Maastricht University

User-controlled pipelines for feature integration and head and neck radiation therapy outcome predictions

Citation for published version (APA):

Welch, M. L., McIntosh, C., McNiven, A., Huang, S. H., Zhang, B.-B., Wee, L., Traverso, A., O'Sullivan, B., Hoebers, F., Dekker, A., & Jaffray, D. A. (2020). User-controlled pipelines for feature integration and head and neck radiation therapy outcome predictions. *Physica Medica: European journal of medical physics*, *70*, 145-152. https://doi.org/10.1016/j.ejmp.2020.01.027

Document status and date:
Published: 01/02/2020

DOI:
10.1016/j.ejmp.2020.01.027

Document Version:
Publisher's PDF, also known as Version of record

Document license:
Taverne

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.umlib.nl/taverne-license

Take down policy
If you believe that this document breaches copyright please contact us at:

repository@maastrichtuniversity.nl

providing details and we will investigate your claim.

Download date: 28 Apr. 2024

Physica Medica 70 (2020) 145–152

Contents lists available at ScienceDirect

Physica Medica

journal homepage: www.elsevier.com/locate/ejmp

Original paper

User-controlled pipelines for feature integration and head and neck radiation therapy outcome predictions

Mattea L. Welch[a,d,e,f,*], Chris McIntosh[a,e,f,g,h,i], Andrea McNiven[b,e], Shao Hui Huang[b,e], Bei-Bei Zhang[b,e], Leonard Wee[d], Alberto Traverso[d,e], Brian O'Sullivan[b,e], Frank Hoebers[d], Andre Dekker[d], David A. Jaffray[a,b,c,e,f]

[a] Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada
[b] Department of Radiation Oncology, University of Toronto, Toronto, Ontario, Canada
[c] IBBME, University of Toronto, Toronto, Ontario, Canada
[d] Department of Radiation Oncology (MAASTRO), GROW School for Oncology and Developmental Biology, Maastricht University Medical Centre, Maastricht, The Netherlands
[e] Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada
[f] The Techna Institute for the Advancement of Technology for Health, Toronto, Ontario, Canada
[g] Vector Institute, Toronto, Ontario, Canada
[h] Peter Munk Cardiac Centre, University Health Network, Toronto, Ontario, Canada
[i] The Joint Department of Medical Imaging, University Health Network, Toronto, Ontario, Canada

ARTICLE INFO

Keywords:
Head and neck
Outcome prediction
Deep learning
Machine learning
User-controlled
Bias

ABSTRACT

Purpose: Precision cancer medicine is dependent on accurate prediction of disease and treatment outcome, requiring integration of clinical, imaging and interventional knowledge. User controlled pipelines are capable of feature integration with varied levels of human interaction. In this work we present two pipelines designed to combine clinical, radiomic (quantified imaging), and RTx-omic (quantified radiation therapy (RT) plan) information for prediction of locoregional failure (LRF) in head and neck cancer (H&N).
Methods: Pipelines were designed to extract information and model patient outcomes based on clinical features, computed tomography (CT) imaging, and planned RT dose volumes. We predict H&N LRF using: 1) a highly user-driven pipeline that leverages modular design and machine learning for feature extraction and model development; and 2) a pipeline with minimal user input that utilizes deep learning convolutional neural networks to extract and combine CT imaging, RT dose and clinical features for model development.
Results: Clinical features with logistic regression in our highly user-driven pipeline had the highest precision recall area under the curve (PR-AUC) of 0.66 (0.33–0.93), where a PR-AUC = 0.11 is considered random. CONCLUSIONS: Our work demonstrates the potential to aggregate features from multiple specialties for conditional-outcome predictions using pipelines with varied levels of human interaction. Most importantly, our results provide insights into the importance of data curation and quality, as well as user, data and methodology bias awareness as it pertains to result interpretation in user controlled pipelines.

1. Introduction

Prognostics are an important part of cancer care [1,2] that estimates the risk of an individual's outcome based on multiple variables (e.g. tumour, patient and environmental). It aids in treatment decisions and differs from aetiological research where the goal is to explain whether an outcome can be attributed to a specific risk factor [3]. In addition to the traditional prognostic factors mentioned above, integration of treatment information to form a treatment specific-conditional prognosis is highly beneficial.

We define treatment specific-conditional prognosis as the prediction of a treatment's effect, if administered as intended, on the patient's outcome [4]. The volume and variety of features available for inclusion in these types of predictions is expanding rapidly. This is in part the result of a hypothesis in cancer management that by analyzing an extensive set of features that encompass the nuances of disease processes and treatments that we can achieve "Precision Medicine" [5–7]. Features can vary from highly cited and tested measurements designed to

* Corresponding author.
E-mail address: mattea.welch@rmp.uhn.ca (M.L. Welch).

https://doi.org/10.1016/j.ejmp.2020.01.027
Received 16 September 2019; Received in revised form 23 December 2019; Accepted 28 January 2020
Available online 02 February 2020
1120-1797/ © 2020 Associazione Italiana di Fisica Medica. Published by Elsevier Ltd. All rights reserved.

probe and describe the nuances of both a patient and corresponding disease [8–11], to more experimental imaging, tissue and treatment features that describe the activity of tumours before and during treatment [12]. These features can even include those generated through automation that explore disease outcome correlations with image signal values (i.e. radiomics) [13–15]. Quantified interventional features have the potential to be combined with these features for a truly comprehensive view of a patient's treatment-influenced course of disease.

In head and neck (H&N) radiation therapy (RT) it is known that the dose fractionation and quality of an RT plan can impact overall and locoregional failure (LRF) free survival [16–18]. Dose volume histograms (DVH) are calculated to evaluate a RT plan based on RT dose delivered to volumes of tissue [19,20]. Metrics calculated using the DVHs are known predictors of both toxicity and outcome [21], but lack spatial dose information that is indicative of a patient's disease and surrounding intrinsic anatomical variations. Recent research quantifying spatial dose distributions for patients has found utility in toxicity prediction [22,23] and may similarly benefit treatment-specific conditional prognosis outcome prediction. However, as the number of prognostic factors that we consider increases, our methods for knowledge integration must change.

