# Radiation therapy outcome prediction using statistical correlations & deep learning

**André Diamant**

Department of Physics
McGill University

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of
*Doctor of Philosophy*

March 2020

# Acknowledgements

Firstly, I would like to sincerely thank my supervisor and mentor Dr. Jan Seuntjens for providing me the opportunity to pursue a more uncommon research direction within the medical physics landscape. He gave me the opportunity to find 100s of rabbit holes, while also providing guidance into which ones to concretely jump into. The freedom he provided me in doing so cannot be understated. I would also like to thank Dr. Issam El Naqa for his support and guidance, and the initial push into machine learning he gave me.

I had the pleasure of working with many medical physicists, radiation oncologists, medical residents, summer students, radiologists and more throughout the course of my research. They all contributed in some way to this thesis, reflected by their co-authorship on the manuscripts presented. In no particular order, thank you to Drs. Monica Serban, Reza Forghani, George Shenouda, Sergio Faria, Houda Bahig, Edith Filion, Cliff Robinson, Hani Al Halabi and Robert Doucet. Thank you specifically to the 'outcomes' group within the MPU, Dr. Avishek Chatterjee and Dr. Martin Vallières. The hours of discussion we had each week was instrumental in guiding the direction of my research and ensuring the statistical analysis was of exceptional quality. I'd also like to specifically thank Veng Jean Heng for co-writing one of the manuscripts within this thesis, and providing a significant amount of his time to gathering the data and performing a multivariable analysis for said manuscript.

I would be remiss to not thank Margery Knewstubb and Tatjana Nisic, who ensure the MPU runs like a well-oiled machine. Anytime there were logistical/administrative matters, they were there to help me quickly resolve them.

The social environment and the friends I made at the MPU made my time here truly enjoyable. There are too many to list individually (and I'd definitely omit some by accident), but I

would like to specifically thank those who were working alongside me for the entire duration of my PhD. Thanks to Gabriel Famulari, Haley Patrick, Veronique Fortier, Logan Montgomery, Joel Mullins & Tanner Connell. Although many left the unit at some point throughout my research, their support has carried on. Thanks to Susannah Hickling, Ali Toltz, Yana Zlateva, Jessica Perez and Marc-André Renaud for their continued support from afar.

Thank you to my mom, dad, brothers and sister for providing constant love, support and encouragement while providing me numerous opportunities which shaped me into the person I am today. Last but certainly not least, thank you to my wife Rachael and son Felix for always being there, listening, encouraging me when things are chaotic and always providing an unwavering pillar of support.

# Abstract

Prognosis after cancer treatment is a constant concern for physicians, patients and their surrounding friends and family. This is one of the reasons that treatment outcomes prediction is such a critical field of research. The sheer magnitude of data generated within a typical radiation oncology clinic each year facilitates the development and eventual validation of predictive and prognostic models. Furthermore, the technological advances driven by data science have enabled the usage of advanced machine learning techniques which can far exceed the performance of previously used conventional techniques.

Most cancer patients follow a standard radiation oncology workflow, which among other things includes medical imaging (CT/PET) and the creation of a radiation therapy treatment plan. As these sorts of data are (in theory) present for every patient, they are ideal variables to input into a predictive model. The goal of this thesis was to investigate these two types of pre-treatment input data (diagnostic imaging and dosimetric data) along with patient characteristics to identify associations and create models capable of predicting a cancer patient's treatment response following radiation therapy.

The first objective was to investigate dose-volume metrics as predictors of clinical outcomes in a cohort of 422 non-small cell lung cancer (NSCLC) patients who received stereotactic body radiation therapy (SBRT). A correlation between the dose delivered to the region *outside* the tumor and the occurrence of distant metastasis was revealed. In particular, patients who received above a certain threshold dose were shown to have significantly reduced distant metastasis recurrence rates compared to the rest of the population. This was first shown on 217 patients all of whom were treated with conventional SBRT treatment modalities. Next, a similar analysis was done on 205 patients who were treated with a robotic arm linear accelerator (CyberKnife). It was found that the CyberKnife cohort had both superior distant control and local control, suggesting that under current prescription practices, CyberKnife, as a delivery device, could be superior for treating NSCLC patients with SBRT.

The second objective of this thesis was to investigate the usage of a deep learning framework applied to raw medical imaging data in order to predict the overall prognosis of head & neck cancer patients post-radiation therapy. A *de novo* architecture was built incorporating CT images, resulting in comparable performance to a state-of-the-art study. Furthermore, our model was shown to recognize imaging features ('radiomics') previously shown to be predictive without being explicitly presented with their definition. The final portion of this work was the development of a multi-modal deep learning framework which incorporated CT & PET images along with clinical information. This was compared to the previous architecture built, showing substantial increase in prediction performance for both overall survival and local recurrence. It was also shown to function in the presence of missing data, a common occurrence within the medical landscape.

This work demonstrates that pre-treatment prediction of a cancer patient's post-radiation therapy outcomes is possible by learning correlations and building models from readily available data. Future efforts should be put towards data sharing & data curation to enable the creation and validation of models that eventually can be used in the clinic. Ultimately, predictive models should evolve into generative models whereupon one's treatment could be automatically created with the explicit intention of statistically optimizing that patient's outcomes.

# Résumé

Le pronostic après le traitement du cancer est une préoccupation constante des médecins, des patients et de leurs proches. Ce n'est qu'une des raisons pour lesquelles la prédiction des résultats est un domaine de recherche si critique. L'ampleur des données générées chaque année dans une clinique de radio-oncologie facilite la création et, éventuellement, la validation de modèles pronostiques prédictifs. En outre, les progrès technologiques induits par l'informatique ont permis l'utilisation de techniques avancées d'apprentissage automatique qui peuvent largement dépasser les performances des techniques précédemment utilisées.

La plupart des patients atteints du cancer suivent un flux de travail de radio-oncologie standard, qui comprend entre autres l'imagerie médicale (CT / PET) et la création d'un plan de traitement par radiothérapie. Comme ces types de données sont (en théorie) présents pour chaque patient, ce sont des variables idéales à saisir dans un modèle prédictif. Le but de cette thèse était d'étudier ces deux types de données de prétraitement pour trouver des associations et créer des modèles capables de prédire le pronostic d'un patient après la radiothérapie.

Le premier objectif était d'étudier les paramètres de dose comme prédicteur des résultats cliniques dans une cohorte de 422 patients atteints de cancer du poumon non à petites cellules (CBNPC) qui ont reçu une radiothérapie stéréotaxique corporelle (SBRT). Une corrélation entre la dose délivrée à la région *extérieur* de la tumeur et l'apparition de métastases à distance a été trouvée. En particulier, les patients qui ont reçu une dose au-dessus d'un certain seuil se sont avérés avoir un taux de récidive de métastases à distance considérablement réduit. Cela a été premièrement démontré sur 217 patients qui ont tous été traités avec des modalités de traitement SBRT conventionnelles. Ensuite, une analyse similaire a été effectuée sur 205 patients traités avec un accélérateur linéaire à bras robotisé (CyberKnife). Il a été constaté que la cohorte CyberKnife avait à la fois un contrôle distant et un contrôle local supérieurs, ce qui suggère qu'en vertu des pratiques de prescription actuelles, le CyberKnife pourrait être supérieur lors du traitement des patients CBNPC.

Le deuxième objectif de cette thèse était d'étudier l'utilisation d'un cadre d'apprentissage profond appliqué aux données brutes d'imagerie médicale afin de prédire le pronostic post-radiothérapie des patients atteints d'un cancer de la tête et du cou. Une architecture *de novo* incorporant des images CT a été construite, résultant en des performances comparables aux études de pointe. De plus, il a été démontré que notre modèle reconnaît les caractéristiques d'imagerie («radiomique») précédemment montrées comme prédictives sans en être explicitement informé de leur définition. La dernière partie de ce travail consistait du développement d'un cadre d'apprentissage profond multimodal qui incorporent des images CT & PET ainsi que des informations cliniques. Ce modèle a été comparé à l'architecture précédente, montrant des augmentations substantielles des performances de prédiction pour la survie globale et la récidive locale. Il a également été démontré qu'il fonctionnait en présence de données manquantes, un phénomène courant dans le paysage médical.

Ce travail démontre que la prédiction avant-traitement du pronostic post-radiothérapie d'un patient atteint du cancer est possible en apprenant des corrélations et en construisant des modèles à partir de données facilement disponibles. Des travaux futurs devront être consacrés au partage et à la conservation des données afin de permettre la création et la validation de modèles pouvant éventuellement être utilisés en clinique. Finalement, les modèles prédictifs devraient évoluer vers des modèles génératifs, après quoi le traitement pourrait être automatiquement créé avec l'intention explicite d'optimiser statistiquement le pronostic des patients.

# Contents

# List of Figures

# List of Tables

# List of Abbreviations

| | |
|---|---|
| **AUC** | Area Under the Curve |
| **AAA** | Analytical Anisotropic Algorithm |
| **CBCT** | Cone Beam Computed Tomography |
| **CNN** | Convolutional Neural Network |
| **CK** | CyberKnife |
| **CRT** | Conformal RadioTherapy |
| **CT** | Computed Tomography |
| **CTV** | Clinical Tumour Volume |
| **DICOM** | Digital Imaging and Communications in Medicine |
| **DSB** | Double Strand Break |
| **DM** | Distant Metastasis |
| **EBRT** | External Beam Radiation Therapy |
| **EQD2** | Equivalent Dose in 2 Gy Fractions |
| **ICRU** | International Commission on Radiation Units and Measurements |
| **IGRT** | Image Guided Radiation Therapy |
| **IMRT** | Intensity Modulated Radiation Therapy |
| **GTV** | Gross Tumour Volume |

| | |
|---|---|
| **HI** | Homogeneity Index |
| **HU** | Hounsfield Unit |
| **KM** | Kaplan-Meier |
| **LRC** | Loco-Regional Control |
| **LRF** | Loco-Regional Failure |
| **MRI** | Magnetic Resonance Imaging |
| **MC** | Monte Carlo |
| **NSCLC** | Non-Small Cell Lung Cancer |
| **NTCP** | Normal Tissue Complication Probability |
| **OAR** | Organ At Risk |
| **OR** | Odds Ratio |
| **OS** | Overall Survival |
| **PET** | Positron Emission Tomography |
| **PNG** | Portable Network Graphics |
| **PTV** | Planning Target Volume |
| **REB** | Research Ethics Board |
| **ROC** | Receiver Operating Characteristic |
| **RP** | Radiation Pneumonitis |
| **RT** | Radiation Therapy |
| **SSB** | Single Strand Break |
| **TCP** | Tumor Control Probability |
| **TPS** | Treatment Planning System |

| | |
|---|---|
| **US** | Ultrasound |
| **VMAT** | Volumetric Modulated Arc Therapy |

# Preface and Contribution of Authors

This thesis contains three published manuscripts (Chapters 4, 5 & 6) and one being prepared for submission (Chapter 7). All manuscripts represent original contributions to research. Two papers represent the first thorough investigation into a new predictive metric which could lead to a change in clinical practice (Chapters 4 & 5); the other two represent a novel application, development and refinement of a deep learning algorithm as applied to head & neck cancer outcome prediction (Chapters 6 & 7). A breakdown of each author's contributions to the manuscripts is listed below.

1. *André Diamant, Avishek Chatterjee, Sergio Faria, Houda Bahig, Edith Filion, Issam El Naqa, Cliff Robinson, Hani Al Halabi & Jan Seuntjens, "Can dose outside the PTV influence the risk of distant metastases in stage I lung cancer patients treated with stereotactic body radiotherapy (SBRT)?", Radiotherapy & Oncology, vol. 128, no. 3, pp. 513-519, 2018.* (Chapter 4)
   I designed and performed all analysis, gathered all patients from their respective institutions, developed the in-house software required and wrote the manuscript. S.F., H.B., E.F. and H.A.H. provided patient data. A.C., S.F., I.E.N., C.R. and J.S. provided expert knowledge and consultation throughout the course of the research. All authors reviewed the manuscript.

2. *André Diamant, Veng Jean Heng, Avishek Chatterjee, Sergio Faria, Houda Bahig, Edith Filion, Robert Doucet, Farzin Khosrow-Khavar, Issam El Naqa & Jan Seuntjens, "Does non-coplanar radiotherapy reduce distant recurrence in NSCLC patients compared to conventional SBRT?", accepted for publication in Radiotherapy & Oncology, vol. 144, pp. 201-208, 2020* (Chapter 5)
   I designed and performed the majority of the analysis (with the exception of the multivariable analysis), gathered half the patients (VMAT/CRT) from their respective institutions, developed the in-house software required and wrote the manuscript. V.J.H. gathered the

CK patients, performed the multivariable analysis (and the motion robustness analysis) and co-wrote the manuscript with me. S.F., H.B. and E.F. provided patient data. R.D. provided expert consultation specifically with respect to the CyberKnife. Additionally, he provided motion tracking data for the motion robustness analysis. F.K.K provided expert statistical consultation. A.C., S.F., I.E.N., and J.S. provided expert knowledge and consultation throughout the course of the research. All authors reviewed the manuscript.

3. *André Diamant, Avishek Chatterjee, Martin Vallières, George Shenouda & Jan Seuntjens, "Deep learning in head & neck cancer outcome prediction", Scientific Reports, vol. 9, no.2764, 2019.* (Chapter 6)
   I (along with A.C. and J.S) conceived the initial project: using deep learning to perform outcome prediction on head and neck cancer patients. I developed the framework including the connection to radiomics and the concept of training *de novo*, analyzed the data, and wrote the manuscript. M.V. provided the base of the in-house script which extracted the radiomic variables. A.C., J.S., M.V. and G.S. provided expert knowledge and consultation throughout the course of the research. All authors reviewed the manuscript.

4. *André Diamant, Avishek Chatterjee, Martin Vallières, Monica Serban, Yujing Zou, Reza Forghani, George Shenouda & Jan Seuntjens, "Multi-modal deep learning framework for head & neck cancer outcome prediction", manuscript under preparation for submission.* (Chapter 7)
   I conceived the initial project: combining a PET image and clinical information with the CT image to perform outcome prediction using a deep learning framework on head and neck cancer patients. I developed the framework including the training methodology, analyzed the data, and wrote the manuscript. M.S. and Y.Z. retrieved the testing set data (to be included in the submitted manuscript). G.S. provided access to the testing set data along with curating their outcomes and clinical information. A.C., J.S., M.V., R.F. and G.S. provided expert knowledge and consultation throughout the course of the research. All authors reviewed the manuscript.

# Chapter 1

# Introduction

## 1.1 Cancer

Cancer is both one of the most prevalent diseases in society along with being one of the most life-altering ones. Nearly 1 in 2 Canadians will develop some form of cancer during their lifetime, while nearly 1 in 4 will die from it [1]. Many will have to endure the emotional and physical hardships associated with cancer diagnosis (and treatment) multiple times as cancer often recurs, either locally or through distant metastases. These are but few of the reasons why there is great value in the discovery of predictive characteristics that can be used to model prognosis and inform both the patient and their care team of potential future risk.

Cancer results from abnormal cell growth, which is to some extent caused by stochastic mutations occurring during cell replication [2]. Abnormal cell growth results in the fundamental truth that cancer cannot be eradicated in the manner that a virus such as smallpox can be. However, what *is* achievable is the advancement of diagnosis & treatment mechanisms to the point where cancer is (mostly) *curable*, through the drastic increase of survival rates and reduction of recurrence risk. With the influx of data that is also becoming available in an increasingly organized fashion in the 21$^{st}$ century healthcare environment, along with the advent of artificial intelligence (AI) methodologies, 'outcome prediction' research is becoming incredibly prevalent. By combining multiple forms of information, which is in principle readily available within the clinic, predictions regarding prognosis/diagnosis can be made. These predictions could enable treatment strategy alterations in the light of data collected during treatment or perhaps even guide the creation of a *truly* personalized treatment plan in lieu of the ‘standard of care’ treatment. The field of medical physics in particular concerns the usage of radiation to eradicate cancer cells, known as radiation therapy. The focus of this thesis is specifically with respect to cancer patients treated with radiation therapy, further introduced in the following section.

