Contribution of imaging-genetics to overall survival prediction compared to clinical status for PCNSL patients

Amine Rebei, Agusti Alentorn, Hamza Chegraoui, Vincent Frouin, Cathy Philippe

CONTRIBUTION OF IMAGING-GENETICS TO OVERALL SURVIVAL PREDICTION COMPARED TO CLINICAL STATUS FOR PCNSL PATIENTS

Amine Rebei[⋆]*, Agusti Alentorn*[†]*, Hamza Chegraoui*[⋆]*, Vincent Frouin*[⋆]*, Cathy Philippe*[⋆]

[⋆] Université Paris-Saclay, CEA, Neurospin, 91191, Gif-sur-Yvette, France
[†] Sorbonne Université, Inserm, CNRS, UMR S 1127, ICM, AP-HP
Hôpital La Pitié Salpêtrière, Charles Foix, Service de Neurologie 2-Mazarin, Paris, France

ABSTRACT

Accurately predicting the survival of patients with cancer has the potential to substantially enhance and customize the treatment strategies. Integrating and using all the patients' available data is essential to make the most accurate predictions.
In this work, we gather clinical, imaging and genetic data into one mono-block multivariate survival analysis for patients with primary central nervous system lymphoma (PCNSL). As a first step, we select the best features from each pre-processed dataset. Then we assemble and use the resulting block to predict overall survival with a survival random forest algorithm. The assessment of the proposed method yielded a C-index of 0.776. We thus conclude that multimodal data integration significantly improves prediction performance.

***Index Terms*—** Survival analysis, imaging genetics, MRI, Brain Tumor

1. INTRODUCTION

Primary CNS lymphoma (PCNSL) is a diffuse large B-cell lymphoma (DLBCL), occurring in the brain and that never spreads to the rest of the body. It is a rare tumor that accounts for $\leq 1\%$ of all lymphomas, and approximately 2% of all primary CNS tumors [1]. This disease generally follows an aggressive course and still has very high mortality despite advances in its treatment.
Recently two important contributions have been made in the understanding of this disease. On the one hand, a Genome-wide association study (GWAS) susceptibility study showed that some SNPs are associated with the onset of PCNSL [2]. On the other hand, retrospective survival analysis studies showed that different treatments and clinical characteristics had significant effects on survival [3]. Accurate survival analysis can be very helpful to personalize treatments. For example, patients with poorer prognoses might need closer follow-up and different, more suitable treatment.
To further understand PCNSL tumors, we hypothesize that the data available in clinical research cohorts could improve clinical management if they were jointly analyzed. Several studies showed the predictive power of imaging and/or genetic data [4]. Data integration is a promising approach that can help reduce uncertainty from classic clinical prognostications by providing an automatic, data-driven process that gives accurate and consistent results.
In this study, we apply such methods for the first time on a cohort of patients with PCNSL built by the French national Lymphoma Oculo-Cerebral (LOC) Network. We use concatenated clinical, imaging and genetic data, with the aim of improving survival prediction performance compared to typical clinical-only predictions.

2. MATERIAL

2.1. Patients and clinical block

In the LOC cohort, genotyping data were produced for 346 immunocompetent HIV-negative patients with PCNSL [2], among whom 250 patients had a brain MRI scan at diagnosis available to us. Furthermore, clinical data comprising age, sex, treatment, disease progression and Karnofsky score are available. The Karnofsky score is a way to measure the general health status and the ability of cancer patients to perform ordinary tasks. It ranges from 0 meaning death, to 100 meaning life with normal activity and no complaints. This score is used by the physicians for prognosis and will constitute, with the age and sex, the baseline predictor.
For the 148 patients (80 males) for which we have complete pre-processed data that passed all quality checks (clinical, genetic and at least one MR image), the median age is 66 years and the median survival time is 1157 days (3 years and 2 months), ranging between 36 and 7923 days (21 years and 8 months).

