Physics and Imaging in Radiation Oncology 26 (2023) 100450

Contents lists available at ScienceDirect

Physics and Imaging in Radiation Oncology

journal homepage: www.sciencedirect.com/journal/physics-and-imaging-in-radiation-oncology

Original Research Article

# Multi-centre radiomics for prediction of recurrence following radical radiotherapy for head and neck cancers: Consequences of feature selection, machine learning classifiers and batch-effect harmonization

Amal Joseph Varghese [a], Varsha Gouthamchand [b], Balu Krishna Sasidharan [a], Leonard Wee [b], Sharief K Sidhique [a], Julia Priyadarshini Rao [a], Andre Dekker [b], Frank Hoebers [b], Devadhas Devakumar [c], Aparna Irodi [d], Timothy Peace Balasingh [a], Henry Finlay Godson [a], Joel T [a], Manu Mathew [a], Rajesh Gunasingam Isiah [a], Simon Pradeep Pavamani [a], Hannah Mary T Thomas [a,*]

[a] Department of Radiation Oncology, Christian Medical College, Vellore, Tamil Nadu, India
[b] Department of Radiation Oncology (Maastro), GROW School for Oncology and Reproduction, Maastricht University Medical Centre+, Maastricht, The Netherlands
[c] Department of Nuclear Medicine, Christian Medical College, Vellore, Tamil Nadu, India
[d] Department of Radiology, Christian Medical College, Vellore, Tamil Nadu, India

ARTICLE INFO

*Keywords:*
Head-and-neck cancer
Radiomics
Loco-regional recurrence
Multi-institutional
Prognosis
Machine learning

ABSTRACT

*Background and purpose:* Radiomics models trained with limited single institution data are often not reproducible and generalisable. We developed radiomics models that predict loco-regional recurrence within two years of radiotherapy with private and public datasets and their combinations, to simulate small and multi-institutional studies and study the responsiveness of the models to feature selection, machine learning algorithms, centre-effect harmonization and increased dataset sizes.

*Materials and methods:* 562 patients histologically confirmed and treated for locally advanced head-and-neck cancer (LA-HNC) from two public and two private datasets; one private dataset exclusively reserved for validation. Clinical contours of primary tumours were not recontoured and were used for Pyradiomics based feature extraction. ComBat harmonization was applied, and LASSO-Logistic Regression (LR) and Support Vector Machine (SVM) models were built. 95% confidence interval (CI) of 1000 bootstrapped area-under-the-Receiver-operating-curves (AUC) provided predictive performance. Responsiveness of the models' performance to the choice of feature selection methods, ComBat harmonization, machine learning classifier, single and pooled data was evaluated.

*Results:* LASSO and SelectKBest selected 14 and 16 features, respectively; three were overlapping. Without ComBat, the LR and SVM models for three institutional data showed AUCs (CI) of 0.513 (0.481–0.559) and 0.632 (0.586–0.665), respectively. Performances following ComBat revealed AUCs of 0.559 (0.536–0.590) and 0.662 (0.606–0.690), respectively. Compared to single cohort AUCs (0.562–0.629), SVM models from pooled data performed significantly better at AUC = 0.680.

*Conclusions:* Multi-institutional retrospective data accentuates the existing variabilities that affect radiomics. Carefully designed prospective, multi-institutional studies and data sharing are necessary for clinically relevant head-and-neck cancer prognostication models.

## 1. Introduction

Locoregional recurrence (LRR) is a highly prevalent pattern of relapse seen in about 20 to 50 % of patients with head-and-neck cancers (HNC) within two years after radiation treatment [1–3]. Therefore, the development of prognostic models to accurately identify patients who are at risk for LRR prior to radiotherapy would help the clinicians make better decisions to personalize treatment.

* Corresponding author at: Radiation Oncology, Unit 2, Ida B Scudder Cancer Centre, Christian Medical College, Vellore, Tamil Nadu 632004, India.
*E-mail address:* hannah.thomas@cmcvellore.ac.in (H.M.T. Thomas).

https://doi.org/10.1016/j.phro.2023.100450
Received 28 October 2022; Received in revised form 1 May 2023; Accepted 2 May 2023
Available online 16 May 2023