## 1.2 Radiation therapy

There is a negative stigma surrounding the use of radiation within society, however radiation therapy (RT) has proven to be an particularly effective cancer treatment technique. RT is used in approximately 50% of cancer treatments [3, 4]. This can be in conjunction with chemotherapy, targeted therapy and/or surgery, or utilized as the sole treatment method. The unit which defines the amount of radiation delivered during radiation therapy, and is thus prescribed (akin to a drug) is known as the gray (Gy). The gray is defined as the energy deposited per unit mass, 1 Gy = 1 joule/kilogram.

One form of radiation therapy (RT) involves directing a beam of *ionizing radiation* through a patient’s body, carefully targeted at their cancer. Charged particles (such as electrons, protons or carbon ions) are considered directly ionizing as they deposit energy through Coulombic interactions. Photons are traditionally classified as indirectly ionizing because they must undergo interactions (photoelectric effect[1], Compton scattering and pair production) to produce secondary charged particles which then directly deposit energy [5, 6]. The most common way to deliver radiation to a patient is through external beam radiation therapy (EBRT). The EBRT classification (as opposed to *internal* beam radiotherapy) solely refers to the fact that the radiation source is *external* to the patient. EBRT can be delivered using beams of protons, electrons, heavy ions or photons. By far, the most common of these is the high-energy photon beam, delivered by a linear accelerator (or ‘linac’). A linac accelerates electrons to MeV-energies which strike a tungsten target, emitting photons (said to be in the MV-range) which are directed towards the patient. The majority of treatments (and consequently linacs) deliver 6 MV photons, although there is some variety (for instance, IROC-H data shows that 96.4% of lung stereotactic

[1] The designation “indirectly ionizing” is in fact incorrect since photons can directly ionize atoms; what is more correctly meant is “indirectly energy-depositing” since the majority of the photon energy is *transferred* to orbital electrons removed from the atom which subsequently deposit their energy through Coulombic interactions with multiple atoms in the medium.[5]

treatments are at 6 MV, remainder is mostly 10 MV [7]).

RT could be considered the most fundamental way of eradicating a cell as it functions by damaging the cellular DNA. This occurs through either single-strand breaks (SSBs) or double-strand breaks (DSBs), both of which can cause substantial damage to a cell's ability to survive. While the cell has numerous complex repair mechanisms [2], the delivery of a precise amount of radiation (i.e., prescription) has been empirically determined to provide an adequate probability of eradication. These repair mechanisms (and the biological differences between healthy tissue and cancerous tissue) can also be taken advantage of in order to spare healthy tissue. However, it is an inherent consequence of RT that healthy tissue receives *some* amount of dose. To better understand these effects, concepts such as tumor control probability (TCP) and normal tissue complication probability (NTCP) were developed. This allows the physician to not only prescribe a tumor dose, but also to prescribe *dose constraints* for each organ-at-risk (OAR) which can ensure the NTCP remains sufficiently low. To first order, one can imagine a particular treatment modality could be beneficial to a patient with a tumor in a specific region, based on differing NTCPs (graphically depicted in Figure 1.1 [8]).

*Conceptually*, this should lead to radiation therapy aiming to maximize the TCP (by delivering as much dose as possible to the tumor) while minimizing the NTCP (by avoiding dose to non-cancerous, or 'healthy' tissue). In practice, this optimization problem is not always explicitly performed. Instead, modern day treatments are based on meeting dose-volume constraints as defined by the planner. However, this manifests itself in a 'one-prescription-fits-all' methodology, in which the standard-of-care is (more or less) widely defined for a particular cancer type and a treatment is crafted per patient to ensure that the relevant regions (tumor/healthy tissue) receive (at least/below) a predefined benchmark dose (defined so the TCP/NTCP remain acceptably high/low). The work performed throughout this thesis aligns itself with a healthcare ecosystem where treatment plans are built through the explicit optimization of TCP/NTCP. In the future, TCP/NTCP will not be the metric which is explicitly optimized (as they currently are concepts defined as population averages); however the idea of 'outcomes-based treatment planning on a per-patient basis' (hopefully) will exist. For example, perhaps patient-personalized TCP/NTCP curves could be generated, based on that specific patient's pre-treatment data. Although this will require fundamental changes to the workflow within oncology, numerous areas of research (this thesis included) present novel

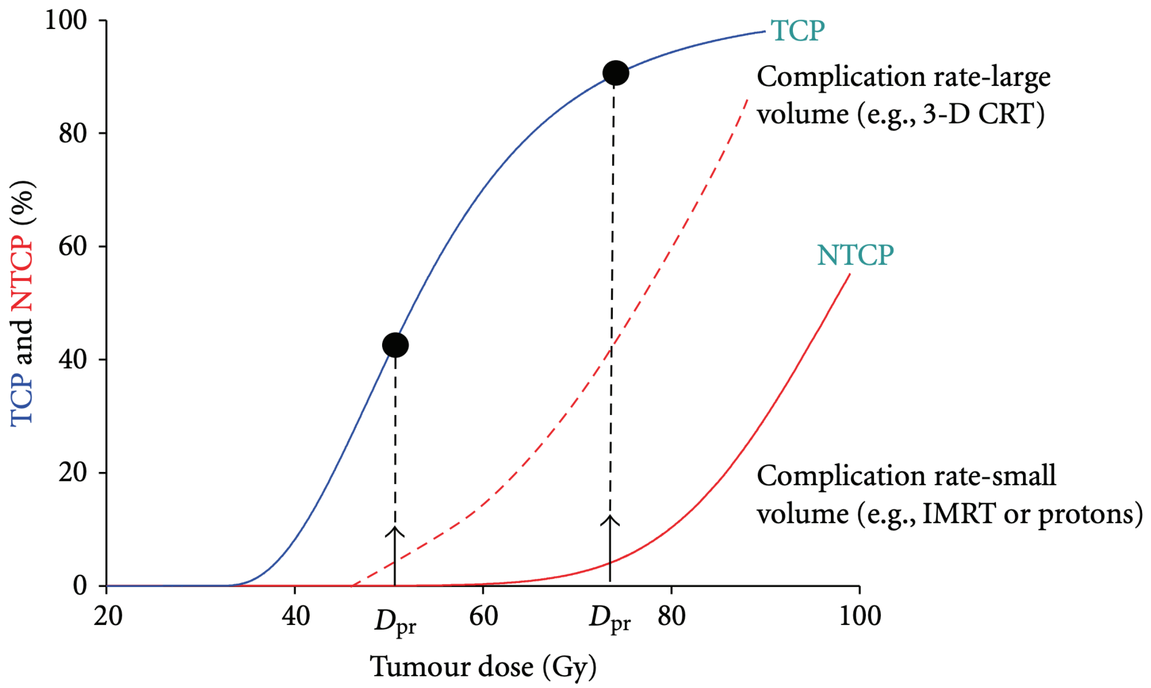


Figure 1.1 **Dose-response relationship for both tumor control probability (TCP, blue) and normal tissue complication probability (NTCP, red)**. In this example, two treatment modalities are shown impacting a differing amount of healthy tissue and thus having differing NTCP curves. Thus, for this hypothetical treatment, IMRT (intensity modulated radiation therapy) or protons would be superior to 3-D CRT (conformal radiation therapy). The dashed curve can only achieve a TCP of  45% while ensuring the NTCP stays below 10%, while the solid curve can achieve nearly double the TCP ( 90%) while maintaining the same 10% NTCP [8].

evidence that such a process could result in significant improvements to clinical outcomes. In particular, this thesis focuses primarily on clinical outcomes as an end-point (local control, distant control, survival, etc...). Treatment complications (and the ability to predict them) are an important area of research, however, they are not explicitly investigated throughout this thesis.

## 1.3 Workflow & available information

The workflow within a typical cancer clinic has multiple complex steps involving a variety of trained professionals and resources (Figure 1.2 [9, 10]). Prior to even entering the radiation oncology department (this workflow assumes radiation therapy is deemed the appropriate treatment), the patient has a consultation with a physician which (nearly always) leads to a computed tomography (CT) scan (within the radiology department). This often also involves other forms of imaging, most commonly magnetic resonance imaging (MRI) & positron emission tomography (PET). Pathology is nearly always done, unless a patient is considered unfit to tolerate its invasive nature. The results are discussed by the medical team and presented to the patient. If cancer is present, the patient formally enters the radiation oncology workflow. Another image (called CT-simulation) is performed, whereupon the relevant anatomy is *contoured*. This image is needed as it mimics the positioning of treatment, allowing the delivery to be as accurate as possible. The most important contours are the gross tumor volume (GTV), the planning target volume (PTV) and any organs-at-risk (OARs). The GTV is defined as the macroscopic disease seen on the CT, while the PTV represents an added margin (typically on the order of a few mm) to account for patient movement and daily setup uncertainties. Occasionally, a clinical target volume (CTV) margin is added representing the incorporation of microscopic spread (and its associated uncertainty). An internal target volume (ITV) margin will be added if motion is of concern. Common OARs are the spinal cord, heart, lung, brain stem, bladder, rectum, etc... Next, the radiation oncologist prescribes a specific amount of radiation to the GTV, along with dose constraints to each OAR. This prescription depends on knowledge of the stage, the pathology, patient condition, sex, age, smoking status, etc... Following the prescription, medical physicists and/or dosimetrists create a *treatment plan* which is verified and approved by a medical physicist. The radiation oncologist then validates the treatment plan, formally approving it for treatment. Delivery of a treatment requires precise setup of the patient (often with immobilization) along with numerous verification steps, some of which may necessitate an adaptation of the originally approved plan ('adaptive radiotherapy'). Most patients receive more than 1 dosage of radiation therapy (defined as a *fraction*), and thus the delivery setup process occurs many times per patient. Note that the above workflow is a *simple* example. Sometimes multiple images (perhaps of multiple modalities) are required throughout the process, or the treatment planning process is repeated multiple times before the final plan is accepted by the radiation oncologist. Occasionally blood tests and/or genotyping is available, resulting in biomarker/genetic data, however this research does not incorporate these forms of data.

As a patient progresses through their cancer diagnosis and subsequent treatment, there is a tremendous amount of data created for a variety of reasons. As mentioned above, medical imag-

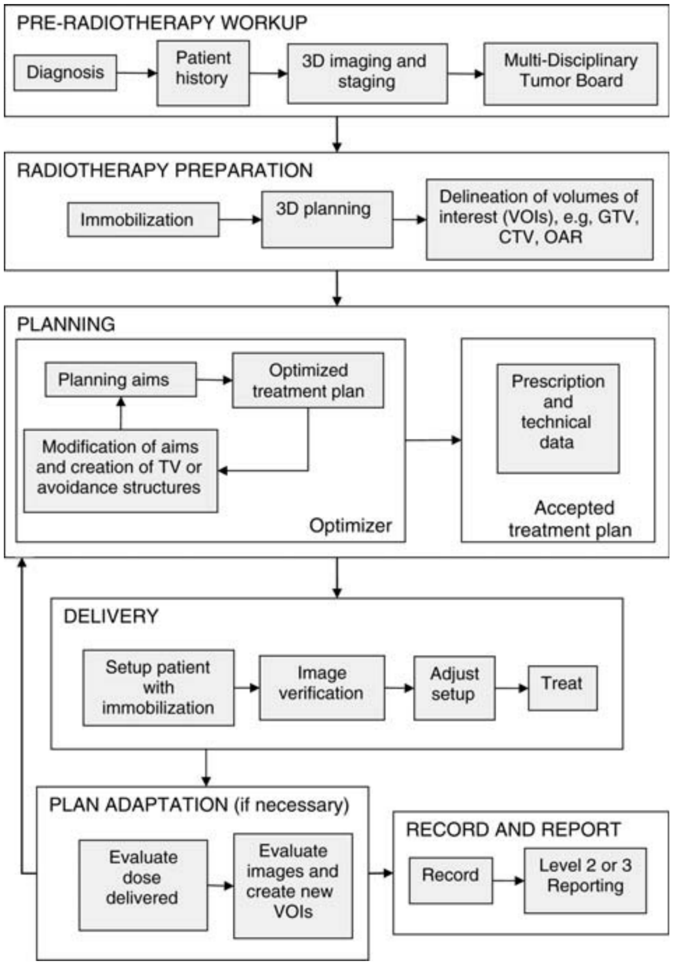


Figure 1.2 **Example workflow within a typical radiation oncology clinic that depicts the numerous individuals and resources involved [9]**.

ing & pathology (if available) is used to diagnose, locate and subsequently plan the treatment of the disease. Additionally, it is used to follow-up on the patient, to ensure that the cancer is in recession or otherwise determine the appropriate actions. All patients receive at *least* one type of 3D or 4D image, however many receive multiple. There are a wide variety of imaging modalities used within an oncology department, however the most common are computed tomography (CT), positron emission tomography (PET), magnetic resonance imaging (MRI) and ultrasound (US). Although these images are created for the aforementioned reasons, there is a wealth of potentially predictive outcome information hidden within. Additionally, the patient's clinical history and a variety of other laboratory tests (e.g., HPV/EBV) can be considered. This is what the entire field of outcome prediction research and big data revolves around; utilizing the otherwise 'discarded' information within a cancer clinic to improve prediction accuracy and eventually facilitate personalized treatment. A summary of the most commonly used forms of data is shown in Figure 1.3 [11]. The first two manuscripts of this thesis focus on the usage of treatment plans (dose distributions), while the final two manuscripts focus on the usage of medical imaging (specifically CT & PET).

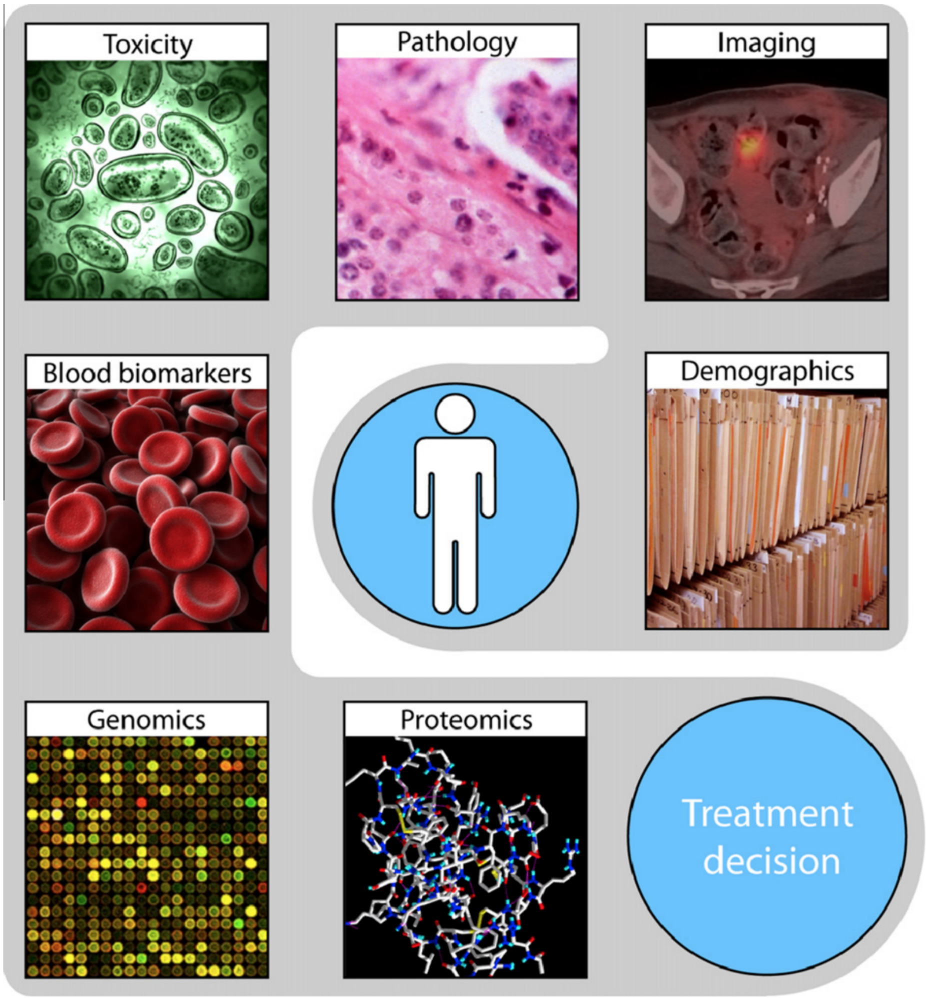


Figure 1.3 **Summary of prevalent forms of information which could be used within healthcare [11]**. Note this graphic does not represent every type of information within healthcare, but rather some of the most commonly used.

