prognosis are characterized by an higher percentage of patients presenting these characteristic. Therefore, this can validate our stratification of patients affected by ICC.

According to these findings, the three phases provide complementary information that have proven their importance to achieve a good performance in both cancer subtyping and survival analysis. Pertinently, a new frontier of texture analysis is currently rising, that is the delta-texture analysis (DTA). In fact, evaluating the difference between two region of interest (spatial DTA) or the same region of interest in separate clinical time instant (temporal DTA) has been shown to be more robust in oncological predictive task. Moreover, the most undiscovered underpinnings of tumor evolution would be explained with models encompassing both delta-radiomic and genomic tumor information.

6. Conclusions

In this work we proposed a distant supervision application for radiomics in Intrahepatic Cholangiocarcinoma patients. We performed cancer subtyping for stratification of patients into clinically relevant subpopulations. We provided radiomic characterization of groups of patients at different risk and we assessed the different contributions of the information provided by the three phases of the CT scans, together with a comparison between the contributions of the core cancer information ant the margin information. Such application could pave the way to both temporal and spatial delta-texture analysis in cancer research.

References

[1] R. J. Gillies *et al.*, "Radiomics: images are more than pictures, they are data," *Radiology*, vol. 278, no. 2, pp. 563–577, 2016.

[2] F. Fiz *et al.*, "Contrast administration impacts ct-based radiomics of colorectal liver metastases and non-tumoral liver parenchyma revealing the "radiological" tumour microenvironment," *Diagnostics*, vol. 11, no. 7, p. 1162, 2021.

[3] C. Liu *et al.*, "Supervised graph clustering for cancer subtyping based on survival analysis and integration of multi-omic tumor data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2020.