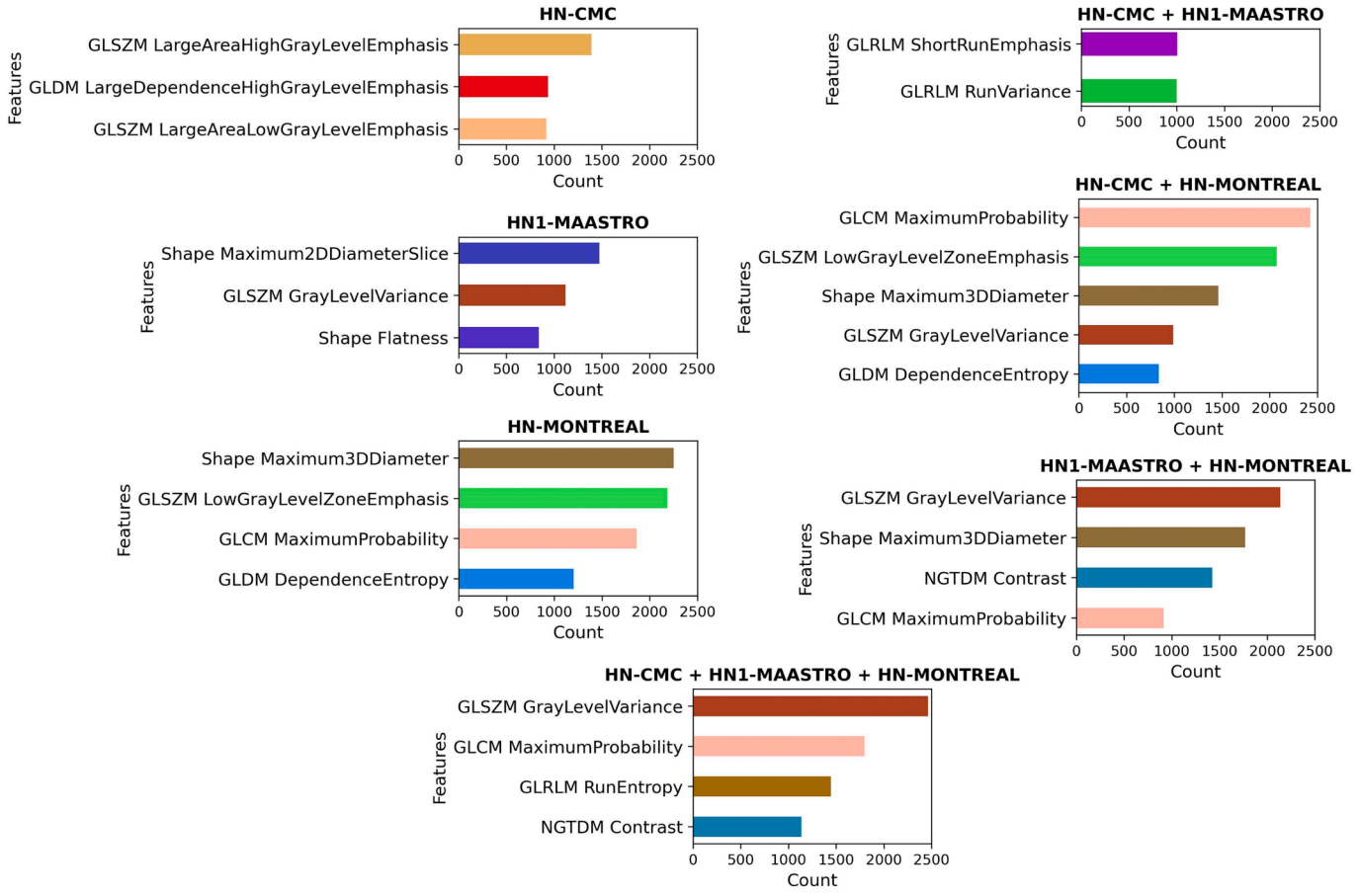


**Fig. 1.** Features selected by LASSO for the single and multi-institutional pooled datasets.

Radiomics, which is a machine learning driven quantitative analysis of medical images, has shown promising results in predicting risk of treatment outcomes in various cancers, including head-and-neck cancer [4–6]. However, radiomics extracts many features from the volume of interest leading to high dimensionality of the feature space [7] and attendant risk of over-fitting. Thus, feature selection is recommended to remove redundant and irrelevant features that do not contribute to the prognostic model. Therefore, selecting highly reproducible and stable features for building the model is important. In HNC, the optimal feature selection methods and classifiers remain to be studied for multicentre radiomics-based prognostic models [8–10].

Like all machine learning applications, radiomics based prognostic models benefit when trained and validated on as large image datasets as feasibly achievable. However, the burden for disease sites such as head-and-neck cancer are highly variable across the world. For example, in India and Netherlands, HNC accounts for about 20% and 3%, respectively of all cancer diagnoses [11,12]. So, it is quite difficult for single institutional data to account for the variabilities that negatively impacts the generalizability of models to real-life data such as the variations seen across populations, image quality and size of the dataset.

Often models trained on small sample sizes result in model over-fitting and lack generalizability [11,13–15] which makes pooling data from different institutions quite attractive. However, multi-institutional retrospective data will unavoidably introduce heterogeneity due to but not limited to differences in clinical subjects, variations in scanners, model versions, acquisition and reconstruction protocols and target definitions. Radiomic features are particularly sensitive to such variations, often referred to as the centre or site effect [16,17]. Centre effect harmonization is not trivial as different methods have shown satisfactory results in some studies [17,18] but has had no favourable effect in others [19]. The best batch effect removal approaches for radiomics have been reported [20]; however, batch assignment within these approaches can also contribute to further variability [21].

This study investigates the combination of publicly available datasets with single-institutional retrospective data to construct radiomic models for loco-regional recurrence within two years of treatment in head-and-neck cancer. We evaluated the generalizability of the radiomic models when applied to a new real-world dataset. This study was intended to show feasibility for a prospective radiomic study that is currently enrolling patients, and to identify the potential issues arising in multi-institutional modelling studies of this kind. Specifically, we have investigated the role of feature selection method, choice of machine learning architecture, sampling effects and batch harmonization effects with regards to external validation results.

## 2. Materials and methods

### 2.1. Data

Study included 562 patients treated for locally advanced head-and-neck cancer, had pre-treatment CT images and had follow-up for at least two years following radiation treatment. The images and the loco-regional recurrence data were from two public datasets and two private datasets. Description about the data is available Supplementary S.1.

The datasets HN-CMC, HN1-MAASTRO, and HN-MONTREAL were used for feature extraction and model training and HN3-MAASTRO dataset was reserved as the validation dataset for all experiments. The endpoint modelled is the loco-regional recurrence of HNC at two years and is measured from the start of radiotherapy treatment to date of recurrence. The number of events in each cohort is shown in