2.2. Imaging block

Although the rarity of the cancer makes the LOC cohort outstanding, brain MRI data are quite non-homogeneous. MRIs were acquired at diagnosis on systems with very diverse field strengths (1, 1.5 and 3 T) at 40 sites. The tumors were manu-

ally segmented by a neurologist.
For each of the 250 patients with MRI, between 1 and 4 images are available, for a total of 753 MRI scans and an average of 3 different MR sequences per patient. The following sequences were studied: T1-weighted (T1w), contrast-enhanced T1-weighted (ceT1w) with gadolinium, T2-weighted (T2w), T2-weighted fluid-attenuated inversion recovery (FLAIR). The images were bias-corrected then their intensity was normalized using the hybrid white stripe method [5]. At each step of the pre-processing pipeline, quality check was performed and patients with poor merit metrics were eliminated. Since the dataset does not contain all modalities for all patients, we considered the most frequently available ones and study the ceT1w images (148), the FLAIR images (104) and the patients for whom both are available (104). For each image, we used the support of the segmented tumor and radiomic features were extracted from each corrected image using the python package PyRadiomics [6], including Shape-based (8 features/image), first-order statistics (18) and second-order statistics (75). In total 845 radiomics features were extracted from the images: 101 from the original and 93×8 from the eight wavelet-derived images using the Coiflet wavelet. These features were then standardized using a z-score and highly correlated features (correlation threshold 0.99) were eliminated.

2.3. Genotyping block

Genotyping data was acquired from blood samples as described in [2] and imputed, yielding the identification of 5 SNPs associated with PCNSL risk. These SNPs were linked to 5 genes: *EXOC2*, *ANO10*, *PVT1*, *BACH2* and *HLA-DRA*. All SNPs with genomic location matching these genes, as well as those located within 1 megabase upstream and downstream to account for their regulatory regions, were extracted. This block of genotyping data was then pruned using the PLINK software [7] to select SNPs in approximate linkage equilibrium with each other and to avoid strong colinearity (using VarianceInflationFactor=10 and windowSize=40, and WindowStep=10). For each of these SNPs, we derived two variables: the allelic dosage (0/1/2) as well as the dominant component (0/1). The SNPs with the same value across all patients were further discarded, leading to a total number of 45045 SNPs (90090 variables).

3. METHODS

3.1. Machine Learning procedure

The patients of the three-block dataset were classically divided into a TRAIN set (80%) and a TEST set. The averages on the distributions of age, survival and sex were kept roughly identical between the TRAIN and TEST sets (values differ by less than 5% between the different train/test splits). The genetic and imaging blocks went independently through a feature selection step using a 3-fold cross validation (CV) procedure on the TRAIN set - each fold being divided into a training and a validation subsets, defining a nested CV). For each block, in the nested CV framework, a univariate Cox proportional hazard model was first fit for each feature separately in order to rank them according to their C-index. An exhaustive grid search[8] for the optimal number of features k in the multivariate survival model chosen was then performed. The features selected from imaging and genetic blocks, as well as the clinical data, were then concatenated and used for the final performance calculations in the TRAIN and TEST sets. (Figure 1).