The agglomeration of diverse features represents a movement towards precision medicine, but also the utilization of big data in cancer care [5,6,24]. Approaches and pipelines for big data feature integration would provide flexible solutions that could drive data exploration and clinical decision support systems forward; an idea that was demonstrated by Mobadersany et al. [25] who combined deep learning and traditional user defined features. Additionally, these 'big machines' can be thought of as user-controlled pipelines requiring a spectrum of user interaction and assurance while evaluating intermediate byproducts and tuning various operating parameters.

In this work, we build two generalized feature integration pipelines for cancer treatment specific conditional prognosis; one of which is a highly user-driven process, while the other is substantially automated. Both pipelines leverage clinical, radiomic, and interventional features extracted from personalized RT plans (henceforth referred to as RTx-omic features). As a proof of concept, we applied our pipelines to a H&N dataset to determine how the conditional-prognostic performance of clinical features may be impacted by RTx-omic and radiomic features during LRF prediction, and whether conclusions could be drawn regarding the influence of user bias on user-controlled pipelines.

2. Methods

Our methods are designed to build and explore two pipelines for patient information integration and outcome prediction: 1) A machine learning pipeline that is inherently user-driven. Features are explicitly defined and informed by prior-knowledge, and classification models are finalized as a separate step in the pipeline; 2) a deep learning pipeline that is more automated and allows spontaneous emergent features to be learned by the machine, while simultaneously developing a classifier. Both pipelines explored the impact of clinical, radiomic and RTx-omic features on outcome predictions. In this study we use LRF prediction at three years in H&N cancer as our case study. Predictions are performed using different modeling methods, each of which has specific benefits to our research question. This section details the data curation, and pipelines used for our analysis.

3. Data curation and preparation

We used a single dataset from the Princess Margaret Cancer Centre with institutional research board approval. The dataset contained planning computed tomography (CT) DICOM images, DICOM RT Structures, DICOM RT Dose, and clinical variables for 190 patients. Gross tumour volumes (GTV) in the DICOM RT Structure file were contoured by radiation oncologists (experience levels ranging between 5 and 30 years) for intensity modulated radiation therapy (IMRT) treatment based on clinical-radiological evidence of disease extent. Often during contouring, simulation magnetic resonance imaging (MRI) was fused with the planning CT to aid in target delineation. Additionally, H&N Radiation Oncology Quality Assurance Rounds occurred weekly for the opportunity to peer-review RT target volumes, including the GTV and clinical target volumes (CTV). Additional patient details can be found in Table 1 of the Supplementary Material.

The inclusion criteria for this study was an oropharynx disease site, squamous cell carcinoma pathology, 70 Gy prescribed dose in 35 fractions to the primary GTV, and full delivery of prescribed dose. Application of inclusion criteria reduced our dataset from 190 patients to 160 patients with 18 LRF events at three years. This resulted in an imbalanced dataset with an event rate of 11%; a challenging problem for most modeling methods, but one we believe can be modeled using appropriate disease, image and treatment descriptors. Furthermore, for our highly user driven pipeline (hence forth referred to as Machine Learning Pipeline and described below), only patients who did not have dental artifacts (DA) were included. This was to safeguard our radiomic features against spurious image signals [15] and reduced the dataset further to 64 patients with 7 LRF events. For our more automated pipeline (hence forth referred to as the Deep Learning Pipeline and described below) all 160 curated patients were used with the assumption that the convolutional neural network (CNN) would learn to distinguish between important and irrelevant machine generated features regardless of the DA status of the image.

4. Machine learning pipeline

Our Machine Learning Pipeline (MLP) (Fig. 1) was designed to allow a researcher to have control over all aspects of feature definition, feature space reduction, and model building, and validation. This is typical of the traditional clinical modeling or radiomics pipelines where features are defined based on prior knowledge of the disease, or hand-engineered and extracted from images.

4.1. Patient-specific features

The complexity of a patient's disease generates a variety of characteristics that may be of benefit when determining a treatment specific conditional-prognosis. In our MLP we aim to describe the disease using a variety of features that are known predictors of H&N patient LRF (clinical features), as well as more exploratory features quantifying a patient's planning CT signal (radiomic features) and their personalized RT treatment plan designed based on their intrinsic anatomical variants (RTx-omics features). Following is a description of the features found in the different feature classes. It should be noted that due to the small number of events found in our dataset that our MLP suffers from the 'curse of dimensionality' – this is mitigated using feature space reduction which is described as a later step.

4.2. Clinical features

Clinical features for our patients were collected from the Princess Margaret Cancer Center H&N Anthology. Patient Age, smoking status, drinking status, disease subsite, T stage, N stage, overall stage, and p16 status were included as clinical features.

4.3. Radiomic features

PyRadiomics 2.0 [26] was used to extract radiomic features (n = 99) that quantified the planning CT (in Hounsfield Units, HU) within a patient's GTV. Images were resampled to an isotropic pixel size of 1 mm using BSpline interpolation, and a bin width of 25 was used for texture feature calculation [27]. All features from the first order statistics (n = 18), shape (n = 12) and texture (GLCM (n = 23), GLSZM