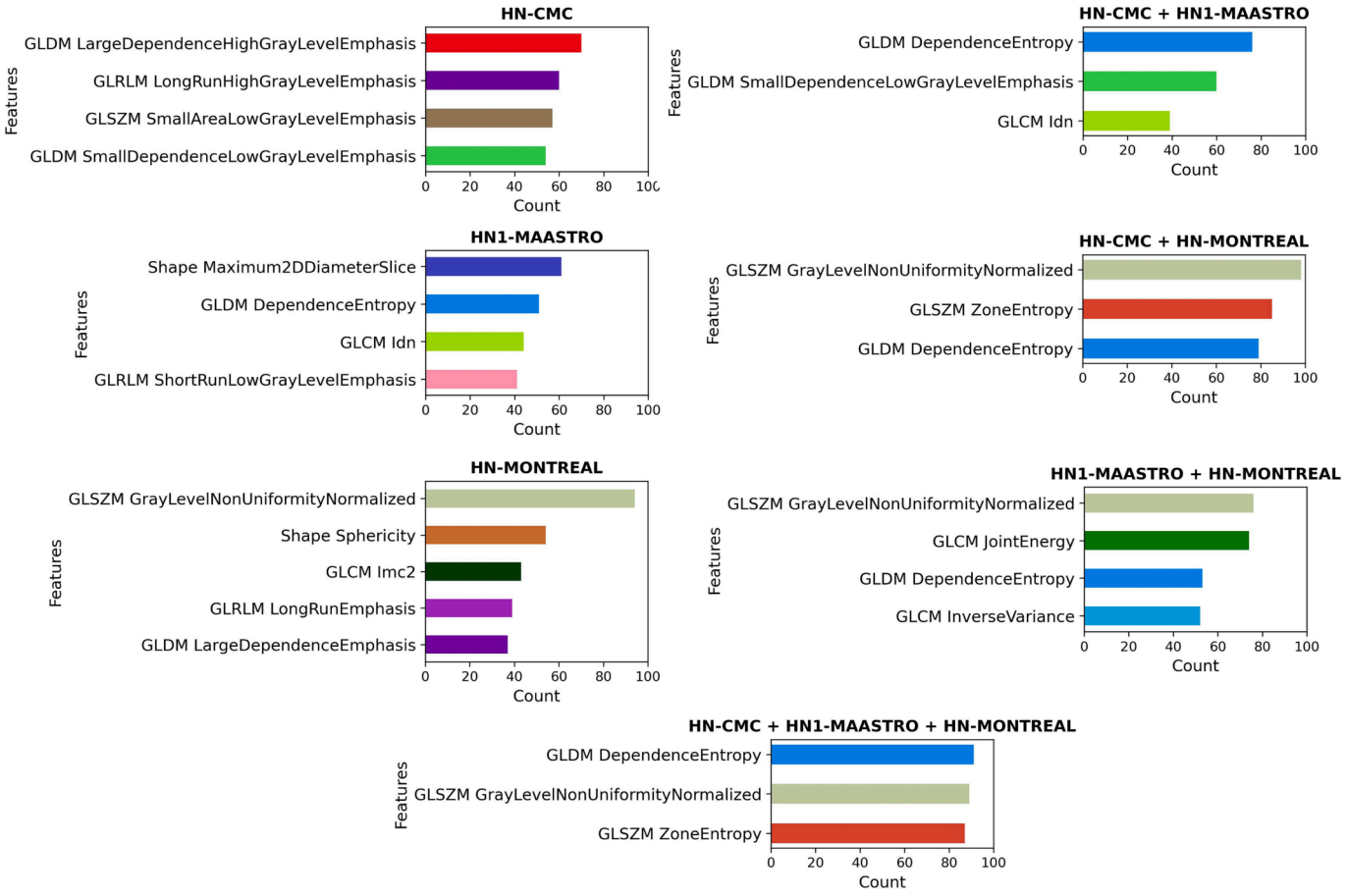


**Fig. 2.** Features selected by SelectKBest for the single and multi-institutional pooled datasets.

Supplementary Fig. S1. Ethical approval for the use of the private datasets was granted by the respective institutional Review Boards with consent waiver (CMC Vellore: IRB No.11640 and MAASTRO Ref. No. 0415).

### 2.2. Feature extraction

The radiomic features were extracted using PyRadiomics (Version v3.0.1), which is mostly compliant with the Image Biomarker Standardization Initiative (IBSI) [22], the deviations including Fixed bin Discretization is mentioned in their documentation [23]. A single parameter file was used for feature extraction (Supplementary Script S1) with further details about the extracted features in Supplementary S.3. The radiomics features for HN3-MAASTRO were extracted at MAASTRO and only the anonymized features were shared with CMC Vellore.

### 2.3. Experiment 1 (Feature selection)

Feature selection methods included Least absolute shrinkage and selection operator (LASSO) and SelectKBest (scikit-learn V1.0.2) and implementation details are in Supplementary S.4). For both methods, the features were selected within each dataset or across the pooled datasets included in the training cohort, depending on how the dataset was used in the different experiments. The validation cohort was unexposed to feature selection to avoid any data leakage.

### 2.4. Experiment 2 (Centre effect harmonization)

Out of 103 features, only the selected features from the four pooled datasets were compared to test the effect of ComBat harmonization. Publicly available PyComBat was used in this study [24] and default values included parametric estimation of batch effects with individual batch adjustments and without covariates, reference batch selection or precision computing. Before and after ComBat harmonization, the feature distribution and mean values from each of the pooled datasets and the AUC of the prognostic models were evaluated.

### 2.5. Prognostic modelling

HN-CMC, HN1-MAASTRO, and HN-MONTREAL were used for training and HN3-MAASTRO was the validation dataset. The synthetic minority over-sampling technique (SMOTE) algorithm was used to oversample the minority class to account for the unbalanced data in the 'LRR-positive' and 'LRR-negative' and produce class-balanced training datasets before training the models.

#### 2.5.1. Experiment 3 (machine learning classifier)

Two frequently used supervised machine-learning based algorithms, namely Logistic Regression which is a linear method and Support Vector Machine with a non-linear kernel, were used to build models for predicting LRR. The primary measure of the models' performance was evaluated based on the AUC of the hold-out validation dataset (HN3-MAASTRO).

