2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Impact of ComBat and a Multi-Model approach to deal with multi-scanner and missing MRI data in a small cohort study. Application to H3K27M mutation prediction in patients with DIPG.

Fahad Khalid[1], Jessica Goya-Outi[1], Vincent Frouin[2], Nathalie Boddaert[3], Jacques Grill[4] and Frédérique Frouin[1]

Abstract— **Radiomics was proposed to identify tumor phenotypes non-invasively from quantitative imaging features. Calculating a large amount of information on images, allows the development of reliable classification models. In multimodal imaging protocols, the question arises of adding an imaging modality to improve model performance. In addition, in the implementation of clinical protocols, some modalities are not acquired or are of insufficient quality and cannot be reliably taken into account. Furthermore, multi-scanner studies generate some variability in the acquisition and data. Some methodological solutions using ComBat and a multi-model approach were tested to take these two issues into account. It was applied to a cohort of 88 patients with Diffuse Intrinsic Pontine Glioma (DIPG). Sixteen models using radiomic features computed using 0, 1, 2, 3 or 4 MRI modalities were proposed. Based on Leave-One-Out Cross-Validation, F1 weighted scores ranged from 0.66 to 0.85. A model of majority voting using the prediction of all the models available for one given patient was finally applied, reducing drastically the number of unclassified patients.**

Clinical relevance— **In case of patients with DIPG, the prediction of H3 mutation is of prime importance in case of inconclusive biopsy or in the absence of it. It could suggest orientations for new chemotherapy drugs associated with the radiation therapy.**

I. Introduction

Diffuse intrinsic pontine glioma (DIPG) is a highly aggressive pediatric tumor, with a median survival of 9–11 months [1]. Due to its position in the brainstem, surgical intervention is not an option, and conventional chemotherapy has proven to be ineffective. Currently, radiation therapy is the only standard care that temporarily mitigates the symptoms, delays disease progression, and extends median survival by a few months. Recent studies have shown that approximately 80% of DIPG harbor mutations at genes encoding histone H3K27M. Most current mutations are H3.1 (HIST1H3B) and H3.3 (H3F3A). These mutations are currently identified from biopsy samples and are associated with patient response to therapy [2]. Some clinical trials to assess therapy options according to these mutations are currently under investigation. In a previous work, we have proposed to predict the two types of histone H3K27M mutations non-invasively using MRI-based radiomic features [3]. The ultimate achievement would be to define whether this could avoid biopsy, or at least replace it when it is not feasible or not conclusive, and guide patient care from diagnosis time. The present work aims at optimizing this predictive model [3] in a larger cohort. However, the introduction of new patients (coming from the same center) introduces some variability in the image database due to the use of two different scanners (1.5T and 3T scanners). Furthermore, our first model was based on the joint use of clinical data and four types of structural MR images: T1-weighted (T1w), T2-weighted (T2w), T1-weighted post-contrast injection (T1c) and T2-weighted FLAIR (FLAIR) images. To increase the number of patients, and the predictive power of the prediction models, we consider patients having less than four modalities (at least one among the four).

Radiomics consists in the extraction of quantitative imaging features to identify tumor phenotypes with some predictive values. It faces the critical issue of lack of reproducibility that hampers the successful translation of radiomic model discovery into better diagnosis, patient classification or monitoring. With the introduction of an additional cohort with differences in scanner field and settings, radiomic features were expected to differ between the two scanners, hence ComBat harmonization was introduced [4]. In addition to image intensity standardization [5], ComBat is dedicated to the harmonization of the radiomic features which are associated with one specific tissue, the tumor in the present case. Furthermore, to take advantage of all the MR modalities available for each patient, a multi-model approach is built, using the 16 combinations of the four MR modalities.

II. Patients and Methods

A. Clinical and image data

This monocentric retrospective study (2014-2019) included 88 patients with DIPG, scanned at the diagnosis time with one of the two scanners of our center and at least one of the four structural MRI modalities: T1w, T2w, T1c and FLAIR (see Table I).