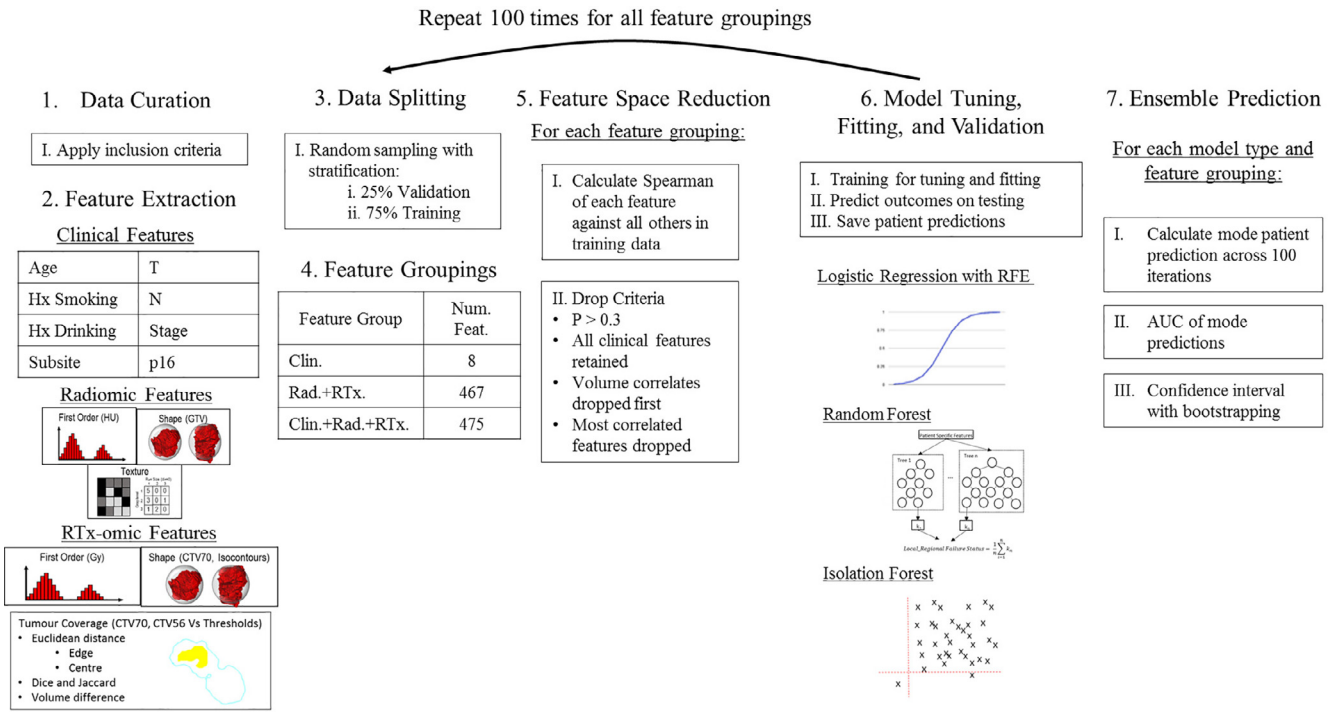

Fig. 1. Machine Learning Pipeline for generalized feature analysis and outcome prediction in H&N patients. Step 1: application of inclusion and exclusion criteria; Step 2: extraction of generalized features – clinical, radiomic and RTx-omic; Step 3: random sampling of patients into training and validation datasets; Step 4: feature grouping based on goal of determining added benefit of radiomic and RTx-omic features; Step 5: reduction of feature set based on spearman rank values calculated within a specific feature grouping; Step 6: tuning, fitting and validating of three different modeling techniques; Step 7: calculation of PR-AUC based on model prediction of patient outcome across 100 iterations of Steps 3–6.

(n = 16), GLDM (n = 14) and GLRLM (n = 16)) classes were extracted. For details on feature equations please see the extensive PyRadiomics documentation [28].

4.4. RTx-omic features

RTx-omic features were extracted using PyRadiomics 2.0 and a custom PyRadiomics module designed to quantify relationships between two ROIs. First order statistical features were extracted from the planned dose volume, where the voxels represent planned RT dose (Gy), instead of HU as was quantified with the radiomic features. These features were extracted from the GTV, CTV70 (clinical target volume at 70 Gy), CTV56 (elective clinical target volume at 56 Gy) and isocontours at 95 and 100% of 70, 63 and 56 Gy, which were generated by thresholding the planned dose volume. Shape features for the isocontours and CTV70 were also calculated.

Tumour coverage was quantified using the custom PyRadiomics module. The module was developed to calculate the Euclidean distance between two ROI edges and centers, as well as Dice and Jaccard metrics, and volume differences. These metrics were calculated to compare all isocontours against CTV70 and CTV56. RTx-omic features were defined in collaboration with a medical physicist, radiation therapist and radiation oncologist.

4.5. Ensemble LRF prediction and validation

We explored the impact of radiomic, RTx-omic and clinical feature groups on prediction of LRF at three years using a multi-step process that explored a variety of modeling methods.

4.6. Data splitting and feature grouping

Our dataset was split into 75% training and 25% testing sets. The data was randomly sampled and stratified to ensure equal distribution of LRF events in each set; this resulted in the training and testing sets containing 5-6 and 1-2 patients, respectively. Feature groups were combined to explore the added predictive value of radiomic and RTx-omic features on accepted clinical factors. This resulted in the following combinations of features 1) clinical, 2) radiomic and RTx-omic; and 3) clinical, radiomic and RTx-omic. (Step 4 in Fig. 1).

4.7. Feature space reduction

Each of the three training set feature groupings underwent feature space reduction to decrease the number of correlated features and the chances of overfitting to the training data. Feature space reduction involved calculating the Spearman rank value for each feature against all other features in the feature group of interest. If the Spearman rank value between two features was greater than or equal to 0.3 the features were considered correlated and one of them was dropped/removed. Clinical features were never dropped, since the goal was to determine added value above accepted clinical features, features correlated to volume were dropped first, and if two features still remained, the feature that was correlated to the most number of other features was dropped (Step 5 in Fig. 1).

4.8. Model tuning, fitting and validation

After feature space reduction, model tuning, fitting and validation was performed using the training dataset. Three different modeling techniques available in Python's Scikit Learn package [29] were explored: a) logistic regression with recursive feature elimination (LOG) [30] – a highly interpretable method of modeling widely accepted in the clinical environment; b) random forest (RF) [31,32] – a more complex method aggregating multiple decision trees together to reduce bias and variance; c) isolation forest (IF) [33,34] – an ensemble of isolation trees designed to detect data anomalies, such as an LRF event in our dataset. Tuning parameters and methods can be found in our

Supplementary Materials.