### 2.6. Experiment 4 (single vs pooled datasets)

Given the multi-centric nature of studied datasets, we chose to study the effect of more data on the performance of the models. The LR and SVM models were trained with a single institutional data and performance was compared with models trained with pooled data and

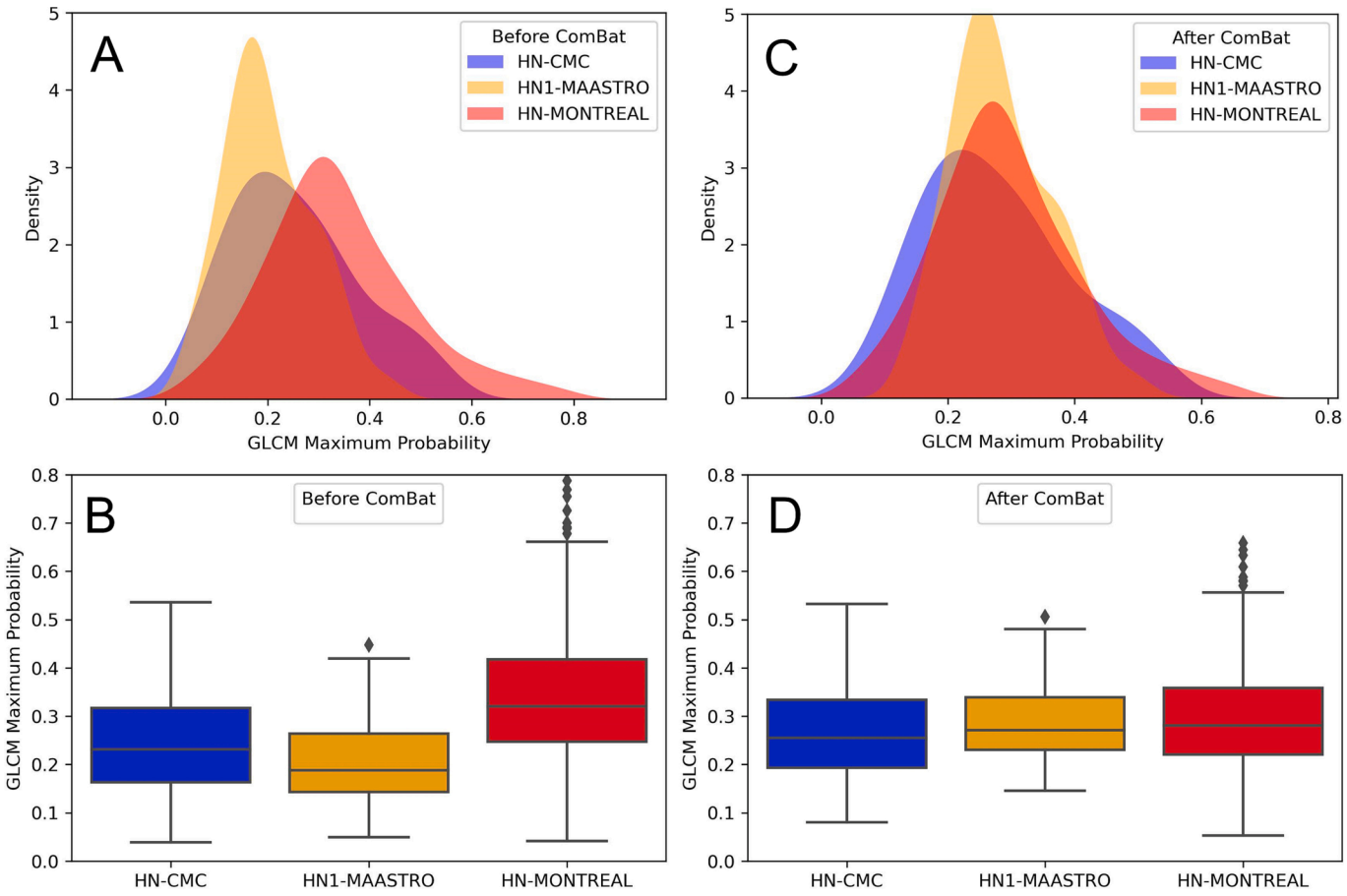


**Fig. 3.** The Kernel Density Estimate (KDE) (A and C) and Box (B and D) plots for one representative feature (GLCM Maximum Probability) before (A and B) and after (C and D) ComBat harmonization.

validated on HN3-MAASTRO.

### 2.7. Statistical analysis

Descriptive statistics were used to summarize the salient features selected. Following feature selection using either LASSO or SelectKBest, Spearman Rank Correlation was performed to eliminate any mutually correlated features; rank correlation for feature being taken forward to model building was set at rho $< |0.7|$. A two-sample Kolmogorov-Smirnov test was performed to compare the distributions and Welch's *t*-test was performed to compare the means of selected features between the individual datasets in the pooled dataset before and after ComBat. The models' performance was evaluated using Area under the Receiver-operating-curve (AUC) and median AUC over 1000 bootstrapping models was reported with 95% confidence intervals (CI). All statistical analyses were done using Python (Python Version 3.7.11 and Scikit-learn Version 1.0.2).

## 3. Results

The demographics of the patients in the training datasets were comparable (Supplementary Table S1).

### 3.1. Experiment 1

Figs. 1 and 2 show that different prognostic features were selected by LASSO and SelectKBest, respectively for each dataset independently in the training data. When the datasets were pooled, some features from the independent datasets were selected but in varying order of importance and frequency of occurrence, along with features that were unique to the pooled dataset and not present in the independent datasets.

### 3.2. Experiment 2

The distributions of the selected features were evaluated before and after ComBat harmonization. Kolmogorov-Smirnov 2 Sample test showed that Combat harmonisation made the distributions significant less different. In most of the selected features, distribution was no longer detected following harmonization (e.g., GLCM Maximum Probability p-val changed from 0.022 to 0.197), while few features either became significantly different or remained unchanged (See Supplementary Table S5).

In Fig. 3, the distribution and the variations of one representative feature namely, GLCMMaximumProbability is shown using a Kernel Density estimate plot (A&C) and Box Whisker plots (B&D). This change in distribution was not noted as drastic in other features (e.g., NGTDMContrast and GLSZMGrayLevelVariance) (Supplementary Table S5).

In Fig. 4 and Supplementary Table S3 the model performance based on AUC with 95% CI before and after ComBat harmonization shows variability and did not necessarily improve the performance of the models.