*This work was supported by Gustave Roussy grant CAJ 2020_088

[1]LITO U1288, Inserm - Institut Curie, 91400 Orsay, France fahad.khalid@inserm.fr, jessicaouti@gmail.com frederique.frouin@inserm.fr

[2]GAIA, Neurospin, CEA, 91191 Gif-sur-Yvette, France vincent.frouin@cea.fr

[3], Pediatric Radiology Department, Hôpital Necker Enfants Malades, AP-HP; IMAGINE Institute, Inserm, Université de Paris, 75015 Paris, France nathalie.boddaert@aphp.fr

[4]Department of Pediatric and Adolescent Oncology, Gustave Roussy, Inserm, Université Paris-Saclay, 94800 Villejuif, France jacques.grill@gustaveroussy.fr

TABLE I
MRI DATASET PROPERTIES

	1.5T	3.0T
Patients	**71**	**17**
T1w	55	10
T1c	52	12
T2w	51	15
FLAIR	50	12

A total of 88 patients were scanned using either a 1.5T (Signa HDxt, GE Medical Systems) or a 3T scanner (MR-750, GE Medical Systems). Among them, 17 presented H3.1 mutation type, 47 H3.3 mutation type and 24 were wild type (WT), or presented another mutation type, or their mutation status was unknown. A total of 17 patients was scanned using the 3T scanner; among them, 9 were H3.3 mutated, 2 were H3.1 mutated, and 6 mutations were unknown or wild type. The clinical feature set consisted of age at the time of diagnosis and sex of patients (see Table II). Patients with H3.3 mutations were older at diagnosis than patients with H3.1 mutation (Wilcoxon test, $p<0.01$). Table III indicated the number of patients with H3.1 or H3.3 mutation available by considering each combination of 1 to 4 MR modality. Only 47/64 patients (73%) presented the four MR modalities.

TABLE II
CLINICAL CHARACTERISTICS OF PATIENTS INCLUDED IN THE STUDY

	H3.1	H3.3	WT/Unknown
Patients	**17**	**47**	**24**
Age (years)	4.9 ± 1.7	8.7 ± 3.6	8.8 ± 6.1
Girls/boys	10g/7b	22g/25b	9g/15b

B. Image feature extraction

Images were pre-processed by a dedicated pipeline [5] including intensity standardization according to the hybrid white stripe approach, resampling to isotropic voxels (1 mm^3) and multi-modal image registration to each T2w scan (when available, T1 or FLAIR otherwise). For each patient, a large spherical region was drawn inside the tumor on the T2w scans (if available, T1 or FLAIR otherwise) and reported in T1w, T1c and FLAIR scans. Radiomic features were extracted using PyRadiomics [6]. A total of 79 features including first-order and texture features were computed for each MRI modality.

C. Harmonization of features using ComBat

In brain MR, standardization approaches have been proposed to correct for the intensity variability. For instance, the hybrid white stripe method proved to be successful in the context of neurodegenerative diseases and brain cancer [5]. However, we showed that this procedure was not sufficient to explain differences in radiomic features observed for the same patients undergoing 1.5T and 3T scans [4]. The further use of ComBat to harmonize radiomic features has been validated. For each radiomic feature y, computed for one given modality, measured in patient j with scanner i, the scanner effect on feature y_{ij} can be modelled as (1):

$$y_{ij} \sim \alpha + \gamma_i + \sigma_i \epsilon_{ij}, \tag{1}$$

where α is the overall value of the radiomic feature y, γ_i is an additive scanner effect and σ_i a multiplicative scanner effect associated to an error term ϵ_{ij}. ComBat estimates the $\hat{\alpha}$, $\hat{\gamma_i}$, and $\hat{\sigma_i}$ terms, and corrected values y^*_{ij} are computed by (2):

$$y^*_{ij} = \frac{y_{ij} - \hat{\alpha} - \hat{\gamma_i}}{\hat{\sigma_i}} + \hat{\alpha} \tag{2}$$

Finally, values y^{**}_{2j} obtained for the second scanner (3T) are realigned to the values obtained for the first scanner (1.5T) according to (3), with $\hat{\mu} = \hat{\sigma_1}/\hat{\sigma_2}$:

$$y^{**}_{2j} = \hat{\mu} y_{2j} + \hat{\alpha}(1 - \hat{\mu}) + (\hat{\gamma_1} - \hat{\mu}\hat{\gamma_2}) \tag{3}$$