After tuning based on the feature grouping of interest, a LOG, RF and IF model were fit to the training data for the same feature group of interest. The fit and tuned LOG, RF and IF models were used to predict the probability of a testing patient experiencing an LRF event, which was saved in an array (Step 6 in Fig. 1). Steps 3–6 of Fig. 1, Data Splitting and Feature Grouping, Feature Space Reduction, and Model Tuning, Fitting and Validation, were repeated 100 times for different splits of the data.

4.9. Ensemble prediction

After fitting 100 models for each of the feature groupings, and each of the modeling types, we performed an ensemble prediction of treatment specific conditional-prognosis for H&N patient LRF at three years. Each combination of feature grouping and modeling method had an array where a row represented a patient and a column represented one of the 100 iterations. The average probability of a patient experiencing an LRF event across the 100 iterations was taken to be that patient's probability of experiencing an LRF event. The precision recall area under the curve (PR-AUC) (described below) for a given feature grouping and modeling method was calculated on the average patient LRF event probability. Confidence intervals (CI)were calculated using bootstrapping.

4.10. Deep learning pipeline

In our Deep Learning Pipeline (DLP) a deep learning network was utilized to minimize user influence (Fig. 2). This gave the system control over what features to extract and how to combine them in the most beneficial way for LRF prediction. Three deep learning networks (DLNs) were trained: 1) Clinical, 2) Radiomic + RTx-omic, and 3) Clinical + Radiomic + RTx-omic. Patient image, RTDose and contour volumes were used in models 2 and 3.

5. Data encoding, generation and pre-processing

5.1. Clinical data encoding

The categorical clinical features (i.e. smoking status, drinking status, disease subsite, T stage, N stage, overall stage, and p16 status) were one-hot-encoded using the function "OneHotEncoder" from Python's scikit-learn 0.22 package [29] to obtain binary categorizations that are easier for the machine to interpret. Age is a continuous variable and remained unaltered.

5.2. Contour volume generation

The three-dimensional (3D) contour volumes were generated by combining each patient's GTV, CTV56 and CTV70 into a single volume, where the intersection of all three regions of interest was denoted by a 1, the intersection of CTV56 and CTV70 was denoted by a 2, and the remaining portion of CTV56 was denoted by a 3.

5.3. Image, dose and contour pre-processing

The 3D image, dose and contour volumes were processed prior to usage in CNN training or testing using a multistep procedure:

1) Voxels in the CT image, RTDose volume and contour volume were interpolated to isotropic 1 mm^3 sizes. SimpleITK's linear resampling image filter was used for the CT image and RTDose, and a nearest neighbor resampling filter was used for the contour volume. This reduced variability in the images and therefore improved processing by the CNN.
2) The CT image and RTDose volume were normalized based on the mean and standard deviation of the population CT and RTDose volumes, respectively. Normalization ensures similar data distributions, allowing for fast convergence during network training.

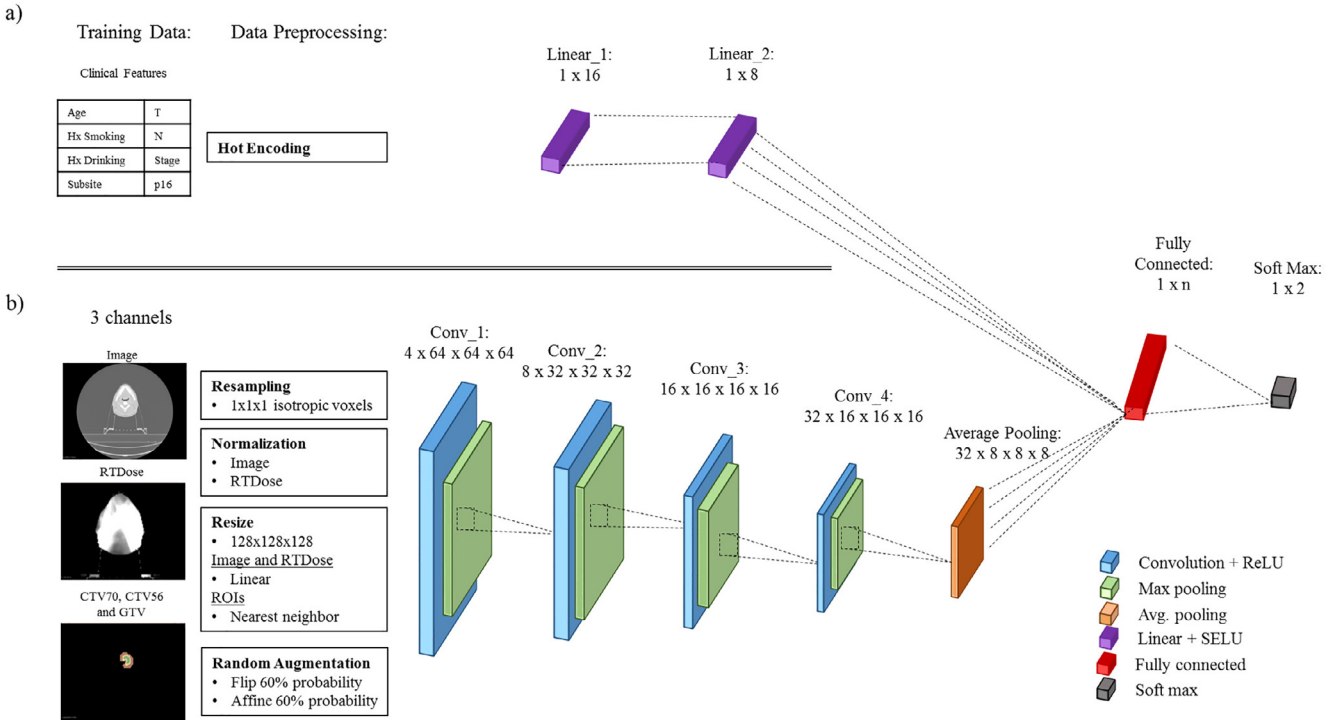

Fig. 2. Deep Learning Pipeline for generalized feature analysis and outcome prediction in H&N patients. a) the features, pre-processing, and linear layers used for our Clinical network. n in the fully connected layer is 8. b) the data, pre-processing steps and CNN layers for our Radiomic + RTxomic network. n in the fully connected layer is 16384. The final network described is the Clinical + Radiomic + RTx-omic network which combines both a) and b). In this network, n in the fully connected layer is 16392.