Unfortunately, the present ecosystem results in it being extremely difficult to acquire, curate and actually make use of most of these data. The process is often incredibly long, and in many ways parallels difficulties experienced in the experimental method. First, approval must be granted by Health Canada's and the Public Health Agency of Canada's Research Ethics Board (REB) [12]. This ensures that the research meets the highest ethical standards, and that the greatest protection is provided to participants serving as research subjects. Often, the application to the REB (or more accurately the Internal Review Board (IRB) of the parent institution) can be difficult to define particularly when the data is being analyzed retrospectively, in comparison to a prospective study where explicit consent forms could be made. Furthermore, when a study begins, it isn't always explicitly clear as to what methodologies will be applied or how many patients will be used. These aspects, among others, result in a slow approval process prior to even beginning any research. Once approval is granted, one of the most time-consuming (and grueling) tasks begins: *data curation*. Data curation is a multi-step process, that first requires learning the schema(s) of the (sometimes multiple) database(s) from which you are retrieving the data. Often, each patient has each form of data (e.g., images, clinical notes, outcomes) in different databases, requiring care to be taken in order to ensure that each variable gets correctly linked to the other variables. Furthermore, each type of data requires domain knowledge to analyze. In particular, CT/PET images and dose distributions (of which are used throughout this thesis) are stored in the 'Digital Imaging and Communications In Medicine' (DICOM) format [13]. Similarly, outcome data is typically not stored in a directly parsable format. Instead, it requires one to manually go through consultation notes and determine whether or not the outcome of interest occurred. When this process is repeated for multiple variables and 100s of patients, it becomes an incredibly time-consuming task. As will be discussed in the final chapter, data availability within the hospital environment is a major hurdle to overcome before outcome prediction models similar to those within this thesis have a realistic chance of being used in a clinical setting. The current environment requires the expenditure of far too many resources in order to make models that include hundreds of patients, let alone the thousands (or hundreds of thousands) that would be required to properly validate such models. Ideally, this thesis (and research similar to it) helps motivate the need for large-scale data sharing and standardization.

## 1.4 Thesis objectives

As the concept of personalized medicine is garnering more and more attention, methodology enabling accurate outcome prediction becomes essential. The hypothesis of this thesis is that by properly using the vast & diverse information that is (in principle) available pre-treatment, one can accurately determine whether a cancer patient is of particularly high/low-risk and thus could warrant a more aggressive/conservative treatment.

There are two primary objectives of this thesis (each of which have sub-objectives) related to outcome prediction using different methodologies, different cancer sites & different types of pre-treatment information:

1. Determine novel predictive dose metrics with respect to distant recurrence in non-small cell lung cancer patients treated with stereotactic body radiation therapy.
   (a) Build an algorithm capable of evaluating dose metrics to a particularly shaped region outside the tumor (Chapter 4).
   (b) Investigate the dose fall-off outside the tumor, for multiple treatment modalities (VMAT/CRT vs. CyberKnife) (Chapters 4 & 5).
2. Use deep learning (convolutional neural networks (CNNs)) applied to pre-treatment data to predict clinical outcomes of head & neck cancer patients.
   (a) Build a novel end-to-end CNN architecture trained *de novo* using CT images as input (Chapter 6).
   (b) Further develop the framework by building a multi-modal architecture capable of jointly considering clinical information, PET & CT images (Chapter 7).

## 1.5 Thesis outline

Chapter 2 provides an overview of the statistical concepts required to understand the various tests (and their statistical significance, or lack thereof) used throughout the research. Chapter 3 introduces machine learning, motivates deep learning while providing a literature review of its applications within medical physics and presents the theory behind convolutional neural networks (CNNs), a tool used extensively throughout the second half of this thesis. Chapters 4-7 are original manuscripts each describing outcome prediction within a variety of contexts. Chapter 4 investigates the dose *outside* of the tumor as a predictive factor within a cohort of non-small cell lung cancer patients. Chapter 5 performs a similar analysis to Chapter 4, on a separate cohort of patients who were treated with a robotic treatment modality in order to determine the impact of a specific treatment modality on the correlation. Chapter 6 examines the predictive performance of directly applying a CNN to pre-treatment CT images of a head & neck cancer patient cohort. This methodology is further expanded on in Chapter 7, with the incorporation of the same cohort's PET images and clinical information. Chapter 8 concludes with a summary of the entire thesis and a discussion on the envisioned future of outcome prediction, particularly how it could and should be applied to personalized medicine.

## References

[1] Government of Canada, Canadian Cancer Society, and Statistics Canada, "Canadian Cancer Statistics," tech. rep., 2018.

[2] E. J. Hall and S. Willson, *Radiobiology for the Radiologist*. Philadelphia: Wolters Kluwer, 2012.

[3] R. Atun, D. A. Jaffray, M. B. Barton, F. Bray, M. Baumann, B. Vikram, T. P. Hanna, F. M. Knaul, Y. Lievens, T. Y. M. Lui, M. Milosevic, B. O'Sullivan, D. L. Rodin, E. Rosenblatt, J. Van Dyk, M. L. Yap, E. Zubizarreta, and M. Gospodarowicz, "Expanding global access to radiotherapy.," *The Lancet. Oncology*, vol. 16, pp. 1153–86, sep 2015.

[4] Y. Lievens, M. Gospodarowicz, S. Grover, D. Jaffray, D. Rodin, J. Torode, M. L. Yap, E. Zubizarreta, and GIRO Steering and Advisory Committees, "Global impact of radiotherapy in oncology: Saving one million lives by 2035.," *Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology*, vol. 125, pp. 175–177, nov 2017.

[5] P. Andreo, D. Burns, A. Nahum, J. Seuntjens, and F. H. Attix, *Fundamentals of Ionizing Radiation Dosimetry*. No. June, 2017.

[6] E. B. Podgorsak and K. Kainz, *Radiation Oncology Physics: A Handbook for Teachers and Students*, vol. 33. Vienna: IAEA, 2006.

[7] IAEA, “Dosimetry of Small Static Fields Used in External Beam Radiotherapy: TRS-483,” tech. rep., 2017.

[8] A. E. Nahum and J. Uzan, “(Radio)biological optimization of external-beam radiotherapy,” *Computational and Mathematical Methods in Medicine*, vol. 2012, p. 13, 2012.

[9] C. Program, F. Quantities, I. Quality, and R. Ion-beam, “ICRU 83: Prescribing, Recording, and Reporting Photon-Beam Intensity-Modulated Radiation Therapy (IMRT),” *Journal of the ICRU*, vol. 10, no. 1, pp. NP–NP, 2010.

[10] N. Niroumandrad and N. Lahrichi, “A stochastic tabu search algorithm to align physician schedule with patient flow,” *Health Care Management Science*, vol. 21, pp. 244–258, jun 2018.

[11] P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R. G. Van Stiphout, P. Granton, C. M. Zegers, R. Gillies, R. Boellard, A. Dekker, and H. J. Aerts, “Radiomics: Extracting more information from medical images using advanced feature analysis,” *European Journal of Cancer*, vol. 48, pp. 441–446, mar 2012.

[12] Government of Canada, “Health Canada and the Public Health Agency of Canada’s (PHAC) Research Ethics Board - Canada.ca,” 2018.

[13] “DICOM Library - About DICOM format,” 2019.

# Chapter 2

# Statistical concepts

## 2.1 Introduction

A fundamental understanding of statistics is required to correctly interpret and utilize the results of any outcome prediction research. Although ‘statistical significance’ is often strived for due to the perceived community acceptance, caution and rigor are required when interpreting the results of any statistical test. This necessity applies to either end of the ‘statistically significant’ spectrum, whether one dismisses results because they’re not ‘statistically significant’ or vice-versa, where one believes results *solely* because they’re ‘statistically significant’ [1]. It is for this reason that this work always attempts to form a hypothesis as to *why* the results are what they are. A crucial statistical concept to stress is that *correlation does not inherently imply causation*.

Throughout the course of this thesis, various statistical tests and concepts were used. Rather than make some attempt to explicitly define them directly prior to their usage, brief descriptions will be in the following subsections. Ideally, this will provide the reader with the statistical background necessary to properly understand the presented results. First, odds ratios will be introduced, a commonly used statistical metric in a medical context that quantifies the association between two events. The concept of $p$-values will also be introduced, a metric used to assess the significance of the result of a statistical test. Next, the receiver operating characteristic (ROC) curve will be introduced, along with its associated metrics. The concept of Kaplan-Meier curves (and their analysis) will be motivated and discussed. Finally, the Cox proportional-hazards model (the multivariable analysis technique used in Chapter 5) will be explained.

## 2.2 Odds ratios & $p$-values

An *odds ratio* (OR) is a measure of association between an exposure and some outcome [2]. The exposure (or variable of interest) is typically some health characteristic or aspect of medical history (i.e. smoking history), but can also be any user-defined metric such as the mean radiation dose to a particular region. The outcome is often the presence of some disease or disorder (i.e., lung cancer given smoking history), but can also represent the occurrence of an event, such as death or cancer recurrence. Intuitively, the odds ratio represents the odds that a particular outcome will occur given some particular exposure, compared to the odds that the same outcome will occur without said exposure. Mathematically, this is represented by the following [2]:

| | **Diseased** | **Healthy** |
|---|---|---|
| **Exposed** | $D_E$ | $H_E$ |
| **Not exposed** | $D_N$ | $H_N$ |

$$\text{OR} = \frac{D_E/H_E}{D_N/H_N} \tag{2.1}$$

thus resulting in the following logical statements:

$$\text{OR} = 1 \implies \text{Exposure does not affect odds of outcome} \tag{2.2}$$

$$\text{OR} > 1 \implies \text{Exposure is associated with higher odds of outcome} \tag{2.3}$$

$$\text{OR} < 1 \implies \text{Exposure is associated with lower odds of outcome} \tag{2.4}$$

A very important consequence of this definition is that an OR $\neq 1$ does *not* necessarily imply a correlation and certainly does not imply causation. For example, it is very possible (probable with a small sample size) that an odds ratio of 1.05 could be entirely due to random chance (or statistical noise). For this reason, one should employ 95% confidence intervals and/or $p$-values. A 95% confidence interval estimates the precision of the odds ratio. Often, this is interpreted as representative of statistical significance *if* the range does not overlap the null value (OR = 1). However, this does not necessarily imply that a range that *does* overlap the null value is indicative of a lack of association.

*P*-values are used to quantify significance for many statistical tests. *Hypothesis testing* is when one uses a data-set to test the validity of a claim (the *alternative hypothesis)* against the *null hypothesis*. If the data-set supports the alternative hypothesis, the null hypothesis is rejected. The *p*-value, for a given statistical model, represents the probability that when the null hypothesis is true, the statistical summary would be greater than or equal to the observed results [3]. In other words, the *p*-value represents the probability of obtaining an effect at least as extreme as the one in your data-set, *assuming the null hypothesis is true*. This means that a *p*-value does *not* represent the validity of the alternative hypothesis, rather represents the null hypothesis' lack of validity. This is a subtle distinction, but nonetheless *p*-values are still a widely used metric to test for statistical significance and will be employed throughout each manuscript within this thesis. Often, 0.05 is defined as a significance threshold within medical physics literature, and more broadly within the clinical environment [3]. It is reiterated that reducing the results of an analysis to exclusively whether or not a threshold was passed can result in erroneous conclusions and should be avoided.

## 2.3 Receiver operating characteristic (ROC) curves

A receiver operating characteristic (ROC) curve is a visual representation of a model's diagnostic ability over a varying range of discriminative thresholds [4]. Mathematically, the *x*-axis represents the false positive rate (FPR) while the *y*-axis represents the true positive rate (TPR). Alternatively, the TPR is known as *sensitivity*. *Specificity* is defined as 1-FPR. Both of these metrics are often used to quantify the performance of a model in conjunction with the *area under the curve (AUC)*. The area under the curve is a measure of discrimination equal to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one (assuming normalized units). In other words, the AUC represents the ability for a model to properly classify input data. An AUC of 0.5 represents a completely random model, while an AUC of 1.0 represents a model with perfect prediction performance. These concepts are depicted in Figure 2.1. A represents a perfect model (AUC = 1), D represents a completely random model (AUC = 0.5) while B and C represent models of intermediate strength (AUCs ≈ 0.85 and 0.7 respectively).

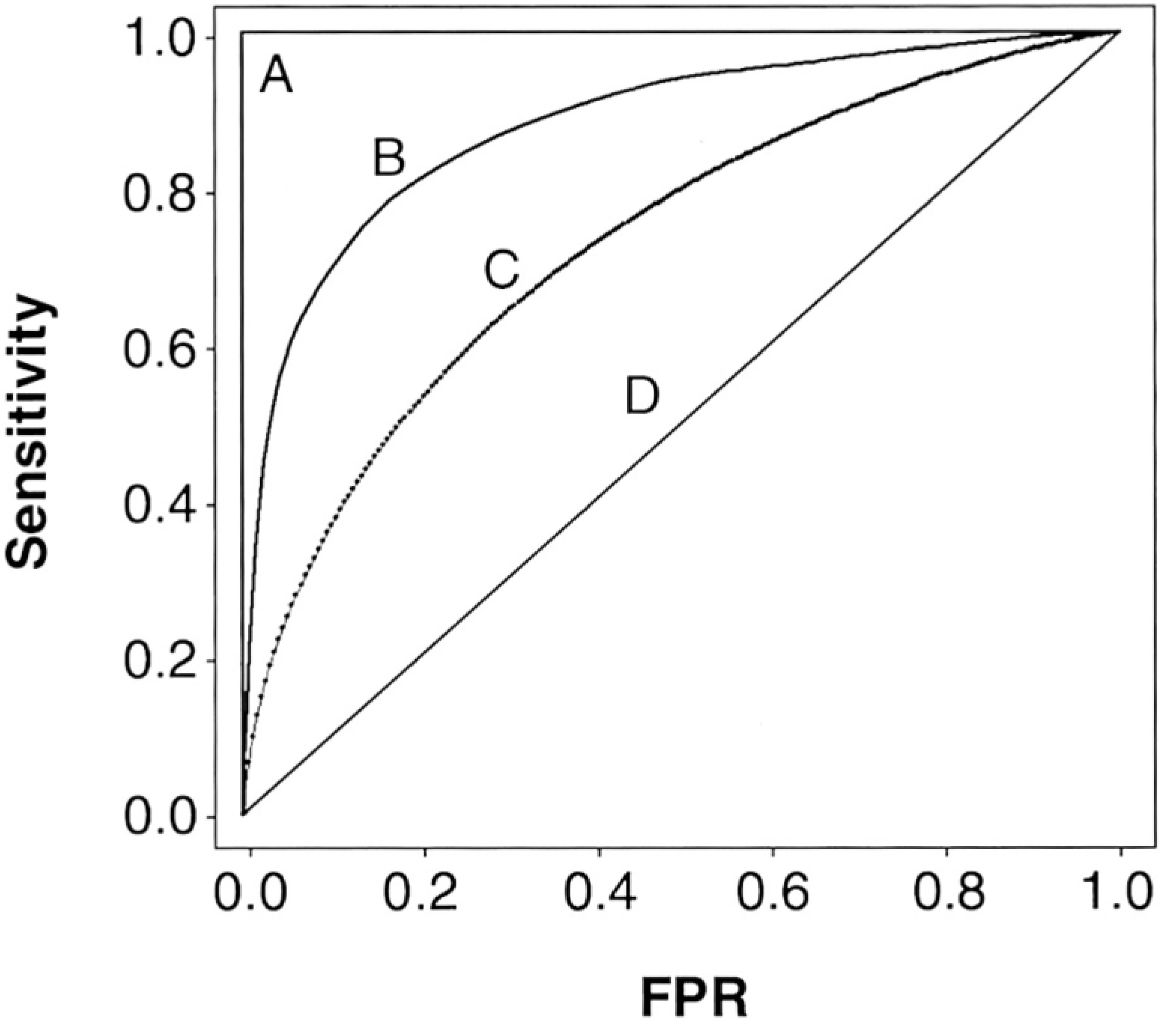


Figure 2.1 **Comparison of four receiver operating characteristic curves of differing quality.** A perfect curve (A) has an AUC of 1.0. The chance diagonal (D) has an AUC of 0.5 and represents a completely random test. Curves with some discrimination ability lie between the two extremes (B, C).