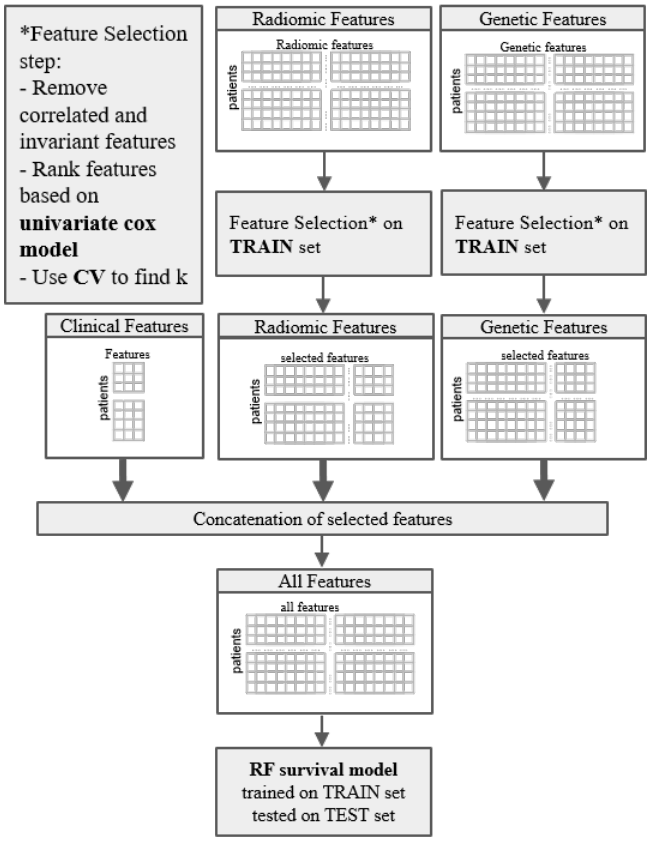

Fig. 1. Data integration process.

3.2. Survival Analysis

To predict the overall survival, we used the random survival forests, as implemented in scikit-survival [9]. Random survival forests are dedicated to right-censored survival data analysis and show good performance in dealing with high dimensional data by performing feature selection while remaining robust to outliers and noise[10]. The results depending on the random starting point, the performance scores were then

averaged over 10 runs with fixed and reproducible random states. To estimate the prediction performance, we use Harrell's concordance index. The C-index (concordance index) is related to the area under the ROC curve. It estimates the probability that, in a randomly selected pair of cases, the case that fails first had a worst predicted outcome. As such, the concordance index focuses on the predictions ranking rather than on predictions themselves and accounts for censored data. To study the importance of the features used for the prediction, the permutation feature importance weights were computed. The permutation feature importance is defined to be the decrease in a model score when a single feature value is randomly shuffled.

4. RESULTS

In this section, we compare overall survival prediction performance of imaging-genetics model versus a simpler model based on immediately available clinical data.

4.1. Clinical Data

Training a random survival forest estimator on 80% of the cohort (118 patients) yields an average C-index of 0.69 ± 0.011 when the estimator is tested on the remaining 20% (see Table 1). Using the feature importance permutation, the weights for age, Karnofsky score and sex were 0.076 ± 0.0027, 0.089 ± 0.0032 and 0.034 ± 0.0022, respectively. Actually, a model fitted only with age and Karnofsky score yields the same C-index of 0.69 ± 0.008 on average.

4.2. Prognostic Imaging Features

Joint features selection using the algorithm described in section 3.1 with both imaging modalities resulted in 24 features, comprising 2 features from FLAIR images (Mean Absolute Deviation and GLDM Low Gray Level Emphasis). Using these features, a C-index of 0.67 ± 0.018 is obtained (Table 1). Based on permutation importance procedure, The most important features are the FLAIR GLDM Low Gray Level Emphasis (0.036 ± 0.0031) and the ceT1w GLDM Large Dependence Low Gray Level Emphasis (0.033 ± 0.0038).

For the 148 ceT1w images taken alone, 7 radiomic features were selected: 3 second-order features and 4 wavelet-filtered features. Using these features, the average C-index over 10 runs is 0.561 ± 0.014. The two most important features were the wavelet LLL GLDM Small Dependence Low Gray Level Emphasis with a weight of 0.054 ± 0.0026 and the wavelet HLL GLCM Cluster Shade with a weight of 0.052 ± 0.0039.