Table 1
PR-AUC and corresponding confidence intervals (CI) for both pipelines.

Random performance PR-AUC = 0.11	User-Driven Pipeline			Automated Pipeline
	Random Forest • Random stratified subsampling • 75/25% split DA patients excluded	Logistic Regression • Random stratified subsampling • 75/25% split DA patients excluded	Isolation Forest • Random stratified subsampling • 75/25% split DA patients excluded	Deep Learning • 10 fold cross validation • 20 epochs DA patients included
Clin.	0.61 (0.25–0.96)	0.66 (0.33–0.93)	0.42 (0.18–0.75)	0.38 (0.23–0.54)
Rad. + RTx.	0.12 (0.05–0.22)	0.19 (0.07–0.56)	0.26 (0.15–0.62)	0.36 (0.17–0.54)
Clin. + Rad. + RTx.	0.33 (0.12–0.73)	0.15 (0.08–0.48)	0.20 (0.12–0.50)	0.32 (0.20–0.45)

3) CT Images, RTDose volumes and contour volumes were resized to a grid size of 128^3. Resizing was performed using the open-source scikit-image library [35], which preserves the image's HU distribution. The aspect ratio of the volumes were maintained by padding each of the volumes to a uniform size based on the largest dimension in the 3D volume.
4) Two types of data augmentation were performed to introduce randomness to the training data and minimize the chances of overfitting. 1) Flipping of the volumes in the lateral direction. 2) affine transformations with rotations between −16 and + 16 degrees, translation in vertical and horizontal directions by 15% or the volumes width and height, and scaling by factors between 0.85 and 1.25. Each of the two different data augmentation types had a mutually exclusive chance of occuring of 60%.

5.4. CNN architecture and training

We used the open-source python library, PyTorch [36], to train our deep learning networks. A virtual machine from VMware, Inc. with 10 Intel Xeon CPU E5-2690 processors and a NVIDIA Tesla K40m GPU was used for training and testing. Ten-fold stratified cross validation was performed using the 160 curated patients and 18 LRF events.

1) Clinical DLN: Utilized only the clinical features described above (Fig. 2a). The one hot encoded feature representations, along with the unaltered age feature were pushed through a two linear neural network layers with weighted optimization to account for class imbalance. The first layer underwent scaled exponential linear units (SELU) activation [37], the output of the second layers was used as input to a single fully connected layer. Outcomes were predicted using softmax classification.
2) Radiomic + RTx-omic DLN: Used the patient image, RTDose and associated contour volume in a three-dimensional, three-channel, four-layer CNN (Fig. 2b). The outputs of all layers, except the final layer, underwent batch normalization, rectified linear unit functioning (ReLU) activation and max pooling [38,39]. The output of the final layer of the CNN underwent average pooling followed by a fully connected layer and softmax classification. The first CNN layer had convolutional kernel sizes of 5 with a padding of 2; the remaining layers used a size of 3 and padding of 1. Weighted optimization was used to account for imbalanced class distributions.
3) Clinical + Radiomic + RTxomic DLN: A combination of the two previously described networks (Fig. 2 a and b). The output of the final linear layer from Clinical and the output from the final CNN layer from Radiomic + RTxomic are combined in the fully connected layer prior to softmax classification. Weighted optimization was used to account for imbalanced class distributions.

5.5. Scoring metric

In order to take into account the large class imbalance found in our dataset, the area under the PR-AUC was used for performance evaluation. PR-AUCs are more sensitive to class imbalances, and therefore provide a better metric of evaluation for our study compared to the more commonly used receiver operator characteristic curves [40].

Precision is the ratio of the number of true positives divided by the sum of true positives and false positives (Eqn. (1)).

$$Precision = \frac{True\ \ Positives}{True\ \ Positives + False\ \ Posivites} \quad (1)$$

Recall is the ratio of the number of true positives divided by the sum of true positives and false negatives (Eqn.2).

$$Recall = \frac{True\ \ Positives}{True\ \ Positives + False\ \ Negatives} \quad (2)$$

When determining whether a PR-AUC is better than random the balance of classes must be considered. This is achieved by determining the probability of randomly guessing a positive event, given by the number of positive events divided by the sum of the positive and negative events, which is equivalent to the event rate of the dataset. For our dataset, a PR-AUC of 0.11 is considered random performance. PR-AUCs were calculated for our work using Python's Sci-kit learn library [29].

For additional comparison, the PR-AUC of univariate GTV volume was calculated, a known prognostic factor for H&N cancer [41].

6. Results

PR-AUC values above 0.11 are considered to have better than random performance. Our MLP with clinical features and LOG modeling had the overall highest PR-AUC for LRF prediction at three years of 0.66 (0.33–0.93). RF modeling performed best with clinical features only (PR-AUC = 0.61 (0.25–0.96)), and IF also performed best when utilizing only clinical features (PR-AUC = 0.42 (0.18–0.75)). Our DLP performed best with only clinical features as well (0.38(0.23–0.54)). PR-AUC values for all modeling methods and feature combinations can be found in Table 1. All of the above mentioned models and feature groups performed better than our univariate GTV volume predictor, which had a PR-AUC of 0.21.