D. Machine Learning Models

Five feature sets were considered as the inputs of the predictive models: one feature set per MRI modality and one clinical feature set. To benefit from all the available patient modalities, 16 models (Mk, $1 \leq k \leq 16$) were built. Table III represents for each model the type of modality it accepts. This allows the original data set to stay unchanged while addressing the missing data handling problem. For each model, a three steps selection procedure was applied to the imaging features:

- Features were selected according to their robustness to the spherical delineations. Based on features computed in dilated and eroded versions of the tumor region, the absolute agreement intraclass correlation coefficient (ICC) of each feature was computed. Only features with ICC > 0.9 were kept.
- Only features presenting an individual Area Under the Receiver Operating Characteristic Curve (AUC) greater than 0.75 for the classification task were kept. This threshold was defined to exclude features that could degrade the model.
- To reduce the total number of features, hierarchical clustering was performed, keeping the minimum absolute Spearman's rank-order correlation between cluster members greater than 0.85. The feature with the greatest AUC of each cluster was finally selected.

Classification task was then achieved using a logistic regression model. Leave-One-Out Cross-Validation (LOOCV) was applied systematically to estimate the performance of each of the 16 models. Feature selection and standardization was performed inside each LOOCV fold, as described in [3]. All experiments have been achieved using radiomic features before the use of ComBat and after the use of ComBat. Finally, an additional prediction model (MMV) was defined as the majority voting of all the models in which each patient can participate.

TABLE III

TABLE SHOWING THE INFORMATION (MARKED BY 'X') USED FOR THE DESIGN OF EACH OF THE 16 MODELS, AND THE NUMBER OF PATIENTS THAT ARE AVAILABLE TO ESTIMATE EACH MODEL

	M01	M02	M03	M04	M05	M06	M07	M08	M09	M10	M11	M12	M13	M14	M15	M16
Clinical features	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
T1w features		x				x	x	x				x	x	x		x
T2w features			x			x			x	x		x	x		x	x
FLAIR features				x			x		x		x	x		x	x	x
T1c features					x			x		x	x		x	x	x	x
Nb of Patients	64	54	54	51	53	52	51	49	51	52	48	50	49	47	48	47
H3.1	17	13	13	12	14	12	13	12	12	14	12	12	13	12	12	12
H3.3	47	41	41	39	39	40	38	37	39	38	36	38	36	35	36	35

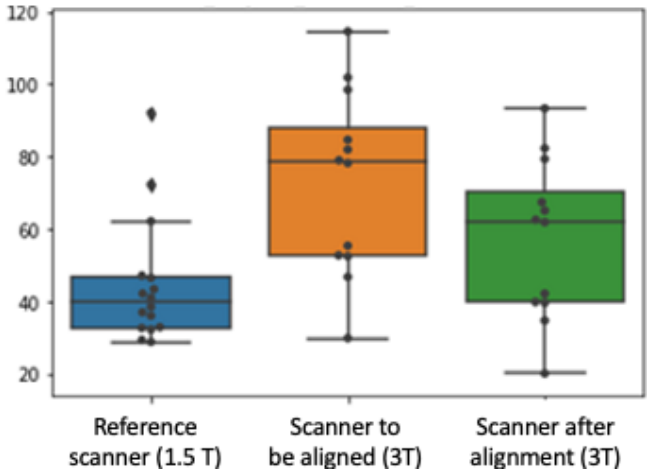

Fig. 1. Box plots of one radiomic feature (FLAIR 90 Percentile) for values coming from the 1.5 T scanner (left), the 3T scanner to be aligned (middle) and the 3T scanner after the realignment procedure using ComBat (right).