For the 104 FLAIR images taken alone, 37 features were selected including 33 features extracted from the wavelet-filtered images, surface to volume ratio, the GLCM Informational Measure of Correlation (IMC) 2, the GLDM Large Dependence High Gray Level Emphasis and the GLDM Small Dependence Low Gray Level Emphasis. The average C-index is 0.564 ± 0.012. By far the most important feature is wavelet LHL GLCM Maximal Correlation Coefficient with a weight of 0.042 ± 0.0029, while the second more important one (wavelet HHL GLDM Dependence Variance) weighs 0.025 ± 0.0013.

C-index	Clinical	Imaging	Genetic
Train	0.89 ± 0.001	0.89 ± 0.001	0.94 ± 0.001
Test	0.69 ± 0.011	0.67 ± 0.018	0.71 ± 0.014

Table 1. Survival prediction performance of each mono-block models. Mean C-indices and standard deviations over 10 runs.

4.3. Prognostic Genetic Features

At the end of the pre-processing pipeline, the genetic dataset contains 450045 SNPs, with two measurements for each SNPs : the allelic dosage (encoding A) and the dominant component (encoding D). We relied on the feature selection step (section 3.1) and the natural feature selection in the random forest algorithm to decide which one of these 2 measurements is the most appropriate.
The feature selection step yielded 2478 genetic features. It should be noted that the SNPs associated with the susceptibility of PCNSL identified in [2] do not appear to be prognostic. The selected features lead to an average C-index of $0.71 \pm$ over 10 runs. The 4 most important ones are located on chromosome 6 (Table2). Two of these SNPs are related to the *MAP3K7* gene, with dominant component. This gene is not one identified by [2] but is situated downstream of *BACH2*. Its protein controls cell functions such as apoptosis and regulates *TNF*, (Tumor Necrosis Factor). Its dysregulation is associated with Alzheimer's disease[11] and cancers. Interestingly, mutations in *MAP3K7* have been associated with DLBCL oncogenesis by regulating the *NF-κB* pathway[12].

Weight	Chr	Position (bp, hg19)	SNP	Nearest Gene	Encoding
0.0200 ± 0.0073	6	91296420	rs282070	*MAP3K7*	D
0.0191 ± 0.0059	6	1054796	rs73716742	*LINC01622*	D
0.0188 ± 0.0056	6	90616785	rs398184	*GJA10*	A
0.0187 ± 0.0108	6	91241659	rs205345	*MAP3K7*	D

Table 2. Permutation weights and gene annotations for the 4 best SNPs. A : Allelic dosage, D : Dominant component.

4.4. Data integration

Compared to the classic clinical predictive model, the imaging data alone are less efficient and the genetic data alone performs a bit better but maybe not significantly.

When concatenating the clinical and the 24 imaging features, the performance over the clinical data does not improve significantly, with an average C-index of 0.704 ± 0.019. Actually, a model based on the 3 best imaging features and the clinical data is enough to reach the same performance (Table 3). Interestingly, integrating the clinical and the 2478 genetic features improves the performance, with a C-index of 0.77 ± 0.019. As with imaging data, this result can be obtained using only the best 781 features, the rest having no impact on performance once the clinical data is factored in.
Adding imaging data to the clinical-genetic block does not significantly improve the performance, with a C-index of 0.776 ± 0.018.
In an integrative point of view, only the genetic data bring an additive predictive value to the baseline model.

C-index	Clinical Imaging	Clinical Genetic	Clinical Imaging Genetic
Train	0.90 ± 0.001	0.93 ± 0.001	0.93 ± 0.001
Test	0.704 ± 0.019	0.77 ± 0.019	0.776 ± 0.018

Table 3. Survival prediction performance of different integrated models. Mean C-indices and standard deviations.