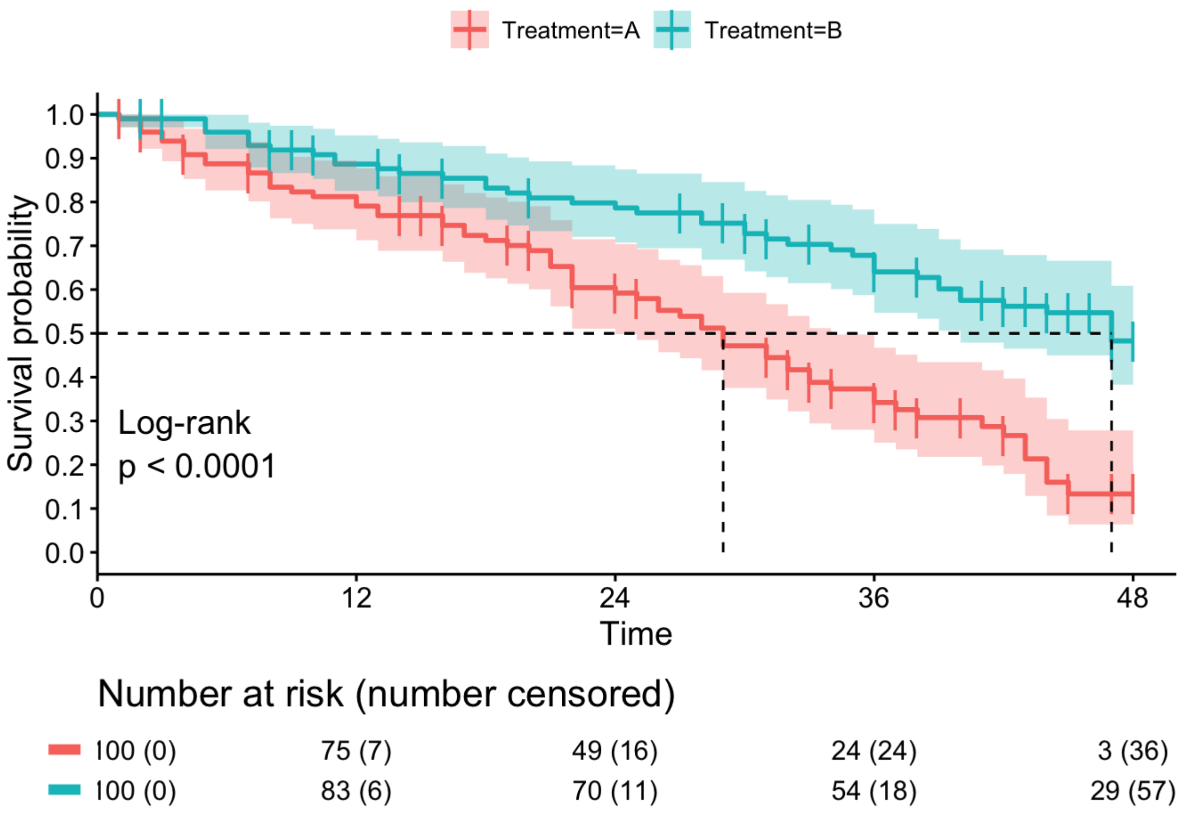
## 2.4 Kaplan-Meier (KM) analysis

Kaplan-Meier (KM) analyses are commonly used to assess the effect an intervention has on survival rates within a clinical trial [5, 6]. In particular, they're an excellent way of analyzing data whereupon subjects have differing survival & enrolment times, especially when some subjects drop out of the study early or are otherwise lost to follow-up. *Time-to-event* is an important concept, defined as starting when a subject receives a particular intervention (in our case, radiotherapy) and ending upon either an event (the outcome of interest) or dropping out of the study (*censored*). This time is also known as the *serial time*, in contrast to the *calendar time*, because two subjects can have an identical serial time even if their interventions occurred on different days.

The outcome of interest can be any clinical end point of interest (progression-free survival, etc...) however for the following example, survival will be used. A KM survival curve depicts the survival function of enrolled subjects over time. For each interval, the survival function is defined as 1 - number of events which occurred in the previous time interval divided by the number of patients at risk. The 'at risk' nomenclature is particularly important as it does *not* include patients who have died or are censored. Thus, the $y$-axis represents the estimated probability of survival for a hypothetical cohort (existing at a specific snapshot in time), *not* the actual overall % of patients who survived.

While valuable information can be gained from a single KM curve, often KM analysis involves the comparison of two groups (thus two curves). Typically, one curve represents the group which received an intervention (ex. a new drug on trial) while the other curve represents the control (ex. placebo). An example of this is shown in Figure 2.2. In this example, treatment B is significantly more effective at improving survival than treatment A.

As discussed in the preceding sections, a statistical test is required to quantify the difference between two curves. The *log-rank* test is widely used in clinical trials to establish the efficacy of a new drug/treatment in comparison to the 'control' standard-of-care. The test statistically evaluates whether the null hypothesis is true, i.e. whether the survival distribution is identical between the two populations (or curves). The log-rank test statistic is defined as [6, 8]:



Number at risk (number censored)

| | | | | | |
|---|---|---|---|---|---|
| 🟥 | 100 (0) | 75 (7) | 49 (16) | 24 (24) | 3 (36) |
| 🟦 | 100 (0) | 83 (6) | 70 (11) | 54 (18) | 29 (57) |

Figure 2.2 **Example of KM curve comparison.** *x* axis represents the serial time, *y*-axis represents the survival probability. Numbers below graph represent the number of patients still at risk each time point, accounting for the number of patients lost to follow-up (number censored) [7].

$$L = \frac{0_1 - {E_1}^2}{E_1} + \frac{0_2 - {E_2}^2}{E_2} \tag{2.5}$$

where $O_i$ represents the total number of observed events in the $i^{th}$ group and $E_i$ represents the expected number of events in the $i^{th}$ group. As the test is evaluating whether the null hypothesis is true, the expected number of events is calculated from the *entire population*, across both curves (as if they were considered the same distribution). Thus, $E_i$ represents the sum of expected events over time; each time an event is observed, the number of expected events also increases. Finally, a $\chi^2$ table (with 1 degree of freedom) is used to determine whether the log-rank test statistic is considered statistically significant and thus the null hypothesis would be rejected [8].

One limitation of the log-rank test is that it only can determine whether the difference in survival times across two groups is significant, but cannot account for other independent variables. An alternative method of comparing two survival distributions is by using the *hazard ratio* and the Cox proportional-hazards model (further explained in the subsequent section). The hazard ratio gives a relative event rate between the two groups (along with 95% confidence intervals), while the Cox model is able to account for other confounding variables.

## 2.5 Multivariable analysis

The methods described above (KM curves, ROC metrics) are examples of *univariable* analyses. A primary limitation is that they only consider the impact that a single variable has on the investigated outcome, while ignoring the potential impact of others. For example, a univariable analysis of the impact smoking has on coronary heart disease does not account for the fact that smokers are more likely to be male, live in poverty or have a sedentary lifestyle. Any or all of these factors could be *confounders*. Furthermore, univariable methods function best for categorical variables (sex, usage of drug, smoker/non-smoker, etc...) when a dataset can be split into two distinct groups (or curves, as seen above). When investigating continuous variables (age, gene expression, weight, etc...), the described approach is not applicable[1]. The *Cox proportional-hazards* model (hereafter referred to as the Cox model) addresses both these problems [9].

The Cox model simultaneously evaluates the impact that multiple variables (*covariates*) have on the occurrence of an event. Notably, rather than solely considering the binary occurrence of said event, the Cox model incorporates the time at which the event occurred. Typically, the Cox model is expressed as a *hazard function* at time $t$, $h(t)$ [9]:

$$h(t) = h_0(t) \times exp(b_1x_1 + b_2x_2 + ...) \tag{2.6}$$

where the $b_i$ coefficients represent the effect size (impact) of each covariate $x_i$ and $h_0(t)$ represents the baseline hazard. In essence, the Cox model represents a multiple linear regression of

[1] Short of applying a threshold, which is not ideal due to the amount of lost information.

the logarithm of the hazard of variables $x_i$.

Each quantity $exp^{b_i}$ represents the *hazard ratio* of the *i*'th covariate. A hazard ratio indicates whether the increase/decrease of a covariate results in an increase/decrease of the event hazard. Similar to the odds ratio, it is helpful to think of hazard ratios above 1 as being positively associated with the event probability (in our case, bad prognostic factor), while a hazard ratio below 1 is negatively associated with the event probability (in our studies, good prognostic factor).

# References

[1] V. Amrhein, S. Greenland, and B. McShane, "Scientists rise up against statistical significance," *Nature*, vol. 567, pp. 305–307, mar 2019.

[2] M. Szumilas, "Information Management for the Busy Practitioner Explaining Odds Ratios," Tech. Rep. 3, 2010.

[3] R. L. Wasserstein and N. A. Lazar, "The ASA Statement on p -Values: Context, Process, and Purpose," *The American Statistician*, vol. 70, pp. 129–133, apr 2016.

[4] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[5] J. T. Rich, J. G. Neely, R. C. Paniello, C. C. Voelker, B. Nussenbaum, and E. W. Wang, "A practical guide to understanding Kaplan-Meier curves," *Otolaryngology - Head and Neck Surgery*, vol. 143, pp. 331–336, sep 2010.

[6] M. K. Goel, P. Khanna, and J. Kishore, "Understanding survival analysis: Kaplan-Meier estimate.," *International journal of Ayurveda research*, vol. 1, pp. 274–8, oct 2010.

[7] R. Van Paemel, "Kaplan Meier curves - Towards Data Science," 2019.

[8] J. M. Bland and D. G. Altman, "The logrank test," *BMJ*, vol. 328, p. 1073, may 2004.

[9] M. H. Katz, "Multivariable Analysis: A Primer for Readers of Medical Research," *Annals of Internal Medicine*, vol. 138, pp. 644–650, apr 2003.

# Chapter 3

# Deep learning in medical physics

## 3.1 Introduction to deep learning

Deep learning is a niche application within the greater domain of machine learning that has rose to prominence in recent years. A form of artificial intelligence (AI), machine learning is ‘the ability for an AI system to acquire their own knowledge, by extracting patterns from raw data ’[1]. Deep learning, a subset of machine learning, uses combinations of artificial neural networks to learn from large amounts of data. There are a wide variety of machine learning algorithms, ranging from simple logistic regression to naive Bayes. These sorts of algorithms have proven successful at many decision-based tasks, such as whether to recommend Cesarean delivery (logistic regression) or separating spam email from regular email (Naive Bayes).

Logistic regression is a statistically based algorithm which assigns a probability to the occurrence of a particular binary outcome based on a number of input variables. Similar to linear regression, it involves the combination of multiple inputs and an output, however contrary to linear regression, the output is a probability. Specifically, $P(y=1|\mathbf{x})$: the probability that the outcome $y = 1$, given the input data $\mathbf{x}$. Instead of a linear combination, logistic regression uses the logistic function:

$$P(y=1|\mathbf{x}) = \frac{1}{1+\exp(\mathbf{w}^T\mathbf{x})} \tag{3.1}$$

where $\mathbf{w}^T$ represents the weight matrix (akin to linear regression), and $\mathbf{x}$ represents the matrix of input data.

This can be quite powerful and is widely used as a basic statistical tool, due to its simplicity and interpretability. One problem which algorithms of this class encounter is that they're heavily dependent on the *representation* of the data they're given. They explicitly require data in a particular format: a matrix of data, where each column represents a data instance and each row represents a variable. Furthermore, they require the user to have specific knowledge of the domain in order to convey certain pieces of relevant information to the algorithm prior to using it. Among other things, deep learning provides a way for the computer to break down complex concepts within a data-set into a simpler representation, which if desired could then be analyzed by a similar algorithm to those described above. Furthermore, some deep learning algorithms are able to function on a shockingly small amount of *a priori* information required, compared to a feature-learning approach where the user is required to pre-define relevant metrics [1]. In other words, some deep learning algorithms are capable of learning attributes of a data-set and conveying that information to the user, rather than the other way around.

The structure of this chapter is as follows. First, the theoretical concepts behind the building block of most commonly used deep learning algorithms, the *multi-layer perceptron* (or alternatively, the *neural network*) will be introduced. Next, the *convolutional neural network*, a derivative of the neural network which is remarkably powerful for computer vision will be described. The training & optimization process fundamentally required to use a deep learning algorithm will be explained. A literature review of concurrent deep learning research within the field of medical physics will be presented. Next, generative modeling will be introduced, along with a similar review of generative modeling as it applies to medical physics. Finally, a practical overview of how to build a deep learning model (specifically using CNNs) within a hospital setting will be presented.

## 3.2 Neural networks

It is difficult to describe a deep learning algorithm without using the term 'neural network' . A *neural network*, also known as a deep feed-forward network or a multi-layer perceptron (MLP) is the quintessential deep learning model. Nearly all algorithms that fall under the deep learning umbrella either involve a/multiple neural network(s), concepts adopted from neural networks or some derivative of a neural network [1]. Thus, before going further, we must understand the neural network.

A neural network is composed of multiple *layers*, each of which contain multiple *nodes*, or *neurons*. Each neuron represents the flow of information from one layer to the next. Mathematically, this is represented by a weighted sum of its inputs followed by some non-linear *activation function* (more on this later) resulting in a single output value. This is shown in the following equations:

$$z = w_1x_1 + w_2x_2 + ... + w_nx_n = \mathbf{w}^t * \mathbf{x} \tag{3.2}$$

$$h_w(\mathbf{x}) = \mathscr{H}(z) = \mathscr{H}(\mathbf{w}^t * \mathbf{x}) \tag{3.3}$$

$$\mathscr{H}(z) = \begin{cases} 0 & \text{if } z < 0 \\ 1 & \text{if } z \geq 0 \end{cases} \tag{3.4}$$

where $x_i$ is the $i^{th}$ input into a neuron, $w_i$ is the $i^{th}$ (learned) weight of a neuron, $h_w(\mathbf{x})$ represents the output of a neuron and the Heaviside function ($\mathscr{H}$) is used as an example non-linear activation function.