III. RESULTS

A. ComBat harmonization

Fig. 1 illustrates the values of one specific radiomic feature (90 Percentile computed in the tumor region of FLAIR image) for the patients acquired with 1.5 T and 3T scanners. After the realignment based on ComBat procedure, we observe a reduction of the values issued from the 3T scanner, which better fit with the values coming from the 1.5T scanner. The ComBat procedure was applied to each radiomic feature independently. Table IV provides the selected features by each of the 16 models without and after the realignment procedure. Following the feature selection, twelve parameters (out of the 318 tested) are involved in the building of the 32 models. Age is selected by all the models. After the realignment procedure based on ComBat, the 15 radiomic models (M02 to M16) are reduced to two features: age and one radiomic feature, this number of features being equal to 2 (7 models), 3 (2 models) and 4 (5 models) when ignoring the realignment procedure. Table V provides the F1 weighted score, obtained by the 16 models following LOOCV without and with realignment. The model showing the lowest performance (M01) is highlighted in red color and the model generating the highest performance (M02) is highlighted in blue color.

B. Prediction of mutation using the multi-model approach

Fig. 2 provides the prediction results obtained by the model showing the highest performance according to LOOCV (M02) and the majority voting process (MMV). Both models used radiomic features after their realignment using ComBat. Of note, for the model MMV, results are quite similar before and after ComBat, one H3.3 case that was misclassified before Combat was classified as undecided after ComBat, with an equal number of votes for each class. The number of undetermined cases is drastically reduced when using MMV : two patients are left undecided, with a equal number of votes for both tumor mutation whereas 10 patients could not be classified using M02 approach, due to missing data (lack of T1 modality in that case).

	M02	MMV
True prediction	43 (13 H3.1 ; 30 H3.3)	48 (15 H3.1 ; 33 H3.3)
False prediction	11 (0 H3.1 ; 11 H3.3)	14 (1 H3.1 ; 13 H3.3)
Undetermined cases	10 (4 H3.1 ; 6 H3.3)	2 (1 H3.1 ; 1 H3.3)

Fig. 2. Prediction results provided by the best radiomic model (M02) and by the majority voting approach (MMV).

IV. DISCUSSION

The presented approach makes it possible to adapt prediction of the H3 mutation to the real conditions of an examination and to propose models that adapt to the available data. Indeed, the prediction of the H3 mutation is of prime interest in cases where it is not possible to perform the biopsy or when its results are not conclusive. The different proposed models will be tested on additional data coming from a new clinical trial. As the number of subjects will increase, it will then be possible to refine the models by possibly incorporating more clues. The model M16 that we initiated in [3] incorporates patients with all the 4 MR modalities, in our augmented database, about 25% of the patients could not be analyzed using this model. Model M01 is based on age only and clearly shows worse performance than other models using LOOCV (see Table V). The development of the multi-model approach fills to main objectives, it avoids any imputation based methods for missing data handling along with the benefit of pooling in additional data if a modality is made available for the study.

The MR scanner affects the radiomic feature values extracted from MR images, introducing major confounding factors in

TABLE IV

FEATURES SELECTED BY EACH MODEL BEFORE (IN BLACK COLOR) AND AFTER (IN BLUE COLOR) REALIGNMENT PROCEDURE USING COMBAT FOR THE CLASSIFICATION TASK (H3.1 VERSUS H3.3 MUTATION).

	M01	M02	M03	M04	M05	M06	M07	M08	M09	M10	M11	M12	M13	M14	M15	M16
Clinical	Age	Age	Age	Age	Age	Age	Age	Age	Age	Age	Age	Age	Age	Age	Age	Age
T1w		Id1[1]				Id6	Id1	Id6				Id1	Id1	Id1		Id1
T2w			Id2													
FLAIR				Id3, Id4			Id3, Id4		Id3	Id3, Id4		Id3	Id3	Id3	Id3	Id3
T1c					Id5					Id5	Id5		Id5	Id5	Id5	Id5
Clinical	Age	Age	Age	Age	Age	Age	Age	Age	Age	Age	Age	Age	Age	Age	Age	Age
T1w		Id6				Id9	Id9	Id9				Id6	Id9	Id6		Id9
T2w			Id2						Id2							
FLAIR				Id7												
T1c					Id8					Id10	Id11				Id11	