5. CONCLUSION

The results presented in this work are two-fold. First, we compared the performance obtained by concatenation of blocks of clinical, imaging and genetic data to predict the survival of PCNSL patients. We showed that an integrative approach surpassed clinical predictions significantly. While a good performer on its own, the imaging block performance was shown to be mostly redundant with the genetic and clinical blocks. This could be remedied by improving the dataset quality and further correcting multisite effects. This work will have to be replicated in other similar studies.
Second, studying the genetic performance showed that SNPs predicting the susceptibility to PCNSL did not necessarily correlate with its clinical evolution. Nonetheless, the selected SNPs yield a predictor with a significant performance uplift over the one with clinical data only. This performance may be improved when we include all the genetic data available, without focusing on the susceptibility genes. Further investigations on molecular implications of our findings may lead to patients stratification improvement and to a potential application in personalized treatments of PCNSL.
This work showed that the joint use of the three blocks of clinical, imaging and genotyping data allows us to train a better predictor than with clinical data alone. Yet, the identification and interpretation of important predictor variables is hindered when blocks are trivially concatenated. Therefore, as a perspective, a multi-block approach [13] could be applied on these data, with various structured sparse constraints, depending on the nature of each block.

6. COMPLIANCE WITH ETHICAL STANDARDS

Collection of patient samples, imaging and associated clinicopathological information was undertaken with written informed consent and ethical review board approval.

7. ACKNOWLEDGMENTS

This work was partially funded by the PRT-K/INCa grant *LOC-model* reference 2017-1-RT-04-CEA-1 with data from patients included in the LOC Network supported by the INCa. The authors have no conflicts of interest to declare.

8. REFERENCES

[1] Khê Hoang-Xuan et al., "Diagnosis and treatment of primary CNS lymphoma in immunocompetent patients: Guidelines from the European Association for Neuro-Oncology," *The Lancet Oncology*, vol. 16, no. 7, pp. e322–e332, jul 2015.

[2] Karim Labreche et al., "A genome-wide association study identifies susceptibility loci for primary central nervous system lymphoma at 6p25.3 and 3p22.1: a LOC Network study," *Neuro-Oncology*, vol. 21, no. 8, pp. 1039–1048, aug 2019.

[3] Yudong Shan et al., "Prognostic factors and survival in primary central nervous system lymphoma: A population-based study," *Disease Markers*, vol. 2018, 2018.

[4] Philipp Kickingereder et al., "Radiomic profiling of glioblastoma: Identifying an imaging predictor of patient survival with improved performance over established clinical and radiologic risk models.," *Radiology*, 2016.

[5] R.T. Shinohara et al., "Statistical normalization techniques for magnetic resonance imaging," *NeuroImage: Clinical*, vol. 6, pp. 9–19, jan 2014.

[6] J.J.M. Van Griethuysen et al., "Computational radiomics system to decode the radiographic phenotype," *Cancer Research*, vol. 77, no. 21, pp. e104–e107, nov 2017.

[7] S. Purcell et al., "PLINK: a tool set for whole-genome association and population-based linkage analyses," *Am. J. Hum. Genet.*, vol. 81, no. 3, pp. 559–575, Sep 2007.

[8] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[9] Hemant Ishwaran et al., "Random survival forests," *Annals of Applied Statistics*, vol. 2, no. 3, pp. 841–860, sep 2008.

[10] Hong Wang and Gang Li, "A Selective Review on Random Survival Forests for High Dimensional Data," *Quantitative Bio-Science*, vol. 36, no. 2, pp. 85–96, nov 2017.

[11] Xin Wang et al., "Therapeutic Potential of AMP-Activated Protein Kinase in Alzheimer's Disease," *Journal of Alzheimer's Disease*, vol. 68, no. 1, pp. 33–38, 2019.

[12] Mara Compagno et al., "Mutations of multiple genes cause deregulation of NF-B in diffuse large B-cell lymphoma," *Nature*, vol. 459, no. 7247, pp. 717–721, jun 2009.

[13] M. Tenenhaus et al., "Regularized Generalized Canonical Correlation Analysis: A Framework for Sequential Multiblock Component Methods," *Psychometrika*, vol. 82, no. 3, pp. 737–777, 9 2017.