### 3.3. Experiment 3

Fig. 4 and Supplementary Table S3 show the variability in the models' performance between the ML algorithms with SVM always

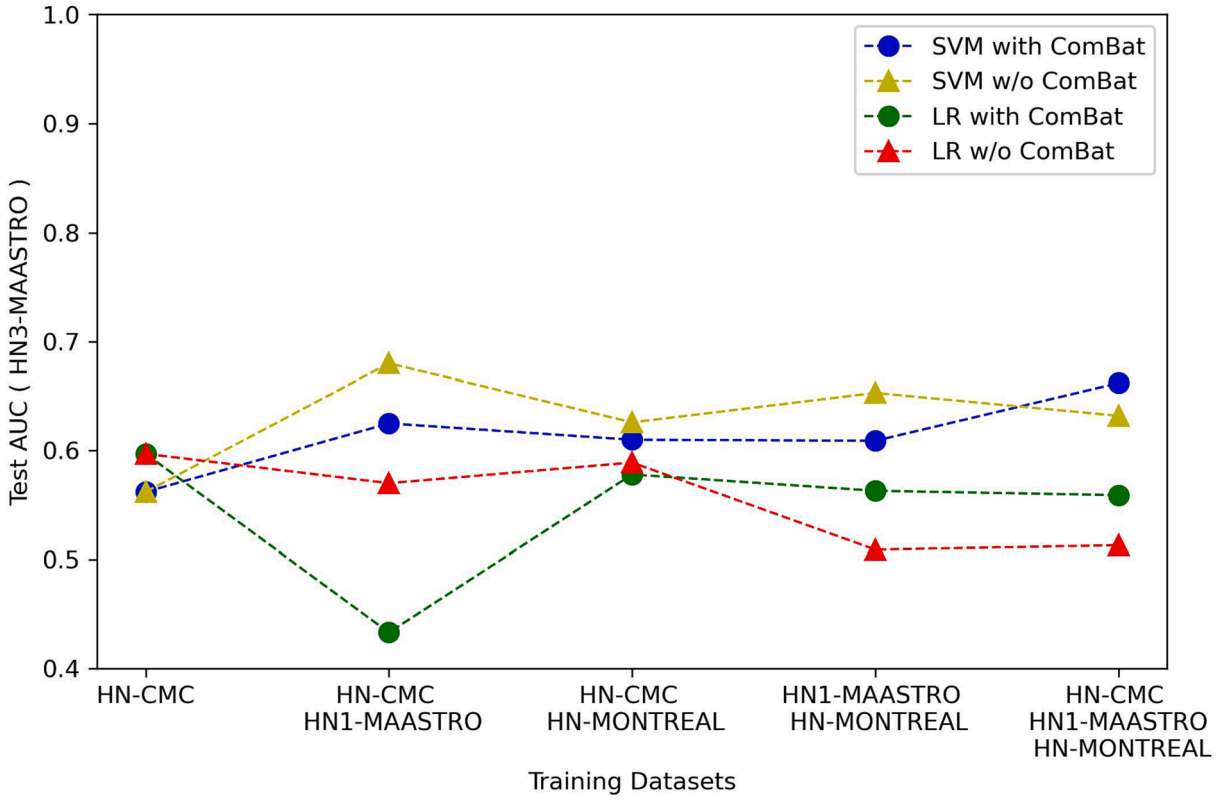


**Fig. 4.** Performance of the LR and SVM models trained on pooled datasets prior to and post ComBat harmonization. Model performance is reported on validation data HN3-MAASTRO.

outperforming the LR models, when pooled datasets were used. The two best performing models were SVM models trained on HN_MAASTRO and HN-CMC + HN-MAASTRO (AUC = 0.63) and HN-CMC + HN1-MAASTRO + HN-MONTREAL (AUC = 0.66) datasets, respectively.

### 3.4. Experiment 4

Fig. 5 and Supplementary Table S4 show the effect of size of a dataset on the performance of the models for a representative single dataset (HN-CMC) against pooled data. Fig. 5A shows that Logistic Regression models do not show significant differences in model performance between single (red curve AUC = 0.60) and all the pooled data (orange curve AUC = 0.56). Fig. 5B shows that SVM models trained better with the added data and the performance improved significantly from AUC = 0.56 (red curve) to AUC = 0.66 (orange curve). This trend was seen for all datasets independently or pooled in some combination (Supplementary Table S4). For all the combinations of pooled data studied, SVM model performance plateaus at AUC~= 0.66, regardless of more data being added.

## 4. Discussion

In this study we simulated a multi-institutional study to build prognostic models of loco-regional recurrence (LRR) in locally advanced head-and-neck cancer (LA-HNC) patients. We designed experiments to study the responsiveness of the models to the choice of feature extraction, machine learning classifiers, batch effect normalization and data size.

This study is important as radiomics has shown potential to personalize patient treatment using routinely acquired clinical images. This is particularly important in clinical management of HNC, since biological heterogeneity inside a tumour that characterizes the inter-patient differences is just as important as heterogeneity or variability seen in terms of determining the clinical outcome [1,9]. CT imaging was opted as it is an indispensable imaging in management of HNC and all patients treated with radiation will have it.

Most radiomics studies including this study rely on retrospective imaging which lends to an unavoidable bias in patient selection and data heterogeneity from scanners, imaging and reconstruction parameters, inter-observer variations in delineations, biological variability across populations etc. In real-life data, these variations manifest more prominently than prospective data. In a multi-centric setting, ensuring the quality of the data can be additionally challenging. For example, we often rely on what is defined as ROI. However, there could be differences in GTV definitions between centres where some include just the primary tumour and others both primary and nodes. In being too conservative, we risk losing a lot of data, and the contrary introduces noise leading to a trade-off between size of the dataset and noise. Also, from the images or their metadata it is often difficult to determine if the head-and-neck CT was imaged with or without contrast. With prospective imaging we might be able to limit some of these variations, but it would be at cost of much smaller samples in exchange for better control over the quality of the data.

Currently, there are many feature selection methods available. However, there is no optimal feature selection method for radiomics yet. In this study, we have chosen 1) LASSO and 2) SelectKBest based on their popularity, and ease of implementation. It was observed that features from a single dataset have quite different distributions (Figs. 1 and 2), may not be applicable in multi-institutional dataset [25] and are dependent on the choice of the selection method [26]. We also observed only three features were common among the two methods (Supplementary Table S2).