[1]Id1: T1w firstorder median Id2: T2w glcm entropy Id3: FLAIR glcm homogeneity1 Id4: FLAIR firstorder 90Percentile Id5: T1c firstorder 10Percentile Id6: T1w firstorder root meansquared Id7: FLAIR glszm small area emphasis Id8: T1c glcm Idn Id9: T1w firstorder mean Id10: T1c glrlm short run emphasis Id11: T1c glcm Entropy

TABLE V

F1 WEIGHTED SCORE OBTAINED BY LOOCV FOR THE 16 MODELS WITHOUT AND WITH REALIGNEMENT BASED ON COMBAT

Model	without ComBat	with ComBat
M01	0.66	0.66
M02	0.88	0.85
M03	0.70	0.69
M04	0.75	0.68
M05	0.71	0.74
M06	0.87	0.78
M07	0.82	0.77
M08	0.84	0.82
M09	0.76	0.74
M010	0.70	0.75
M011	0.77	0.77
M012	0.82	0.75
M013	0.83	0.82
M014	0.81	0.78
M015	0.77	0.77
M016	0.81	0.78

multi-centric studies [4]. Here, we validated a harmonization procedure ComBat realignment for MR radiomic features extracted from different scanners. In patients scanned with 1.5T and 3T scanners, we showed that this harmonization procedure realigns radiomic feature distributions (Fig. 1). In radiomics especially in the light of oncological studies, pooling images acquired using different devices and different acquisition and reconstruction protocols is often needed to increase the size of cohort, or combining different cohorts. In that context, we demonstrated that ComBat could realign feature values so that all data could be analyzed together, even if images had been acquired with different scanners. It is important to highlight that the effects of ComBat on our prediction task are small in with respect to the final decisions. However, changes are seen in feature selection by individual models to decide upon the type of mutation as shown in the Table IV. Fewer features are selected after the feature were subjected to ComBat harmonization. This further reinforces the positive impact of ComBat as it seems to increase the level of robustness. This effect needs to be further investigated, increasing the number of patients with 3T scans.

V. CONCLUSION

The findings of this work demonstrated ComBat harmonization method could efficiently remove the scanner/protocol effect while preserving the individual variations in MR modalities coming from different patients and scanners. Furthermore, it allows the data set to stay unchanged without the need for adding artificially constructed data addressing missing data problem which is commonly used in medical imaging. This approach enables large MR multicentric studies to highlight the added value of radiomic analysis in features acquired from different scans. Furthermore, ComBat harmonization may display visible change in values and rather adds a level of robustness. The multi-model concept utilizes all the available data performs well due to the individual model prediction mechanism. Voting by each model could be associated to a level of confidence for each prediction.

REFERENCES

[1] Hargrave D, Bartels U, Bouffet E, Diffuse brainstem glioma in children: critical review of clinical trials, Lancet Oncol, 7:241–8, 2006. DOI: 10.1016/S1470-2045(06)70615-5

[2] Castel D, Philippe C, Calmon R et al, Histone H3F3A and HIST1H3B K27M mutations define two subgroups of diffuse intrinsic pontine gliomas with different prognosis and phenotypes, Acta neuropathologica, 130(6):815-827, 2015. DOI: 10.1007/s00401-015-1478-0

[3] Goya-Outi J, Calmon R, Orlhac F et al, Can Structural MRI Radiomics Predict DIPG Histone H3 Mutation and Patient Overall Survival at Diagnosis Time?, 2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI):1-4, 2019. DOI: 10.1109/BHI.2019.8834524

[4] Orlhac F, Lecler A, Savatovski J et al, How can we combat multi center variability in MR radiomics? Validation of a correction procedure, Eur Radiol, 31(4):2272-80, 2021. DOI: 10.1007/s00330-020-07284-9

[5] Goya-Outi J, Orlhac F, Calmon R et al, Computation of reliable textural indices from multimodal brain MRI: suggestions based on a study of patients with diffuse intrinsic pontine glioma, Phys Med Biol, 63(10):105003, 2018. DOI: 10.1088/1361-6560/aabd21

[6] van Griethuysen JJM, Fedorov A, Parmar C et al, Computational radiomics system to decode the radiographic phenotype, Cancer Res, 77(21):e104–7, 2017. DOI: 10.1158/0008-5472.CAN-17-0339