Table 2 presents the number of features that were retained after feature set reduction and model fitting in our MLP. The number of features was averaged across all 100 fittings for each of the feature groupings and modeling methods. It can be seen that all clinical features are retained after feature set reduction in the clinical feature grouping, as is expected based on the design of the feature set reduction method. Additionally, radiomic and RTx-omic features are known to correlate to clinical and volume features; therefore, more features were retained in the radiomic + RTxomic model than the clinical + radiomic + RTxomic model.

7. Discussion

The ability to conditionally prognosticate a cancer patient's outcome based on their treatment is foundational to making personalized

Table 2
The number of features remaining after feature set reduction (FSR) and model fitting for the user-driven pipelines. The average number of features and standard deviations are presented.

	User-Driven Pipeline					
	Random Forest		Logistic Regression		Isolation Forest	
	FSR	Modeling	FSR	Modeling	FSR	Modeling
Clin.	8 ± 0	3 ± 1	8 ± 0	6 ± 2	8 ± 0	6 ± 0
Rad. + RTx.	409 ± 0	3 ± 1	409 ± 0	31 ± 56	409 ± 0	6 ± 0
Clin. + Rad. + RTx.	95 ± 13	4 ± 1	128 ± 21	27 ± 29	150 ± 15	6 ± 0

cancer medicine a reality. To accommodate existing and rapidly emerging patient and treatment information, processes are required to integrate the variety of disease features available, including RTx-omic features that precisely quantify the treatment. Our work presents two user controlled pipelines where clinical features with LOG had the highest PR-AUC when predicting LRF at three years for H&N cancer patients. More importantly, our results provide insight pertaining to the development of user-controlled pipelines for outcome prediction. In particular, the importance of curation, and user, data and methodology bias awareness as it pertains to result interpretation.

The clinical features selected for this study provided the highest PR-AUC for H&N LRF prediction at three years when combined with LOG modeling in our highly bespoke user driven pipeline. Although a promising result, large CIs indicate that subsampling was important and too few LRF events were present in our data. Additionally, the large CIs prevent us from definitively stating one model is better than another. Both of these observations suggest that a larger dataset may have resulted in a different final observation. These results are not to say that imaging and RT treatment features do not provide additional information important to the prediction of LRF, only that with the current data and our current features they do not draw immediate conclusions.

When utilizing imaging and RT treatment information only, our DLP performed better than all three MLP modeling methods. This result may indicate that the machine was able to detect and extract features that were not seen and more informative than the hand-engineered/user-knowledge-informed features present in our MLP. Additionally, in our DLP, the Radiomic + RTx-omic DLN had comparable performance to the Clinical DLN (0.36 (0.17–0.54) vs. 0.38 (0.23–0.54), respectively, p-value = 0.97). This indicates that information could be extracted from images and RT treatment plans that is useful for conditional prognostics; we just have yet to obtain enough data to strengthen this signal. Future work may also be able to utilize larger resampling grids to retain more imaging and treatment details, providing more nuanced information to the machine. Despite this promise, LOG prediction with clinical features still performed better than both of these networks, and could be due to the breadth of knowledge included in the curation of clinical feature definition [42–44], therefore requiring less complicated modeling techniques.

Additionally, the DLP had more consistent PR-AUCs and smaller CIs across all feature combinations when compared to MLP modeling methods. This may be affected by differences in training/validation data, but it also seems to indicate that by using a less user-driven approach we are able to obtain more consistent information out of all data types when using our defined topology. These observations also lead the authors to suggest that various modeling methods, feature selection techniques, topology configurations, and levels of human interaction are tested during model development to determine the optimal performance for a given research question. This type of testing has been performed by other groups when utilizing radiomic features for outcome prediction [45,46] and would ensure that the best results for that given research question are achieved.

Predictions utilizing quantitative image analysis and pattern recognition has been an area of study for close to two decades [47,48]. Recent utilization of these methods in cancer prognostics with hand-engineered features has found promising results, particularly in H&N cancer [49–54]. Deep learning is also being researched for its utility in this area [25,55]. In a recent study by Diamant et al. [56]., it was determined that deep learning methods were capable of identifying traditional radiomic features, in addition to newly generated features, that were beneficial in H&N outcome prediction. Although the above mentioned work is promising, a recent study by Ger et al. [57] found that consistent associations between radiomic features and outcome in H&N patients could not be found, even when utilizing large datasets ($n > 600$) with standardized imaging practices.

Obtaining large, high quality clinical datasets that are applicable to a given research question is challenging, as was seen in this study. However, if a strong biomarker or feature is embedded in the data and driving the outcome of interest it should be apparent, regardless of the dataset size, which has a stronger impact on the CIs than the overall performance [58]. When developing predictive models, it is understood that more data is often preferred. Larger datasets improve statistical analysis of the model and have a higher chance of containing heterogeneities that models may encounter during clinical usage. More importantly, small datasets have increased potential for false positive and false negative errors [59] that are detrimental to health care resources and patient outcomes, respectively.

The authors believe that the largest limitation for this study was the number of LRF events. The event rate for LRF was small, and in combination with our dataset size, this left very few examples to learn from during training. To account for the imbalance we used upsampling in our MLP and weighted optimization in our DLP; however the large CIs indicate the importance of subsampling in our study and the need for larger more diverse data. Additionally, utilization of uniform and high quality plans developed using the same planning criteria may have negatively impacted the final conclusions. Namely, it is possible that treatments were consistent enough that it was not possible to observe any LRF causing variations. Despite this, we were able to demonstrate the importance of benchmarking prognostic automated information generation pipelines against clinical variables which already achieve good predictions [15].

Another important limitation to the utilization of automated pipelines and data analysis is that imposed by the operator/human. Human knowledge is at the core of each step of an automated pipeline: data curation and collection, data pre-processing, feature definition – either through explicit definition or definition of a deep learning topology, feature selection, and model tuning, fitting and validation. Curation of the data in our study was guided by expert knowledge of clinical staff, as was definition of our RTx-omic features. Feature selection and modeling relied on prior author knowledge and experience. All of these steps will ultimately be biased by whomever is performing the experiments, which can be both a good and bad characteristic of the study. Until we are able to explore all permutations of potential features and machine learning methodologies within large datasets it is not possible to make definitive statements about the impact that automated pipelines will have on cancer care prognostics.