Radiomic models often run the risk of overfitting (failure to predict in unseen data) and one way to mitigate is to increase the size of the training sample, which is achieved in this study by pooling data from different institutions/sources which also introduces more heterogeneity.

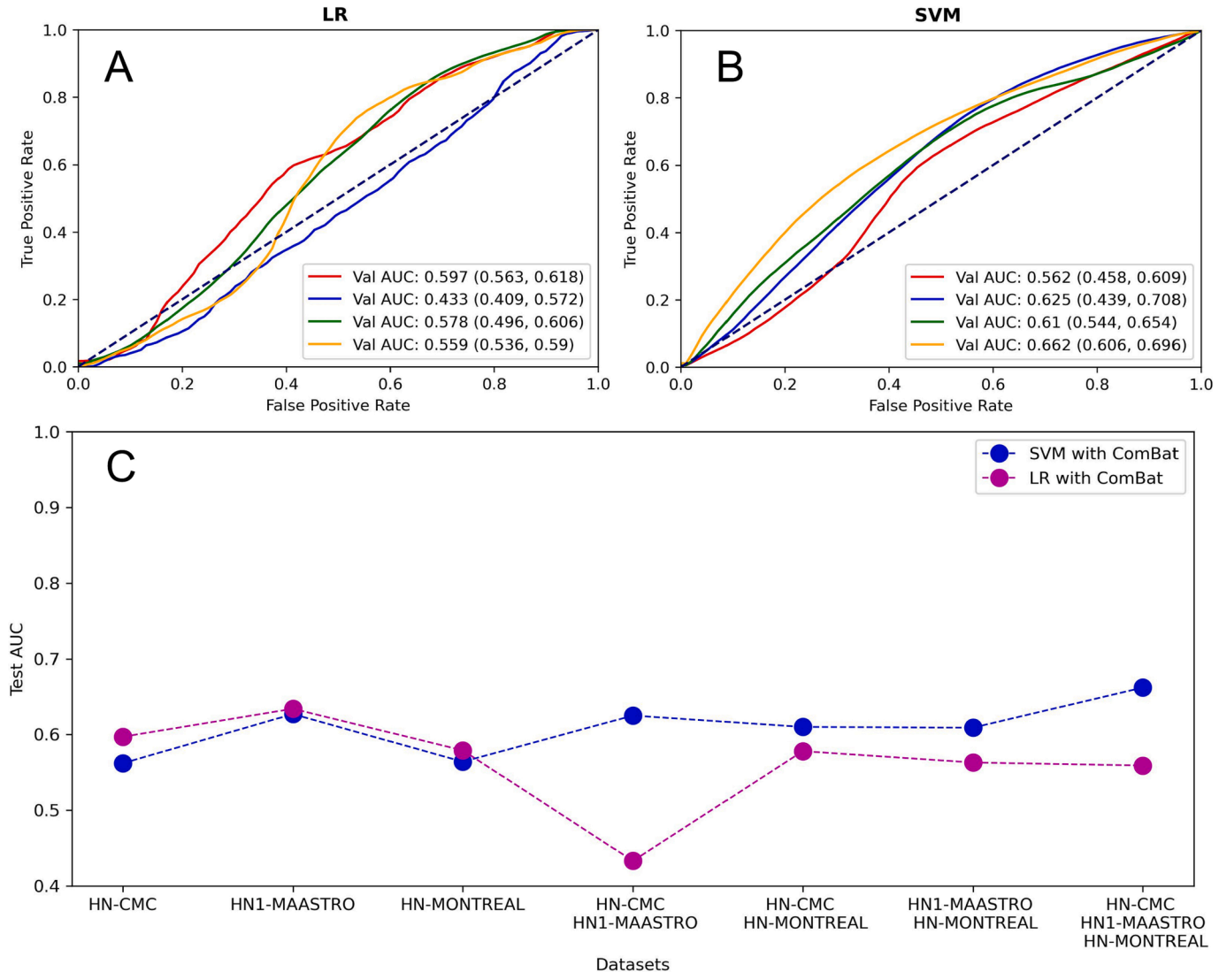


**Fig. 5.** Performance of the models trained with data from single institution versus multi-institutional pooled data. Validation ROC of Logistic Regression (A) and SVM LRR models (B) for an example single dataset (HN-CMC) and its pooled dataset combinations. The ROCs correspond to the HN-CMC (red), HN-CMC + HN1-MAASTRO (blue), HN-CMC + HN-MONTREAL (green) and HN-CMC + HN1-MAASTRO + HN-MONTREAL datasets (orange), respectively. C) Test AUC across all single and pooled datasets in this experiment. Validation data was HN3-MAASTRO. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

We chose ComBat as it is the most popular method reported for Radiomics studies with an easy Python implementation tool available [24]. We saw that ComBat standardized the distributions and the feature mean (Supplementary Table S5), but it leaves us no clearer as to what the cause(s) of the batch effects might be. The danger is that we wipe out real clinical demographic heterogeneities in addition to scanner/acquisition/delineation types of heterogeneities which may lead to potentially dangerous misjudgment. Our study shows that applying ComBat has either similar or better performance in the validation set compared to results without the batch correction suggesting that ComBat has more utility in some combination of multi-centre datasets but less in others. We do not at present have access to enough data to extricate the reasons for this, and we mention this as an important question for follow up studies.

We have not exhaustively searched over all possible classifiers for the best classifier. For the feature selection and classifier methods used, we considered the responsiveness on the hyper-parameters, but this was not the primary purpose of our study. We cannot exclusively say that the hyper-parameters have been optimized, but our results represent what could be seen in future multi-centre studies. We selected Logistic Regression and Radial SVM as being two of the most used classifiers [11,14,15] and their selection motivated differently; LR has linear decision boundary and models are simpler to explain compared to SVM which operates on non-linear boundaries. With the multi-institutional data, we found LR model performed poorly on the validation data, despite harmonization with ComBat, while SVM models outperformed the LR models, irrespective of the ComBat harmonization. Hence, while choosing a ML framework, that it should be considered not only for complexity e.g., linear or non-linear decision boundaries, but also on how well does the framework deals with the different sources of heterogeneity. Although more data yielded significantly better performance for both LR and SVM models compared to single institutional data (Supplementary Table S4), we observed that once the SVM models had sufficiently learnt the heterogeneity, more data did not necessarily improve the performance. So, AUCs achieved with approximately 200 patients and about 500 patients remained at 0.6. This may be an inherent limitation of AUC, where it typically limits it to under 0.8 in small samples as shown by Bahn and Alber [27]. In multi-institutional studies a trade-off in capturing the true biological heterogeneity and achieving a sample size that can statistically account for the true

differences is unavoidable.