The example neuron shown above (making use of a step function) can be used for simple linear binary classification. Evidently, it's just computing a linear combination of the inputs and applying a threshold. If the result exceeds a threshold, it outputs a positive class or else outputs the negative class.

While the above example can be used for simple applications, it is still incredibly primitive. However, the neuron is a very powerful building block as we can begin to combine multiple neurons both horizontally and vertically. A 'horizontal' combination of neurons represents a number of neurons all of which have the same input(s), however could have different learned weights. This is called a *layer*. A single layer of neurons is called a *perceptron*, while multiple layers stacked on top of one another ('vertical' combination whereupon the inputs of one layer are the outputs of the preceding layer) is called a *multi-layer perceptron*. Layers which do not explicitly use external input data or output some form of user-desired classification are called *hidden layers*. This nomenclature represents the fact that the layer's representation of the data is never explicitly seen by the user, barring debugging processes or the pursuit of specific analytics. A multi-layer perceptron is visually depicted in Figure 3.1.

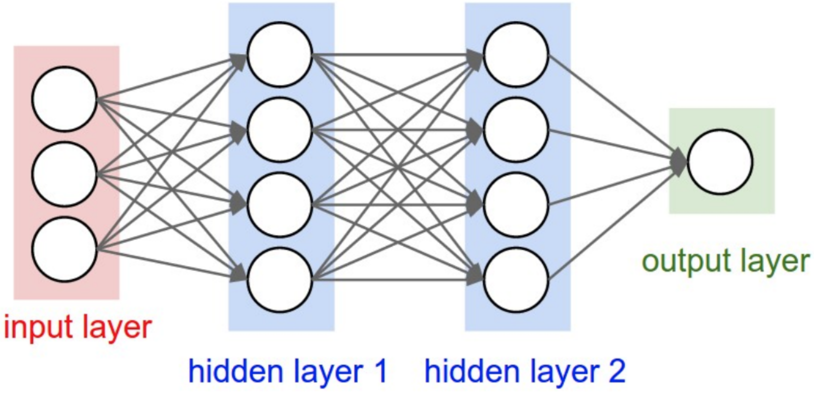


Figure 3.1 **A 3-layer neural network (or MLP) that consists of an input layer (3 nodes), 2 hidden layers (4 nodes each) and an output layer (1 node).** As indicated by the arrows between layers, each node is connected to every other node in the proceeding/preceding layer. This connection is mathematically represented by Equations 1, 2 and 3. Figure reproduced from CS231n - Stanford University [2]

In the example discussed, the activation function used was the Heaviside function. The presence of an activation function is crucial, as it provides the algorithm the ability to represent non-linear relations. Without this, a neural network would simply be a glorified linear regression algorithm. However, *any* non-linear function can be used as an activation function. In practice, the Heaviside function is rarely used. Currently, the most popular activation function is the rectified linear unit (ReLU) [1]. The ReLU computes the function $f(x) = \max(0,x)$. In other words, the activation is thresholded at zero for negative values. Although this seems like a very simplistic non-linear function, in practice it has proven to be adequate in allowing neural networks to far exceed the performance of many other machine learning algorithms.

## 3.3 Convolutional neural networks

So far, only neural networks have been discussed whereupon the input is a single vector which is subsequently transformed through numerous hidden layers. A *convolutional neural network*

(CNN) expands on this idea, by making the explicit assumption that the input data is an image [1]. While it is theoretically possible to flatten image data into an *x* by 1 vector and utilize a neural network, this does not scale well from a computational perspective and ultimately does not achieve good performance. For example, an image of size 200 × 200 × 3 (which doesn't come close to the size of the average medical image) would lead to every single neuron having 120,000 weights! This is very wasteful and often leads to over-fitting [2].

CNNs introduce spatial representation, by using *kernels* (or filters) with learned weights. This results in each neuron being connected to a small spatial region of the previous layer, rather than being fully connected to every neuron of the previous layer. This allows the model to learn spatial representation and drastically reduces the number of parameters. A visual representation of a CNN is shown in Figure 3.2.

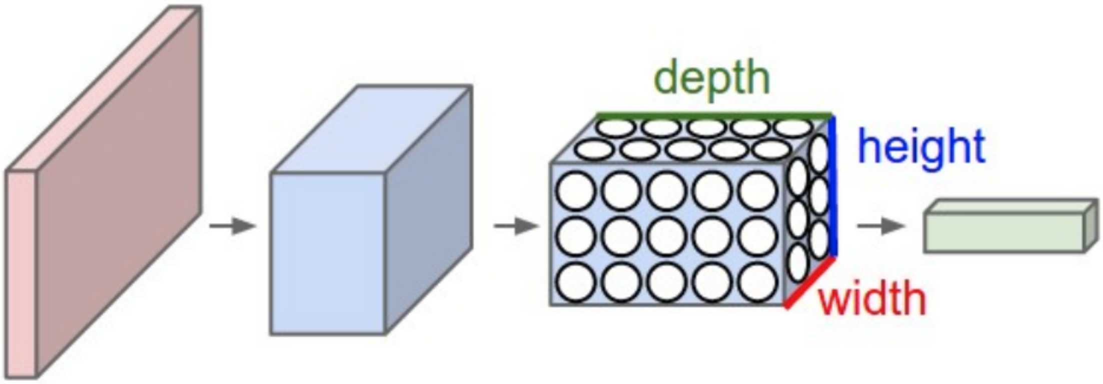


Figure 3.2 **Visual depiction of a CNN.** With each subsequent layer, the 3D input data (shown in red, notably has dimensions of *width* × *height* × *depth*) gets transformed into a smaller, but deeper representation which is used as input into the next layer. Figure reproduced from CS231n - Stanford University [2].

The most important layer of the CNN is the *convolutional* layer [2]. These layers compute the output of neurons, similar to a neural networks' layer, except they have multiple weights which define the aforementioned kernel. Recall that the CNN introduces spatial representation, by connecting each neuron to a small region of the preceding layer. Each kernel computes a dot product between the weights and this small region (called the *input volume*, typically 3×3, 5×5, etc...). This results in each layer reducing the width and height of the image, while increasing the depth (where the depth represents the number of kernels with differing weights). With stacked convolutional layers, the input image (say 512×512×1 for a single slice CT) is transformed into a much smaller, but deeper volume (say 32×32×12). Intuitively, each of these filters ($n$ = 12, in the prior example) will activate strongly when they see a particular feature within an input image. Often, in networks trained on photo-realistic images, this could be a sharp edge, or a patch of some specific color. When you have 100s of these filters working together, the network is able to analyze an input image and find relations that otherwise could be undetectable by more conventional algorithms.

Typically, after a number of convolutional layers, the final volume will be flattened into a single $x$ by 1 vector and input into a (or multiple) fully connected layer(s) (which are mathematically and functionally identical to the layers described in the basic neural network). This is the basic architecture used in Chapters 6 and 7 and results in the output of a single classification score. It is noted that other deep learning methodologies are able to output a continuous variable; the example described merely parallels the methodology used throughout Chapters 6 and 7.

## 3.4 Training & Optimization

Initially, all of the aforementioned weights (whether they are in a fully connected layer or in a convolutional layer) are either randomly initialized or sampled from some distribution[1]. Based on the performance of the model (as defined by a so-called loss/objective function), these weights are iteratively adjusted using a method further defined below. This is called *training* a neural network and involves a number of terms which must be defined to understand the optimization algorithms.

The *error* (or *loss/objective function)* is the mathematical quantity which defines how well the network is performing during the training. The entire goal of the optimizer is to minimize the error, and the weights are adjusted to achieve this task. Two of the most common error functions are mean squared error (for regression problems) and cross entropy (for classification problems). *Gradient descent* is the most commonly used optimization algorithm which iteratively moves

[1] Commonly used is the Xavier initializer. Put simply, it draws samples from a truncated normal distribution with its mean centered on 0 [1, 3].

the weights in the direction of steepest descent as defined by the negative of the gradient (with respect to the error). Calculating this gradient can be challenging, which led to the development of *backpropagation* [4]. By using the derivative chain rule, backpropagation computes the gradient of the error function with respect to the weights. Then, the weights can be updated according to the following equation:

$$\mathbf{W}(t+1) = \mathbf{W}(t) - \varepsilon \mathbf{G(W)} \tag{3.5}$$

where $\mathbf{W}(t)$ represents the weight vector at iteration $t$, the relaxation factor $\varepsilon$ represents the learning rate and $\mathbf{G(W)}$ represents the previously computed estimate of the gradient. The learning rate $\varepsilon$ is a *hyper-parameter* tweaked during the entire process, and as shown depicts the impact the gradient has on each iteration. Hyper-parameters represent user-defined parameters (often related to the training process) which can be modified in between each repetition of the training process[2].

Calculating the output of the model (*forward propagation)*, calculating the loss function, computing the gradient and updating the weights (according to Equation 3.5) makes up one iteration. By repeating this process 100s of times, the deep learning model is trained. Typically, other context-specific metrics such as accuracy, Dice similarity score, AUC, sensitivity or specificity are used post-training to assess the model's performance on an independent data set. In this work, the ROC affiliated metrics (AUC/specificity/sensitivity) are typically used.

## 3.5 Deep learning in medical physics

Deep learning has been applied extensively to nearly every STEM sub-field and medical physics is no exception. The majority of applications within medical physics fall into one of two sub-categories: medical imaging or radiation therapy. Machine learning algorithms have been used in computer-aided diagnosis (CAD) for decades, initially on chest radiographs and mammograms in the 1980s [5, 6]. Through the 90's, Shih-Chung *et al.* used a basic CNN for lung nodule detection [7], although the CNN was very shallow (2 layers) in comparison to modern architectures. Since then, a vast amount of research within either the medical imaging

[2] *Not* between each iteration within the training process, but rather between each set of iterations (could be thought of as a training 'experiment')

or radiation therapy sub-fields has made extensive use of deep learning algorithms. In particular, medical physics machine learning research has greatly benefited from the advent of the GPU and open software platforms (in part leading to the rise of deep learning) throughout the previous decade (visually represented in Figure 3.3).

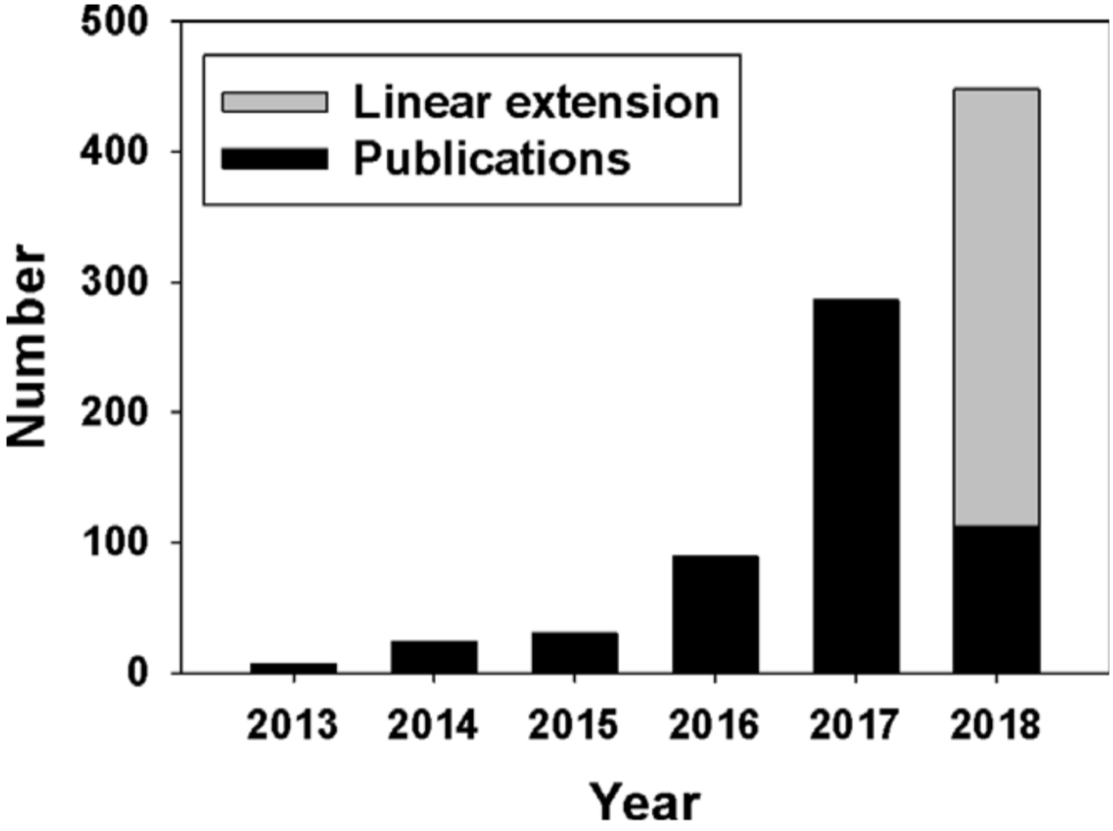


Figure 3.3 **Number of peer-reviewed publications in radiologic medical imaging that involved DL.** Image reproduced from Sahiner *et al.*[8]

Deep learning's applications within medical imaging can be further broken down into 3 categories: image segmentation (contouring), detection & characterization. Segmentation algorithms are often used to define organs of interest as Kovacs *et al.* did while achieving excellent performance in segmenting the whole lung on a cine MRI [9]. The other main class of segmentation algorithms are used to contour lesions; for example, Men *et al.* applied a deep de-convolutional neural network (DDNN) to a cohort of 230 nasopharyngeal cancer patients [10]. Detection algorithms are often similar to segmentation algorithms, however their main purpose is to *locate* an OAR within an image (either 2D or 3D). Typically, this will be used in tandem with a segmentation algorithm. Characterization algorithms often fall into two classes: diagnosis or prognosis (Chapters 6 and 7 involve a prognosis characterization algorithm). Computer-aided diagnosis (CADx) involves the input of some suspicious region, followed by the computer estimating the likelihood that said region is diseased or healthy. A very active area of diagnosis deep learning research (in part due to the more accessible images) uses mammograms to predict the presence of breast cancer [11, 12]. Finally, the idea behind prognostic characterization algorithms (specifically related to medical imaging) is to determine predictive image-based biomarkers, such as: size, shape, texture, kinetics, etc... These range from using an end-to-end CNN (such as in this work) to combining specifically engineered image features (‘radiomics’) [13, 14].

## 3.6 Generative modeling

One of the most discussed topics within the machine learning community at the moment is *generative modeling*. In 2014, Goodfellow et al. [15] proposed a novel type of generative model, known as the generative adversarial network (GAN). GANs are an elegant class of algorithms that perform exceptionally well *if* the user does not explicitly require an understanding of the data’s underlying probability distribution and is instead only interested in sampling high-quality data. GANs effectively learn how to model an input distribution by training two "adversarial" networks [15]. The *generator* continuously improves its ability to generate fake images that can fool the *discriminator*. Meanwhile, the *discriminator* is trained to distinguish between fake and real images. If properly trained, the generator will eventually be able to create images that the discriminator can not tell apart from real images. At this point, the discriminator can be discarded and the generator can be repeatedly used to generate new images.This is visually represented in Figure 3.4.