By not fully understanding the risks associated with applied

methods, we are likely to obtain unstable and misinformed results. From a user-driven pipeline perspective, some researchers [45,46] have done an excellent job of publishing their results as a function of feature selection and modeling method performance. These types of publications are a good starting point when designing an experiment. However, researchers are urged to accurately publish all of their methods, not just the ones that had the best results. Additionally, it is important to understand the risk of data contamination that occurs in these studies. It is not common practice to have a true "Hold-Out" dataset [60], and therefore caution is warranted whenever interpreting the out of sample error rate, value and impact of a publications results.

By exploring the rationale behind various steps of our processes we had important learnings regarding inherent biases present in current user-controlled pipelines; particularly when working with small datasets that contain only a few event of interest examples. There is a desire in this field to move towards the 'Big Machine' paradigm [61] as a way to handle big data and provide a way to analyze and integrate the large and diverse data pools found within healthcare in a consistent and interoperable way. The processes that we have presented in this paper could be considered the 'little machine', a proprietary example of how the big machine would be operated. However, much larger and diverse datasets are needed to make true progress.

8. Conclusion

Our work demonstrates the potential to aggregate together features from multiple specialties for cancer patient outcome prediction in user-controlled pipelines with various degrees of user interaction. Most importantly, our results provide insight pertaining the importance of data curation and quality, as well as operator, data and methodology bias awareness as it pertains to result interpretation in user-controlled pipelines.

Acknowledgements

The authors dedicate this manuscript to Mary Gospodarowicz in honour of her years of leadership in radiation oncology and cancer at the Princess Margaret. Additionally, the authors thank Tom Purdie and Mike Sharpe for their assistance in obtaining the Princess Margaret data used in this study.

Funding

The work was supported by the Natural Sciences and Engineering Research Council to MLW, the Strategic Training in Transdisciplinary Radiation Science for the 21st Century Program to MLW, the Canadian Institutes for Health Research, the Ontario Institute for Cancer Research, and the Terry Fox Research Institute. DAJ is the Mary and Orey Fidani Family Chair in Radiation Physics.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejmp.2020.01.027.

References

[1] Gospodarowicz MK, Henson DE, Hutter RVP, O'Sullivan B, Sobin LH, Wittekind C, editors. Prognostic factors in cancer. 2nd ed.New York: Wiley-Liss; 2001.
[2] Edelstein L. Ancient medicine: selected papers of Ludwig Edelstein. Baltimore and London: Johns Hopkins University Press; 1967.
[3] Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? BMJ 2009;338(feb23 1):b375. https://doi.org/10.1136/bmj.b375.
[4] Buchanan S. The doctrine of signatures. 2nd ed. University of Illinois Press; 1991.
[5] Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Heal Inf Sci Syst 2014;2(1):10.
[6] Wu P-Y, Cheng C-W, Kaddi C, Venugopalan J, Hoffman R, Wang MD. –Omic and Electronic health record big data analytics for precision medicine. IEEE Trans Biomed Eng 2017;64(2):263–73.
[7] Desmond-Hellmann S, et al. Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease. Washington, D.C.: The National Academies Press; 2011.
[8] Cadoni G, et al. Prognostic factors in head and neck cancer: a 10-year retrospective analysis in a single-institution in Italy. Acta Otorhinolaryngol Ital 2017;37(6):458–66.
[9] Wopken K, Bijl HP, Langendijk JA. Prognostic factors for tube feeding dependence after curative (chemo-) radiation in head and neck cancer: a systematic review of literature. Radiother Oncol 2018;126(1):56–67.
[10] Lin BM, et al. Long-term prognosis and risk factors among patients with HPV-associated oropharyngeal squamous cell carcinoma. Cancer 2013;119(19):3462–71.
[11] Mayne ST, Cartmel B, Kirsh V, Goodwin WJ. Alcohol and tobacco use prediagnosis and postdiagnosis, and survival in a cohort of patients with early stage cancers of the oral cavity, pharynx, and larynx. Cancer Epidemiol Biomarkers Prev 2009;18(12):3368–74.
[12] Jaffray DA. Image-guided radiotherapy: from current concept to future perspectives. Nat Rev Clin Oncol 2012;9(12):688–99.
[13] Welch ML, Jaffray DA. Editorial: radiomics: the new world or another road to El Dorado? JNCI J Natl Cancer Inst 2017;109(7):7–8.
[14] Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. Radiology 2015;278(2):151169.
[15] Welch ML, et al. Vulnerabilities of radiomic signature development: the need for safeguards. Radiother Oncol 2019;130:2–9.
[16] Overgaard J, et al. Five compared with six fractions per week of conventional radiotherapy of squamous-cell carcinoma of head and neck: DAHANCA 6&7 randomised controlled trial. Lancet 2003;362(9388):933–40.
[17] Bourhis J, et al. Hyperfractionated or accelerated radiotherapy in head and neck cancer: a meta-analysis. Lancet 2006;368(9538):843–54.
[18] Peters LJ, et al. Critical impact of radiotherapy protocol compliance and quality in the treatment of advanced head and neck cancer: Results from TROG 02.02. J Clin Oncol 2010;28(18):2996–3001.
[19] Shipley WU. Proton radiation as boost therapy for localized prostatic carcinoma. JAMA J Am Med Assoc 1979;241(18):1912.
[20] Lyman JT. Complication probability as assessed from dose-volume histograms. Radiat Res 1985;104(2):S13–9.
[21] Hope AJ, et al. Modeling radiation pneumonitis risk with clinical, dosimetric, and spatial parameters. Int J Radiat Oncol 2006;65(1):112–24.
[22] Jiang W, et al. Machine learning methods uncover radio-morphologic dose patterns in salivary glands that predict xerostomia in head and neck cancer patients. Int J Radiat Oncol 2018;102(3):S212.
[23] Monti S, et al. Voxel-based analysis unveils regional dose differences associated with radiation-induced morbidity in head and neck cancer patients. Sci Rep 2017;7(1). https://doi.org/10.1038/s41598-017-07586-x.
[24] Murdoch TB, Detsky AS. The inevitable application of big data to health care. JAMA 2014;309(13):1351–2.
[25] Mobadersany P, et al. Predicting cancer outcomes from histology and genomics using convolutional networks. Proc Natl Acad Sci 2018;115(13):E2970–9.
[26] van Griethuysen JJM, et al. Computational radiomics system to decode the radiographic phenotype. Cancer Res 2017;77(21):e104–7.
[27] Zwanenburg A, Leger S, Vallières M, Löck S. Image biomarker standardisation initiative, Dec. 2016.
[28] "pyradiomics" https://pyradiomics.readthedocs.io/en/latest/.
[29] Pedregosa F, et al. Scikit-learn: machine learning in python. J Mach Learn Res Jan. 2012;12:2825–30.
[30] Peng C-YJ, Lee KL, Ingersoll GM. An introduction to logistic regression analysis and reporting. J Educ Res 2002;96(1):3–14.
[31] Breiman L. Random forests. Mach Learn 2001;45:5–32.
[32] Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. Ann Appl Stat 2008;2(3):841–60.
[33] Liu FT, Ting KM, Zhou Z-H. Isolation Forest. In: 2008 Eighth IEEE International Conference on Data Mining, 2008, pp. 413–22.
[34] Liu FT, Ting KM. Isolation-based Anomaly Detection, vol. V, pp. 1–44.
[35] van der Walt S, et al. scikit-image: image processing in Python. PeerJ 2014;2:e453.
[36] Shaikh F. An Introduction to PyTorch – A Simple yet Powerful Deep Learning Library, 2018. [Online]. Available: https://www.analyticsvidhya.com/blog/2018/02/pytorch-tutorial/.
[37] Klambauer G, Unterthiner T, Mayr A, Hochreiter S. Self-normalizing neural networks, Adv. Neural Inf. Process. Syst., vol. 2017-Decem, pp. 972–81, 2017.
[38] Nielsen MA. Neural networks and deep learning. Determination Press; 2015.
[39] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436–44.
[40] Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE 2015;10(3):1–21.
[41] Chao KSC, et al. Intensity-modulated radiation therapy for oropharyngeal carcinoma: impact of tumor volume. Int J Radiat Oncol 2004;59(1):43–50.
[42] O'Sullivan B, et al. Outcomes of HPV-related oropharyngeal cancer patients treated by radiotherapy alone using altered fractionation. Radiother Oncol 2012;103(1):49–56.
[43] Brierley J, et al. Global Consultation on Cancer Staging: promoting consistent understanding and use. Nat Rev Clin Oncol 2019.
[44] Tirkes T, Hollar MA, Tann M, Kohli MD, Akisik F, Sandrasegaran K. Response criteria in oncologic imaging: review of traditional and new criteria. RadioGraphics 2013;33(5):1323–41.
[45] Leger S, et al. A comparative study of machine learning methods for time-to-event