There are certain limitations in this study. First, is the retrospective nature of the data. Secondly, the time-to-LRR was dichotomised since one cohort (HN-CMC) did not have event dates for performing the time-to-event analysis (See logic in Supplementary Section S.3). For the prospective observational trial that the institution is accruing currently, care is taken to ensure the follow-up and time to clinical outcomes including loco-regional recurrence are recorded accurately. Thirdly, the heterogeneity of imaging data limits the generalizability of the prognostic model which includes the patients with good prognosis (HPV-positive oropharynx) and poor prognosis (e.g., Stage 4 hypopharynx). Next, there was some discrepancy noted in the GTV definition in the public datasets. However, we have included the patients solely based on the tumour definitions provided by individual centres. HN-MONTREAL dataset included patients treated for nasopharynx and unknown primary. We did not actively try to remove cancer of the nasopharynx as this cancer is more prevalent in some parts of the world relatively more than others, which includes North-East India [28]. However, it could not be captured in the HN-CMC cohort. Similarly, on examining the unknown primary images, at least three were oropharynx, however they were included and reported based on the supporting document that mentioned the presence of a GTV primary. This might have some bearing on the performance of the models. However, no corrections were made and were retained in terms of simulating the real clinical scenario. Next, contrast enhanced CT are not always standard imaging available for head-and-neck cancers. Hence, the datasets included both intravenously injected contrast CTs and non-contrast enhanced CTs and no correction was applied since the effect of contrast on radiomics features in Head-and-neck tumours is not yet fully explored. Next, although PyRadiomics did not exactly comply completely with all IBSI requirements, this would not have affected the present study since all datasets were computed using the same PyRadiomics software and the same feature extraction parameter setting. However, IBSI compliance is needed to allow better reproduction and validation of the results externally. Lastly, both LR and SVM models trained on different combinations of datasets showed decreased performance when validated in the HN3 dataset (Supplementary Table S3). Future studies will look at effect of combinations of datasets instead of just adding more data (Similar to Supplementary Table S4).

Given our results and the growing number of studies on deep learning based oncological prognostication [29,30] and the ability of these models to handle the heterogeneity in the data better compared to machine learning, it would be worth exploring their utility for multi-institutional studies. It would also be interesting to see if with federated architecture [31], where we can leave some centres in training and keep others out for external validation, and easily try different combinations of data, it would be helpful in teasing out the clinical-related heterogeneities and correcting only for the scanner-related heterogeneities.

In summary, the study highlights the variability that occurs when multi-institutional data is pooled for prognostic radiomics models for head-and-neck cancer. Based on our observations, we strongly recommend that future studies mention the scanner models, imaging parameters, use of contrast agents and provide primary and nodal volumes separately along with the other clinical details relevant to the patients. Harmonization techniques may help reduce some variability; however, we are unclear if we are losing key heterogeneity that may be worth preserving. Carefully designed prospective, multi-institutional studies and data sharing will be needed to build clinically relevant radiomics models for prognostication.

## 5. Patient consent

The use of patient data from the private datasets was approved by the respective Institutional Review Board with consent waiver.

HN-CMC; CMC Vellore (IRB No. 11640).

HN3-MAASTRO; MAASTRO IRB (ref number 0415).

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

This work was supported by the DBT/Wellcome Trust India Alliance Early Career Fellowship [Grant number: IA/E/18/1/504306] awarded to HMT. Author BS acknowledges the support by the Foundation I-DAIR. Authors LW and FH acknowledge support by the Hanarth Foundation. LW and AD further acknowledge financial support from the Dutch Research Council (NWO) via the BIONIC, TRAIN and AMICUS grants.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.phro.2023.100450.