The primary challenge in the successful application of a GAN is the training process. In practice, it is very difficult to train the network to a point where it converges properly. This is in part due to the unique objective function needed to train a GAN. The mini-max objective function is:

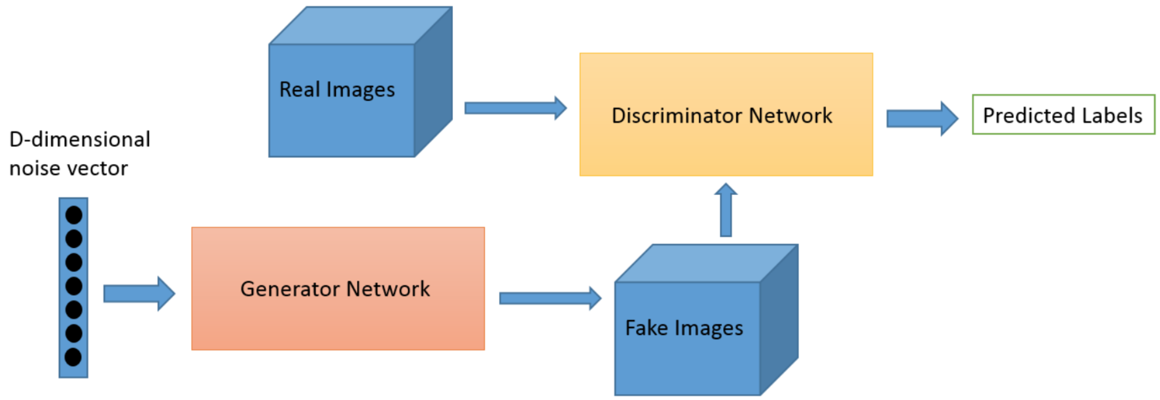


Figure 3.4 **Visual depiction of a basic generative adversarial network (GAN).** Predicted labels represent the discriminator's ability to classify an input as real or fake. Figure reproduced from 6.S191 - MIT [16].

$$\min_{\theta_g} \max_{\theta_d} [\mathbb{E}_{x \sim p_{data}} \log D_{\theta_d}(x) + \mathbb{E}_{z \sim p_z} \log(1 - D_{\theta_d}(G_{\theta_g}(z)))] \tag{3.6}$$

where $D_{\theta_d}(x)$ represents the output of the discriminator (predicted class) and $G_{\theta_g}(z)$ represents the output of the generator (fake image).

The training process ('mini-max game') consists of two alternating steps:

- Gradient ascent on discriminator:

$$\max_{\theta_d} [\mathbb{E}_{x \sim p_{data}} \log D_{\theta_d}(x) + \mathbb{E}_{z \sim p_z} \log(1 - D_{\theta_d}(G_{\theta_g}(z)))]$$

- Gradient descent on generator:

$$\min_{\theta_g} \mathbb{E}_{z \sim p_z} \log(1 - D_{\theta_d}(G_{\theta_g}(z)))$$

Intuitively, the first term of Equation 3.6 represents the ability for the discriminator to properly classify real images. The second term represents the ability for the discriminator to properly classify fake images (or alternatively, the ability for the generator to create images that successfully fool the discriminator). This is represented in the mini-max game, whereupon the first step improves the discriminator's ability and the second step improves the generator's ability.

While deep generative modeling is still relatively recent, there are a few notable applications within medical physics. Nearly all of these fall into the categories of data augmentation/synthesis, image reconstruction or dose prediction. A generative model can be used to create new image samples, thus increasing the training set size and reducing over-fitting. Cui *et al.* used an auto-encoder[3] based framework to simulate dynamic PET emission data then used in a reconstruction algorithm [18]. However, care must be taken as to avoid introducing unacceptable levels of bias in the generation process [19]. Emami *et al.* [20] used GANs to create synthetic CTs from magnetic resonance images in an attempt to facilitate the removal of CT-scans from the diagnostic workflow. Much of the current excitement around GANs is their ability to create incredibly realistic images, particularly of human faces [21]. However, their applications within healthcare could be of tremendous value, especially as the technology is further developed and validated. Mahmood et al. [22] (among others) have used a GAN to create a dose distribution solely from a planning CT. The immediately obvious value from using such a generative model is the time efficiency gained. A properly trained GAN can create a dose distribution (and subsequent treatment plan) on the order of milliseconds, while the current software/workflow used in the clinic typically takes on the order of hours. Furthermore, creating a treatment plan using current software introduces variability and extra planning time due to the individual expertise required in choosing beam angles, etc... As long as the GAN-created dose distribution obeys the prescribed dose constraints, it is reasonable to believe that this sort of automated method will be widely accepted in the future.

Each of these examples highlight the impact and potential that deep generative modeling has on the medical physics community. A more speculative discussion as to how generative modeling could specifically aid in outcome prediction will be presented in Chapter 8.

[3] Auto-encoders compress an input into a smaller representation, which can ultimately be sampled from, resulting in previously unseen generated images. Further details are beyond the scope of this thesis, however the interested reader is referred to Atienza *et al.*[17].

## 3.7 Building a CNN within a hospital environment

The process between beginning with a hypothesis and ending with a deep learning network capable of performing outcome prediction involves numerous steps, which each involve concepts discussed throughout the preceding sections. In this section, these steps will be summarized and put into context with respect to the work performed in Chapters 6 and 7. It is noted that more manuscript-specific information will be presented in the Chapters themselves.

1. *Approval*: Formulate hypothesis and receive REB approval for the proposal.
2. *Data retrieval/curation*: Retrieve all the data from respective sources, and begin curation.
3. *Initial debugging*: Begin building code-base for network, and ensure data is of the correct form to be properly input.
4. *First training*: Once initial debugging is complete, code the desired framework and begin training.
5. *Hyper-parameter tweaking*: By training on one set (training set) and validating on another (validation set), tweak hyper-parameters and re-train to determine the final architecture.
6. *External testing*: Test the final architecture on an external data-set (which went through both *approval* and *data retrieval/curation*) to best estimate the performance.

Throughout the course of building outcome predictions models within a hospital setting, it is wise to always have some form of proposal within the *approval* process. Given it often takes months and can involve numerous levels of bureaucracy, it is common (and occurred over the course of this work!) for delays while waiting for REB approval. Once approval is granted, the data retrieval/curation process begins (described in Chapter 1.3). This will often require numerous queries of the respective data-base, given it is extremely unlikely that one does not accidentally omit the exportation of some relevant piece of information. However, by proper planning this delay can be reduced or avoided! Chapters 4 and 5 involved multiple trips to external institutions to export treatment plans and clinical information of the 422 patients involved. Chapters 6 and 7 primarily made use of a public data-set, however extensive data curation was still required for the images to be in a parseable format. By keeping in mind the schema of whichever deep learning packages are being used (Tensorflow/Keras/PyTorch, etc...), significant time can be saved during the data curation process. Particularly if caution is not taken, there will likely be a *initial debugging* step, whereupon the input data representation will need to be modified. For example, in this work (Chapter 6), the data curation process involved converting the DICOM CT images to PNGs (Portable Network Graphics), encoding the outcome (1/0) in the file title and separating them into folders based on their partitioning scheme (training/validation sets). Once the first three steps have been completed, building (and training) the network can begin.

The final three steps represent building (*first training*), validating (*hyperparameter tweaking*) and testing (*external testing*) a CNN. Unfortunately, the external testing step is not always possible, largely due to lack of data. In Chapters 6 and 7, cross-validation was employed to provide a performance estimate; explicit comparisons were made to the respective benchmark studies using the validation performance. An external data-set is currently being curated, specifically for usage in the manuscript under preparation (Chapter 7).

## References

[1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[2] Stanford University, “CS231n - Convolutional Neural Networks for Visual Recognition.”

[3] F. Chollet, “Keras,” 2015.

[4] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *Nature*, vol. 323, pp. 533–536, oct 1986.

[5] M. L. Giger, K. Doi, and H. Macmahon, “Image feature analysis and computer aided diagnosis in digital radiography. 3. Automated detection of nodules in peripheral lung fields,” *Medical Physics*, vol. 15, pp. 158–166, mar 1988.

[6] H. P. Chan, K. Doi, S. Galhotra, C. J. Vyborny, H. Macmahon, and P. M. Jokich, “Image feature analysis and computer aided diagnosis in digital radiography. I. Automated detection of microcalcifications in mammography,” *Medical Physics*, vol. 14, pp. 538–548, jul 1987.

[7] S. C. B. Lo, S. L. A. Lou, M. V. Chien, and S. K. Mun, "Artificial Convolution Neural Network Techniques and Applications for Lung Nodule Detection," *IEEE Transactions on Medical Imaging*, vol. 14, no. 4, pp. 711–718, 1995.

[8] B. Sahiner, A. Pezeshk, L. M. Hadjiiski, X. Wang, K. Drukker, K. H. Cha, R. M. Summers, and M. L. Giger, "Deep learning in medical imaging and radiation therapy," *Medical Physics*, vol. 46, pp. e1–e36, jan 2019.

[9] W. Kovacs, N. Hsieh, and H. Roth, "Holistic segmentation of the lung in cine MRI," *Journal of Medical Imaging*, vol. 4, p. 1, oct 2017.

[10] K. Men, X. Chen, Y. Zhang, T. Zhang, J. Dai, J. Yi, and Y. Li, "Deep deconvolutional neural network for target segmentation of nasopharyngeal cancer in planning computed tomography images," *Frontiers in Oncology*, vol. 7, dec 2017.

[11] B. Q. Huynh, H. Li, and M. L. Giger, "Digital mammographic tumor classification using transfer learning from deep convolutional neural networks," *Journal of Medical Imaging*, vol. 3, p. 034501, aug 2016.

[12] R. K. Samala, H. P. Chan, L. M. Hadjiiski, M. A. Helvie, K. H. Cha, and C. D. Richter, "Multi-task transfer learning deep convolutional neural network: Application to computer-aided diagnosis of breast cancer on mammograms," *Physics in Medicine and Biology*, vol. 62, pp. 8894–8908, nov 2017.

[13] P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R. G. Van Stiphout, P. Granton, C. M. Zegers, R. Gillies, R. Boellard, A. Dekker, and H. J. Aerts, "Radiomics: Extracting more information from medical images using advanced feature analysis," *European Journal of Cancer*, vol. 48, pp. 441–446, mar 2012.

[14] M. Vallières, E. Kay-Rivest, L. J. Perrin, X. Liem, C. Furstoss, H. J. Aerts, N. Khaouam, P. F. Nguyen-Tan, C. S. Wang, K. Sultanem, J. Seuntjens, and I. El Naqa, "Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer," *Scientific Reports*, vol. 7, no. 1, pp. 1–33, 2017.

[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," *NIPS Proceedings*, 2014.

[16] Massachusetts Institute of Technology, "MIT 6.S191: Introduction to Deep Learning."

[17] R. Atienza, *Advanced Deep Learning with Keras*. Packt Publishing, 2018.

[18] J. Cui, X. Liu, Y. Wang, and H. Liu, "Deep reconstruction model for dynamic PET images," *PLoS ONE*, vol. 12, sep 2017.

[19] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "GAN-based Synthetic Medical Image Augmentation for increased CNN Performance in Liver Lesion Classification," tech. rep., IEEE, 2018.

[20] H. Emami, M. Dong, S. P. Nejad-Davarani, and C. K. Glide-Hurst, "Generating synthetic CTs from magnetic resonance images using generative adversarial networks," *Medical Physics*, vol. 45, pp. 3627–3636, aug 2018.

[21] T. Karras and S. Laine, "A Style-Based Generator Architecture for Generative Adversarial Networks Timo Aila NVIDIA," tech. rep., NVIDIA, 2019.

[22] R. Mahmood, A. Babier, A. Mcniven, and T. C. Y. Chan, "Automated Treatment Planning in Radiation Therapy using Generative Adversarial Networks," tech. rep., 2018.

# Chapter 4

# Can dose outside the PTV influence the risk of distant metastases in stage I lung cancer patients treated with stereotactic body radiotherapy (SBRT)?

**André Diamant**, Avishek Chatterjee, Sergio Faria, Houda Bahig, Edith Filion, Issam El Naqa, Cliff Robinson, Hani Al Halabi & Jan Seuntjens

Article published in: *Radiotherapy & Oncology, vol. 128, no. 3, pp. 513-519, 2018.*

## 4.1 Preface

The first goal of this thesis was motivated by an unexpected observation during my MSc. research [1]. The previous research found an inverse correlation between tumor size and distant metastasis rates. This was seemingly contradictory, as conventional wisdom suggests that *larger* tumors should be associated with a poor prognosis, so we investigated this issue further. This chapter describes the correlation we found between the dose distribution *outside* the tumor and distant metastasis rates. Furthermore, it describes the correlation between tumor volume and this dose distribution, revealing the confounding variable which resulted in the

MSc thesis' observation. Ultimately this chapter presents a statistical outcome prediction model incorporating dose metrics outside of the tumor in non-small cell lung cancer patients treated with SBRT. The work suggests the dose received by the region close to the PTV may be directly relevant to the risk of distant metastases.

## 4.2 Abstract

In an era where little is known about the "abscopal" (out-of-the-field) effects of lung SBRT, we investigated correlations between the radiation dose proximally outside the PTV and the risk of cancer recurrence after SBRT in patients with primary stage I non-small cell lung cancer (NSCLC).

This study included 217 stage I NSCLC patients across 2 institutions who received SBRT. Correlations between clinical and dosimetric factors were investigated. The clinical factors considered were distant metastasis (DM), loco-regional control (LRC) and radiation pneumonitis (RP). The dose (converted to EQD2) delivered to regions of varying size directly outside of the PTV was computed. For each feature, area under the curve (AUC) and odds ratios with respect to the outcome parameters DM, LRC and RP were estimated; Kaplan-Meier (KM) analysis was also performed.

Thirty-seven (17%) patients developed DM after a median follow-up of 24 months. It was found that the mean dose delivered to a shell-shaped region of thickness 30 mm outside the PTV had an AUC of 0.82. Two years after treatment completion, the rate of DM in patients where the mean dose delivered to this region was higher than 20.8 $\mathrm{Gy}_2$ was 5% compared to 60% in those who received a dose lower than 20.8 $\mathrm{Gy}_2$. KM analysis resulted in a hazard ratio of 24.2 (95% CI: 10.7, 54.4); $p < 10^{-5}$. No correlations were found between any factor and either LRC or RP.

The results of this study suggest that the dose received by the region close to the PTV has a significant impact on the risk of distant metastases in stage I NSCLC patients treated with SBRT. If these results are independently confirmed, caution should be taken, particularly when a treatment plan results in a steep dose gradient extending outwards from the PTV.

## 4.3 Introduction

Lung cancer is both the leading cause of cancer deaths worldwide and one of the most frequent cancers [2]. Despite advances in cancer treatment, metastatic lung cancer is still related to a poor prognosis [3]. Cancer recurrence, including distant metastasis, is possibly due to microscopic disease extensions (MDE) of the primary lung tumor, although there has been little research done specifically on MDEs or their possible distribution to verify this hypothesis [4–9].

Stereotactic body radiation therapy (SBRT) is a technique used to deliver a highly accurate dose in a well-defined target volume [10, 11]. The localized nature of this technique is one of the primary advantages: it allows for a maximal dose to the tumor, while minimizing the dose to healthy tissue and thus the risk of treatment complications. However, one clinical difficulty when treating localized lung cancers with SBRT is precisely defining the gross tumor volume (GTV). This inaccuracy may increase the chances that microscopic disease outside of the GTV is not eradicated. Furthermore, as the precise location of MDEs are unknown, the requisite planning target volume/ clinical target volume (PTV/CTV) margins are unclear. If the area outside the tumor contains microscopic disease, minimizing the dose to a region that is overly constrained could increase the probability of microscopic cancer cells surviving after SBRT, increasing the risk of cancer recurrence. The conformal nature of SBRT amplifies this problem due to the sharp dose gradient outside of the PTV.

The present investigation was conceived from a previous study [1] that correlated biomarkers and dosimetric parameters in a small cohort of patients. The study found that target volume size was inversely correlated with distant metastasis. As this seemed counter-intuitive, we investigated further, eventually leading to an analysis of the dose falloff around the PTV. Thus, the aim of the present study was to determine if there is a correlation between the radiation dose immediately outside the PTV and cancer recurrence in patients with stage I non-small cell lung cancer (NSCLC) treated with SBRT.