survival data for radiomics risk modelling. Sci Rep 2017;7(1):13206.
[46] Parmar C, Grossmann P, Bussink J, Lambin P, Aerts HJWL. Machine learning methods for quantitative radiomic biomarkers. Sci Rep 2015;5(1):13087.
[47] Hall EL, et al. A Survey of Preprocessing and Feature Extraction Techniques for Radiographic Images, IEEE Trans. Comput., vol. C–20, no. 9, pp. 1032–1044, 1971.
[48] Harlow CA, Eisenbeis SA. The analysis of radiographic images. IEEE Trans Comput 1973;C-22(7):678–89. https://doi.org/10.1109/TC.1973.5009135.
[49] Bogowicz M, et al. Comparison of PET and CT radiomics for prediction of local tumor control in head and neck squamous cell carcinoma. Acta Oncol (Madr) 2017;56(11):1531–6.
[50] Kuno H, et al. CT texture analysis potentially predicts local failure in head and neck squamous cell carcinoma treated with chemoradiotherapy. Am J Neuroradiol 2017;38(12):2334–40.
[51] Vallières M, et al. Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. Sci Rep 2017;7(1):10117.
[52] Diehn M, et al. Identification of noninvasive imaging surrogates for brain tumor gene-expression modules. PNAS 2008;105(13):5213–8.
[53] Aerts HJWL, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 2014;5(1):4006.
[54] Bogowicz M, Tanadini-Lang S, Guckenberger M, Riesterer O. Combined CT radiomics of primary tumor and metastatic lymph nodes improves prediction of loco-regional control in head and neck cancer. Sci Rep 2019;9(1). https://doi.org/10.1038/s41598-019-51599-7.
[55] Kim DW, Lee S, Kwon S, Nam W, Cha IH, Kim HJ. Deep learning-based survival prediction of oral cancer patients. Sci Rep 2019;9(1):1–10.
[56] Diamant A, Chatterjee A, Vallières M, Shenouda G, Seuntjens J. Deep learning in head & neck cancer outcome prediction. Sci Rep 2019;9(1):1–10.
[57] Ger RB, 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(9):1–13.
[58] Hazra A. Using the confidence interval confidently. J Thorac Dis 2017;9(10):4125–30.
[59] Biau DJ, Kernéis S, Porcher R. Statistics in brief: the importance of sample size in the planning and interpretation of medical research. Clin Orthop Relat Res 2008;466(9):2282–8.
[60] Su X. Practical Machine Learning Course Notes, Feb.
[61] Dekker A. Who is Building the Big Machine, 2016. [Online]. Available: https://github.com/andredekker/BigMachine/wiki/Who-is-building-the-Big-Machine%3F.