## References

[1] Chang JH, Wu Y, Wu ATH. Locoregionally recurrent head and neck squamous cell carcinoma: incidence, survival, prognostic factors, and treatment outcomes. Oncotarget 2017;8. https://doi.org/10.18632/oncotarget.16340. 55600–12.
[2] Alsahafi E, Begg K, Amelio I, Raulf N, Lucarelli P, Sauter T, et al. Clinical update on head and neck cancer: molecular biology and ongoing challenges. Cell Death Dis 2019;10:1–17. https://doi.org/10.1038/s41419-019-1769-9.
[3] Massa ST, Osazuwa-Peters N, Christopher KM, Arnold LD, Schootman M, Walker RJ, et al. Competing causes of death in the head and neck cancer population. Oral Oncol 2017;65:8–15. https://doi.org/10.1016/j.oraloncology.2016.12.006.
[4] Elhalawani H, Mohamed AS, Mulder S, Grossberg A, Smith KE, Gunn GB, et al. Radiomics prediction of radiation treatment outcomes in oropharyngeal cancer: a clinical and image repository in concert with the cancer imaging archive (TCIA). Int J Radiat Oncol Biol Phys 2018;102:e215–6. https://doi.org/10.1016/j.ijrobp.2018.07.748.
[5] Kalendralis P, Shi Z, Traverso A, Choudhury A, Sloep M, Zhovannik I, et al. FAIR-compliant clinical, radiomics and DICOM metadata of RIDER, interobserver, Lung1 and head-Neck1 TCIA collections. Med Phys 2020;47:5931–40. https://doi.org/10.1002/mp.14322.
[6] Vallières M, Kay Rivest E, Perrin LJ, Liem X, Furstoss C, Aerts HJWL, et al. Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. Sci Rep 2017;7:10117. https://doi.org/10.1038/s41598-017-10371-5.
[7] van Timmeren JE, Cester D, Tanadini-Lang S, Alkadhi H, Baessler B. Radiomics in medical imaging—"how-to" guide and critical reflection. Insights Imaging 2020;11:91. https://doi.org/10.1186/s13244-020-00887-2.
[8] Gangil T, Shahabuddin AB, Dinesh Rao B, Palanisamy K, Chakrabarti B, Sharan K. Predicting clinical outcomes of radiotherapy for head and neck squamous cell carcinoma patients using machine learning algorithms. J Big Data 2022;9:25. https://doi.org/10.1186/s40537-022-00578-3.
[9] Wong AJ, Kanwar A, Mohamed AS, Fuller CD. Radiomics in head and neck cancer: from exploration to application. Transl Cancer Res 2016;5:371–82. https://doi.org/10.21037/tcr.2016.07.18.
[10] Parmar C, Leijenaar RTH, Grossmann P, Rios Velazquez E, Bussink J, Rietveld D, et al. Radiomic feature clusters and prognostic signatures specific for Lung and Head & Neck cancer. Sci Rep 2015;5:11044. https://doi.org/10.1038/srep11044.
[11] Francis D. Trends in incidence of head and neck cancers in India. Eur J Cancer 2018;92:S23. https://doi.org/10.1016/j.ejca.2018.01.056.
[12] 528-the-netherlands-fact-sheets.pdf n.d. https://gco.iarc.fr/today/data/factsheets/populations/528-the-netherlands-fact-sheets.pdf (accessed August 22, 2022).
[13] Devakumar D, Sunny G, Balu K, Bowen SR, Nadaraj A, Jeyseelan L, et al. Framework for machine learning of CT and PET radiomics to predict local failure after radiotherapy in locally advanced head and neck cancers. J Med Phys 2021;46:181. https://doi.org/10.4103/jmp.JMP_6_21.
[14] Giraud P, Giraud P, Gasnier A, El Ayachy R, Kreps S, Foy JP, et al. Radiomics and machine learning for radiotherapy in head and neck cancers. Front Oncol 2019;9:174. https://doi.org/10.3389/fonc.2019.00174.
[15] Ger RB, Zhou S, Elgohari B, Elhalawani H, Mackin DM, Meier JG, et al. Radiomics features of the primary tumor fail to improve prediction of overall survival in large cohorts of CT- and PET-imaged head and neck cancer patients. PLoS One 2019;14:e0222509.
[16] Zhao B. Understanding sources of variation to improve the reproducibility of radiomics. Front Oncol 2021;11:633176. https://doi.org/10.3389/fonc.2021.633176.

[17] Orlhac F, Frouin F, Nioche C, Ayache N, Buvat I. Validation of a method to compensate multicenter effects affecting CT radiomics. Radiology 2019;291:53–9. https://doi.org/10.1148/radiol.2019182023.
[18] Orlhac F, Eertink JJ, Cottereau AS, Zijlstra JM, Thieblemont C, Meignan M, et al. A guide to ComBat harmonization of imaging biomarkers in multicenter studies. J Nucl Med 2022;63:172–9. https://doi.org/10.2967/jnumed.121.262464.
[19] Da-ano R, Lucia F, Masson I, Abgral R, Alfieri J, Rousseau C, et al. A transfer learning approach to facilitate ComBat-based harmonization of multicentre radiomic features in new datasets. PLoS One 2021;16:e0253653.
[20] Castaldo R, Brancato V, Cavaliere C, Trama F, Illiano E, Costantini E, et al. A framework of analysis to facilitate the harmonization of multicenter radiomic features in prostate cancer. J Clin Med 2023;12:140. https://doi.org/10.3390/jcm12010140.
[21] Carré A, Battistella E, Niyoteka S, Sun R, Deutsch E, Robert C. AutoComBat: a generic method for harmonizing MRI-based radiomic features. Sci Rep 2022;12:12762. https://doi.org/10.1038/s41598-022-16609-1.
[22] van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational radiomics system to decode the radiographic phenotype. Cancer Res 2017;77:e104–7. https://doi.org/10.1158/0008-5472.CAN-17-0339.
[23] Welcome to pyradiomics documentation! — pyradiomics v3.0.1.post15+g2791e23 documentation n.d. https://pyradiomics.readthedocs.io/en/latest/ (accessed January 27, 2023).
[24] Behdenna A, Haziza J, Azencott CA, Nordor A. pyComBat, a Python tool for batch effects correction in high-throughput molecular data using empirical Bayes methods. Bioinformatics 2020. https://doi.org/10.1101/2020.03.17.995431.
[25] Demircioğlu A. Benchmarking feature selection methods in radiomics. Invest Radiol 2022;57:433–43. https://doi.org/10.1097/RLI.0000000000000855.
[26] Demircioğlu A. Evaluation of the dependence of radiomic features on the machine learning model. Insights Imag 2022;13:28. https://doi.org/10.1186/s13244-022-01170-2.
[27] Bahn E, Alber M. On the limitations of the area under the ROC curve for NTCP modelling. Radiother Oncol 2020;144:148–51. https://doi.org/10.1016/j.radonc.2019.11.018.
[28] Chang ET, Ye W, Zeng YX, Adami HO. The evolving epidemiology of nasopharyngeal carcinoma. Cancer Epidemiol Biomarkers Prev 2021;30:1035–47. https://doi.org/10.1158/1055-9965.EPI-20-1702.
[29] Coccia M. Deep learning technology for improving cancer care in society: new directions in cancer imaging driven by artificial intelligence. Technol Soc 2020;60:101198. https://doi.org/10.1016/j.techsoc.2019.101198.
[30] Diamant A, Chatterjee A, Vallières M, Shenouda G, Seuntjens J. Deep learning in head & neck cancer outcome prediction. Sci Rep 2019;9:2764. https://doi.org/10.1038/s41598-019-39206-1.
[31] Dekker A. Personal Health Train for Radiation Oncology in India and The Netherlands. clinicaltrials.gov; 2020.