## 4.4 Methods

### 4.4.1 Patient selection

Two hundred and seventeen patients with primary stage I NSCLC treated using SBRT (either 3D-CRT or VMAT) between 2011 and 2015 were included in this study. 44% of patients were treated by the McGill University Health Centre (MUHC), the remainder were treated at the Centre Hospitalier de l'Université de Montréal (CHUM). Diagnosis of NSCLC was confirmed through histology in 139/217 (64%) of patients. The remainder (36%) had no biopsy due to either refusal or risk considerations and the diagnosis of primary NSCLC was assumed based on clinical history and PET-CT images. 195/217 of all patients (90%) received a PET-CT scan, which was used for staging. None of the patients received any form of chemotherapy or alternative cancer treatment. All patients had only one well-defined lung tumor and did not have a prior cancer within the past 5 years.

### 4.4.2 Follow-up and clinical factors

The clinical factors considered for this study were distant metastasis (DM), loco-regional control (LRC) and radiation pneumonitis (RP) (grade $\geq 3$ based on Common Terminology Criteria for Adverse Events (CTCAE) v4.0 [12]). Follow-up was performed 1-3 months after treatment completion and then every 3-6 months. Only patients with at least 12 months of follow-up data were included in the analysis. A patient was considered to have LRC if they had a radiographic response to treatment on CT images and no progression of the tumor was seen in the CT or PET scans done at each follow-up visit. If evidence of progression was observed at any point, the patient was considered to have loco-regional failure (LRF).

### 4.4.3 Sub-cohort selection

To confirm the significance of the observations, analyses were performed on sub-cohorts of patients with an additional set of selection criteria as shown:

- Sub-cohort A: The group containing 139 patients whose histology was confirmed by biopsy, 28 of whom developed distant metastasis (20%). This group was chosen to determine if the lack of biopsy for some patients influenced the results.

- Sub-cohort B: The group of patients who had a more sharply defined follow-up time, between the first and third quartile, i.e., month 14 - month 32 (23 metastasis events in 122 patients, 19%). This group was chosen to determine if the wide range of follow-up times influenced the results.

- Sub-cohort C: The group of patients who had a PTV volume between the first and third quartile, i.e., 16 cc - 40 cc (16/109, 15%). This group was chosen to determine if the wide range of PTV volumes influenced the results.

- Sub-cohort D: The group of patients who received one of the majority fractionation schemes, either 48 Gy/3 fractions or 60 Gy/3 fractions (27/159, 17%). This group was chosen to determine if the variation in fractionation schemes influenced the results.

- Sub-cohort E: The group of patients treated at MUHC (19 metastasis events in 96 patients, 20%).

- Sub-cohort F: The group of patients treated at CHUM (18 metastasis events in 121 patients, 15%). This group and sub-cohort E were chosen to determine if the variation in modality (3D-CRT versus VMAT), contouring style and dose prescription (described below) influenced the results.

### 4.4.4 CT acquisition and tumor segmentation

Target delineation was performed on radiation therapy planning CTs using Eclipse (Varian Medical Systems, Palo Alto, CA, USA). The internal target volume (ITV) was drawn based on the inspiration, expiration and the maximum intensity projection (MIP) images obtained from the 4DCT taken in conjunction with the planning CT. The CTs were acquired per a standard scanning protocol with a resolution of $512 \times 512$ pixels and 3 mm slice thickness. The contours were drawn manually and individually verified by an expert radiation oncologist. The PTV was a 3-5 mm extension margin (due to institutional variability) to the ITV. The MUHC patients' prescription isodose surfaces were chosen such that 95% of the PTV was covered by the prescription dose and 99% of the PTV received at least 90% of the prescription dose with a 5 to 11-field 3D conformal technique using Novalis TX 6 MV photon beams (Varian Medical Systems, Palo Alto, CA, USA; Brainlab, AG, Munich, Germany). CHUM patients' prescription isodose surfaces were chosen such that 95% of the PTV was covered by the prescription dose while maintaining the requirement that the prescription isodose must be within 65% - 85% of the maximum dose. SBRT was delivered using RapidArc (Volumetric Modulated Arc Therapy) on a 6 MV photon beam (Varian Medical Systems, Palo Alto, CA, USA). Dose calculation was originally performed using the Analytical Anisotropic Algorithm (AAA) for all patients. The dose analysis was verified on recalculated Monte Carlo plans (EGSnrc, Ottawa, Canada, [13]). Image verification was performed prior to and during each treatment using cone-beam CT (CBCT).

### 4.4.5 Region of interest (ROI) creation

An algorithm to evaluate the dose parameters in a region of varying size outside of each patient's PTV was developed using Python's *pydicom* module [14]. The algorithm consisted of five primary steps, described in detail within the supplementary material:

1. Superimpose PTV point cloud onto dose grid, generate convex hull of PTV.
2. Grow the PTV point cloud in 3D space isotropically.
3. From the grown point cloud, generate a convex hull.
4. Exclusive OR (XOR) logic applied to both convex hulls, resulting in a shell-shaped region radially extending outwards from the PTV.
5. Analysis of dose applied to this region.

The algorithm generated two types of regions. $ROI_{cont}(x\ mm)$ represents a shell-shaped region including the entire volume up to $x$ mm outside of the PTV. $ROI_{diff}(x\ mm)$ represents a 1 mm thick shell-shaped region at a distance of $x$ mm away from the PTV boundary. $x$ was varied from 1 mm to 100 mm in 1 mm increments. This maximal range was chosen to be significantly beyond the range which is suggested to contain microscopic spread [7, 8, 15]. A realistic depiction of the algorithm is shown in the left side of Figure 4.1 (green inner volume represents PTV, black represents the boundary of $ROI_{cont}(30mm)$). A two-dimensional example of the ROIs created for a hypothetical circular tumor is shown in the right side of Figure 4.1.

The potential error due to the difference between the planned dose distribution and the delivered dose distribution due to patient setup error and/or movement was estimated as follows.

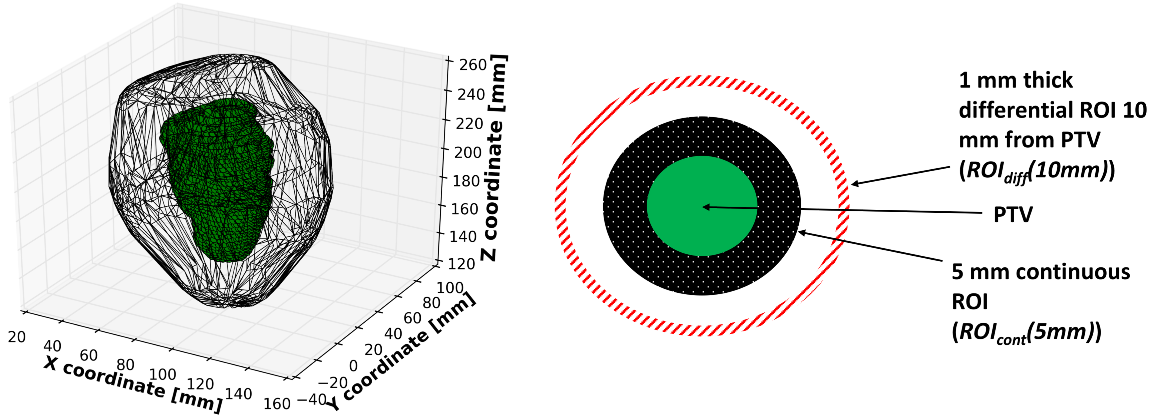


Figure 4.1 **Left side: Depiction of the ROI algorithm.** Inner green volume represents the PTV, black represents the boundary of the 30-mm thick shell-shaped region. **Right side: Two-dimensional example of the region of interest creation algorithm.** Shown are the PTV, $ROI_{cont}$(5 mm) (the continuous region up to 5 mm outside the PTV) and $ROI_{diff}$(10 mm) (1 mm region, 10 mm away from the PTV).

The ROI creation algorithm was rerun with the inclusion of a dose grid shift applied prior to superimposing the PTV point cloud onto the dose grid. For each patient a dose grid shift of one voxel along a random axis was applied. As the voxel spacing in all patients is greater than 2 mm, and the literature reports a maximum error of 2 mm [16–19], this method determines an upper limit uncertainty caused by the planned versus delivered dose discrepancy.

### 4.4.6 Parameters from the dose distribution

All doses were converted to the equivalent dose in 2 Gy fractions (EQD2) assuming an $\alpha/\beta$ of 10 [10, 20–22] in order to correct for biological effects between fractionation schemes. The dosimetric parameters computed for each ROI were the arithmetic mean dose delivered and the median dose delivered. The correlation coefficient between these parameters and the PTV volume was also computed. The homogeneity index (HI) for each patient was computed to analyze whether it had any correlation with the outcome. The HI was computed using the definition from the *International Commission on Radiation Units and Measurements* (ICRU) Report 83 [23–25]:

$$HI = \frac{D_2 - D_{98}}{D_{50}} \times 100\% \tag{4.1}$$

where $D_2$, $D_{98}$ and $D_{50}$ represent the near maximum, near minimum and median PTV dose respectively. A homogeneity index close to zero indicates a near homogeneous dose distribution. Additionally, the arithmetic mean dose to the PTV was computed.

### 4.4.7 Statistical Analysis

The mean dose fall-off for each patient, from ROI$_{diff}$(1mm) to ROI$_{diff}$(100mm) was computed. Each point along the $x$-axis thus represents the mean dose received by a 1 mm thick region $x$ mm away from the PTV. The patients were separated into two groups based on their treatment outcome. For each group, mean dose was averaged and plotted. The difference between the two curves was calculated and plotted.

All statistical analyses were done on Matlab R2017a software. The area under the curve (AUC) of Receiver Operating Characteristic (ROC) curves with respect to each clinical factor was calculated for each parameter. The odds ratio for each parameter was also computed using logistic regression from the *Dose Response Explorer System* (DREES) toolbox [26, 27]. 95% confidence intervals and a $p$-value (two tailed) for each odds ratio were additionally computed. The $p$-value significance threshold was adjusted by a factor of 4 to account for the testing of mean/median dose for both *ROI$_{diff}$(x mm)* and *ROI$_{cont}$(x mm)* (Bonferroni correction [28, 29]). Thus, a $p$-value below 0.01 was considered statistically significant. Factors with an AUC above 0.70, an odds ratio below 0.25 and a $p$-value below 0.01 were considered predictive. Additionally, a mean threshold dose was computed from the optimal operating point of the ROC curve and used to generate a Kaplan-Meier distant-metastasis-free survival curve. Hazard ratios and 95% confidence intervals were generated using Cox proportional hazard regression. Additionally, a multivariate logistic regression analysis of possible confounding factors including tumor location, tumor size (both 2D maximum tumor extension and 3D volume), cancer type, homogeneity index and fractionation scheme was performed.

## 4.5 Results

The follow-up time, defined as the time from treatment completion to last follow-up, had a median of 24 months (range: 12-36 months). DM was observed in 37 out of the 217 patients (17%). RP was observed in 18 patients (8%). LRF was observed in 26 patients (12%). Both LRF and DM was observed in 17 patients (8%). There was no correlation found between histology and any specific outcome. The site of the DM was quite diverse across the patient cohort (10 sites), however the most frequent DM sites were contralateral lung (35%), bone (19%) and brain (14%). The characteristics of the total patient cohort (with respect to DM) are summarized in Table 4.1 (range in parenthesis).

Table 4.1 **Patients and tumor characteristics**

| | DM | Did not develop DM | Percentage of total cohort [%] |
|---|---|---|---|
| **Total** | 37 | 180 | 100% |
| **Median Follow-up [months]** | 21 (3 - 56) | 24 (12 - 58) | |
| **Institution** | | | |
| McGill University Health Centre (MUHC) | 19 | 77 | 44% |
| Centre Hospitalier de l'Université de Montréal (CHUM) | 18 | 103 | 56% |
| **Median PTV volume** [cc$^3$] | 18.5 [5.7 - 51.5] | 37.4 [6.3 - 305.4] | |
| **Histology** | | | |
| NSCLC (verified by PET) | 11 | 67 | 36% |
| NSCLC (not otherwise specified) | 1 | 9 | 4% |
| Adenocarcinoma | 14 | 83 | 45% |
| Squamous Cell Carcinoma | 11 | 21 | 15% |
| **Dose fractionation** | | | |
| 34 Gy/1 fr | 1 | 1 | 1% |
| 34 Gy/2 fr | 0 | 1 | 1% |
| 45 Gy/5 fr | 1 | 0 | 1% |
| 48 Gy/3 fr | 16 | 58 | 34% |
| 48 Gy/4 fr | 7 | 33 | 18% |
| 50 Gy/5 fr | 1 | 13 | 6% |
| 54 Gy/3 fr | 2 | 0 | 1% |
| 60 Gy/3 fr | 9 | 74 | 38% |

The mean dose fall-off for each of the two patient sets (DM or no DM) is shown in Figure 4.2. The difference in mean dose fall-off between the two sets is also shown (inset). The red crosses (bottom curve) represent the patient group that developed DM, the blue circles represent the patient group that did not (top curve). The maximum difference between the groups was found to be 6.6 Gy$_2$ at a distance of 16 mm away from the PTV. Observing the point of inflection on the difference curve to be at 30 mm, it is postulated that $ROI_{cont}(30mm)$ represents the region most indicative of the risk of distant metastasis. A similar analysis for LRC and RP was found to show no statistical difference between their respective curves.

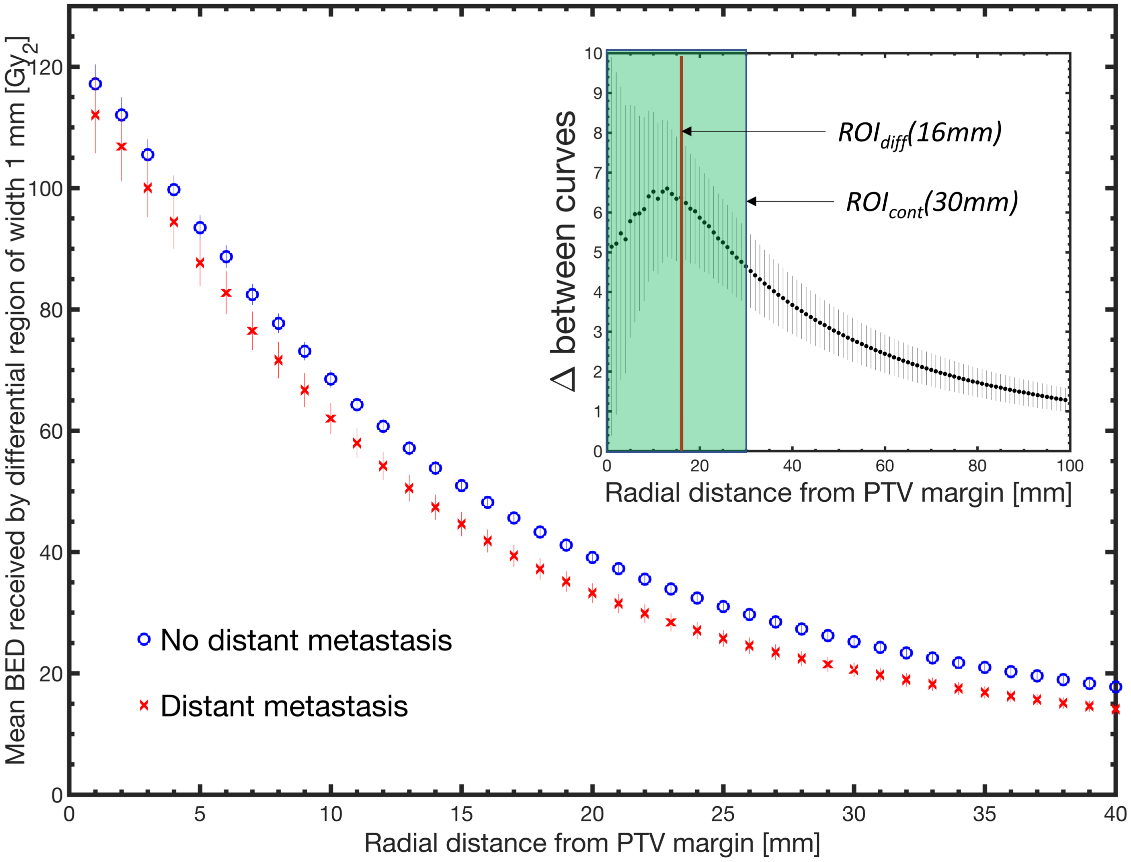


Figure 4.2 **Mean dose fall-off separated into patients whom developed distant metastasis (crosses) and those who did not (circles).** X-axis represents a 1 mm thick spherical shell *x* mm away from the PTV.

The mean dose received by a region of thickness 30 mm ($ROI_{cont}(30mm)$) was found to have an AUC of 0.81 and an odds ratio of 0.09 (95% CI: 0.03 - 0.24), $p$-value $< 10^{-5}$ with respect to DM. The threshold dose was computed from the optimal operating point of the ROC to be 20.8 Gy$_2$. Two years after treatment completion, the rate of DM in patients where the mean dose delivered to this region was higher than 20.8 Gy$_2$ was only 5%, compared to nearly 60% in those who received a dose lower than 20.8 Gy$_2$. The Kaplan-Meier DM-free survival curve separating patients whom received more than the threshold dose (blue) and less (red) is shown in Figure 4.3. The hazard ratio between the two curves was found to be 24.17 (95% CI:

10.74 - 54.36), $p$-value $< 10^{-5}$.

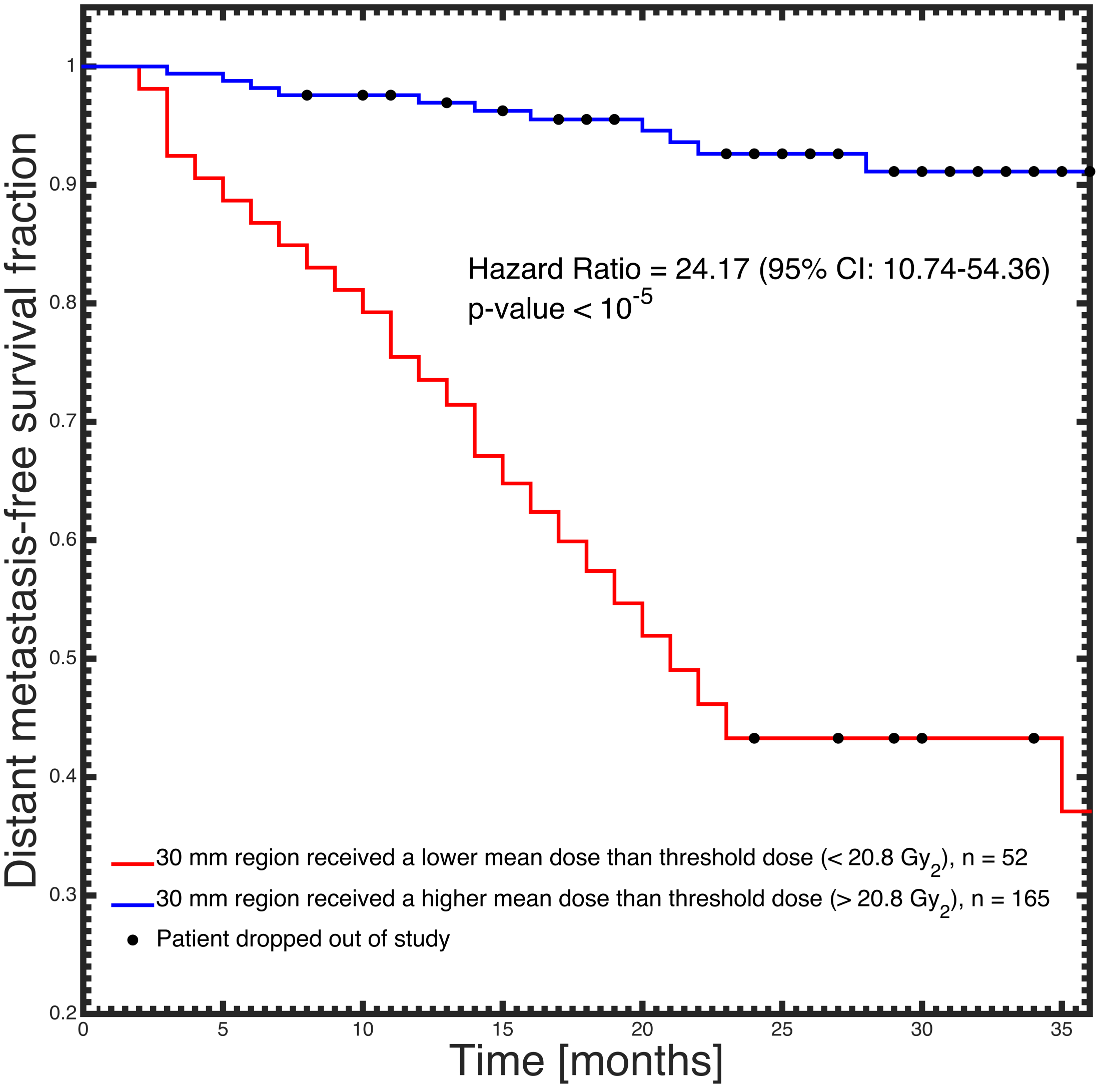


Figure 4.3 **Mean dose fall-off separated into patients whom developed distant metastasis (crosses) and those who did not (circles).** X-axis represents a 1 mm thick spherical shell $x$ mm away from the PTV.

Every other parameter from the dose distribution analyzed with respect to DM had an AUC greater than 0.70 and an odds ratio less than 0.20 ($p$-value $< 10^{-5}$), far exceeding the predictive benchmark. With respect to RP, no dosimetric parameter exhibited an AUC of over

0.55. Similarly, with respect to LRC, no dosimetric parameter exhibited an AUC of over 0.55. The uncertainty caused by potential differences between planned and delivered dose distributions due to patient setup and/or movement was found to be on the order of ± 0.01 for both the odds ratios and the AUCs and was therefore deemed negligible. A summary of the statistical computations with respect to DM performed on the total cohort (217 patients) is shown in Table 2. The correlation coefficient between tumor volume and the mean dose received by *$ROI_{cont}$(30mm)* was found to be 0.62, $p$-value $< 10^{-5}$. The mean dose received by *$ROI_{cont}$(30mm)* was found to not have any correlation with treatment technique/institution.

Table 4.2 **Summary of statistical analysis with respect to distant metastasis (total patient cohort).** Part A represents data extracted from the differential dose analysis, part B is data from the continuous region analysis.

| | AUC | OR [95% CI] | $p$-value |
|---|---|---|---|
| **A. Differential 1 mm thick region at distance *x*** [$ROI_{diff}(x)$] | | | |
| Mean dose delivered to $ROI_{diff}(16\ mm)$ | 0.80 | 0.09(0.04,0.25) | $< 10^{-5}$ |
| Median dose delivered to $ROI_{diff}(16\ mm)$ | 0.76 | 0.16(0.07,0.36) | $< 10^{-5}$ |
| **B. Continuous 30 mm region** [$ROI_{cont}(30\ mm)$] | | | |
| Mean dose delivered to $ROI_{cont}(30\ mm)$ | 0.81 | 0.09(0.03,0.24) | $< 10^{-5}$ |
| Median dose delivered to $ROI_{cont}(30\ mm)$ | 0.75 | 0.17(0.07,0.39) | $< 10^{-5}$ |
| **PTV volume** | 0.73 | / | / |
| **Homogeneity Index** | 0.51 | / | / |
| **Median dose delivered to PTV** | 0.55 | / | / |
| **Mean dose delivered to PTV** | 0.55 | / | / |

With respect to DM, the homogeneity index (HI) was found to have an AUC of 0.51. The mean and median dose to the PTV were both found to have an AUC of 0.55. Therefore, none of the dose coverage factors were deemed predictive.

Detailed results of the statistical computations performed on each patient sub-cohort are presented in the supplementary material. In summary, every dose parameter analyzed within each sub-cohort maintained its correlation with distant metastasis. The mean dose received by a region outside the PTV of thickness 30 mm (ROI$_{cont}$*(30mm)*) was found to have an AUC of 0.80, 0.84, 0.76, 0.83, 0.79 and 0.81 with respect to DM for sub-cohorts A, B, C, D, E and F respectively. Thus, the choice of institution (i.e. modality/prescription method) did not have a significant effect on the conclusions.

## 4.6 Discussion

The main objective of SBRT is local tumour control and there is a lack of literature specifically investigating distant metastasis and its potential risk factors. Nonetheless, our cohort had loco-regional failure rates (10%) and distant metastasis rates (20%) consistent with published literature [30–34]. This implies that our cohort was nothing out of the ordinary, but rather a typical collection of NSCLC patients. The "abscopal" effect or the "out-of-the-field effect" of radiation is thought to be related to immune modulation, but ultimately controversial [35]. Similarly, the interaction between the radiation dose received by the tumor and the potential MDEs outside of the PTV is still unclear. Our findings show a significant correlation between the dose just outside the PTV and the occurrence of distant metastasis in patients treated with SBRT for stage I NSCLC. Patients for whom the dose outside but proximal to the PTV was higher had less incidence of distant metastasis. Figure 2 represents the dose fall-off as we spatially move away from the PTV. It shows that the areas that received the largest difference in mean dose, when comparing patients who developed metastasis versus those who did not, are the most strongly correlated with the occurrence of distant metastasis. Thus, the observed correlation merits urgent scrutiny by independent researchers in prospectively designed clinical studies, and if verified, a change in dose prescription is warranted to potentially reduce the rate of DM. We suggest the addition of a secondary dose margin, on the order of a 20 mm extension, with a lower dose requirement (on the order of 21 $Gy_2$). This may minimize the risk of DM without increasing the risk of treatment complications. Multivariate analysis including dosimetric parameters, tumor location, tumor size including 2D maximum tumor extension and 3D volume, cancer type, homogeneity index and fractionation scheme was performed and did not confer any extra information or reveal any confounding factors.

Microscopic disease extensions (MDEs) have been reported at least 26 mm beyond the gross tumor edge [7]. Studies, performed after lobectomies, measured the extension of the microscopic disease outside the tumor [7, 8, 15]. Although results varied, in part due to the lack of ability to accurately count cells in 3-dimensions, the range of the maximum MDE distance was reported to be 8-26 mm [7, 8, 15, 36–38]. One could speculate that MDEs within this range were more efficiently killed in the group who received a higher dose of radiation to this region, decreasing the risk of spreading cancer cells and causing distant metastases. This is consistent with Figure 2 and part A of Table 2 which suggest that the region that would benefit most from dose escalation is at a distance of 16 mm ($ROI_{diff}(16mm)$) outside the PTV. This is additionally consistent with the shaded region of Figure 2 (inset) and part B of Table 2 showing that the dose delivered to a region of up to 30 mm outside the PTV ($ROI_{cont}(30\ mm)$) is significantly related to the occurrence of distant metastasis. The existence of MDEs suggest that a CTV may be required when treating stage I NSCLC patients with SBRT [19] and that escalating the dose to a region that extends 10-20 mm outside of the PTV may sterilize MDEs and decrease the risk of recurrence. The difference in the mean dose to the region outside the PTV between the two groups was found to be no more than 6.6 $Gy_2$, which corresponds to a dose escalation of less than 5% of the prescribed dose. No correlations were found between any of the dosimetric factors evaluated and the risk of radiation pneumonitis. This implies that dose escalation of this magnitude could be performed without adverse consequences.

There is a growing body of data suggesting that radiotherapy has an impact on the tumor's microenvironment even when radiation treatment is given in small fields such as the ones used in lung SBRT [35]. This impact is complex and not well understood. Radiation may cause not only cell death, but also changes in the phenotype of surviving cells and may make the tumor cells more sensitive to immune-mediated cell death [35]. Radiotherapy can also be immunosuppressive and we speculate as to whether this immunosuppression could have an influence on the spreading of microscopic tumor cells after SBRT. This could result in the phenomenon observed throughout this study, a correlation with distant metastasis without seeing any correlation with loco-regional control.

This study's conclusions apply to primary stage I NSCLC patients receiving SBRT without any systemic treatment. Statistical analysis performed on the sub-cohorts with more restrictive characteristics provided further confidence in the findings. Of particular interest, when analyzing a restricted range of PTV volumes (16 cc - 40 cc), the correlation between PTV volume and DM disappeared, while the correlation with the dose distribution outside the PTV remained. This implies that the dose falloff outside of the PTV is the causal factor, rather than the PTV volume. The choice of institution did not have a statistical impact on the findings, reinforcing that they are not biased by institution, but rather directly influenced by the dose distribution.

Consequently, the choice of technique (3D-CRT versus VMAT) did not impact the mean dose to the region investigated. Thus, no conclusions can be made as to whether one modality is superior.

Whether partially automated or entirely performed by a professional, the treatment planning process is designed to maximize the dose to the PTV while minimizing the dose to healthy tissue. There are several factors (patient size, tumor size/location/density/homogeneity, lung size/density) which will drastically impact the dose gradient outside the PTV. Consequentially, some treatment plans deliver a significantly lower dose to potential MDEs outside the PTV. In particular, this applies to patients with small tumors (less than $\sim$20 cc) who generally receive more conformal treatment plans in the interest of reducing treatment complications. Due to the correlation found between PTV volume and the mean dose to this region, the treatment plans of smaller tumors should be closely examined to ensure that this region is not coincidentally receiving an inadequate dose. In oligometastatic NSCLC patients, it is known that SBRT may increase the risk of developing new distant metastases [39, 40], but this is typically explained by the fact that patients with macroscopic oligometastases often also have occult microscopic tumor cell deposits at other sites [41]. In stage I NSCLC treated by SBRT, our finding of a correlation between the dose proximally outside the PTV and the occurrence of distant metastases must be both confirmed and better understood. The results of this study suggest that a secondary margin outside the PTV subject to a dose constraint of at least 20.8 $\mathrm{Gy}_2$ could drastically reduce distant metastasis rates without adverse consequence.

This study shows a significant correlation between the radiation dose outside of the PTV and the occurrence of distant metastases in stage I NSCLC patients treated with SBRT. Patients who received higher doses (escalation of less than 5% of the prescription dose) outside the PTV had less risk of developing distant metastases. If independently confirmed, this correlation suggests that current practice, particularly the inexistence of a CTV, may cause inadequate coverage in stage I NSCLC patients treated with SBRT. Dose escalation to an area beyond the PTV may be beneficial, ensuring that the probability of complete microscopic disease eradication is adequate.

**Conflicts of Interest:** None.

**Acknowledgments**: André Diamant acknowledges majority support from grant FRQNT-205163 and partial support from the grants NSERC CREATE 432290, CIHR FDN-143257 and MOP-114910. None of the funding sources had any role in the creation or execution of the study.

# References

[1] A. Diamant, “Modelling lung SBRT treatment outcomes using Bayesian network averaging,” 2016.

[2] Canadian Cancer Society, “What is non-small cell lung cancer?,” 2015.

[3] N. C. Institute, “Metastatic Cancer Fact Sheet - National Cancer Institute,” 2016.

[4] L. Kim, C. Wang, A. Khan, and M. Pierce, “CTV: The Third Front,” *International Journal of Radiation Oncology*Biology*Physics*, vol. 95, no. 2, pp. 800–801, 2016.

[5] M. Van Herk, “Errors and Margins in Radiotherapy,” *Seminars in Radiation Oncology*, vol. 14, no. 1, pp. 52–64, 2004.

[6] T. C. Mineo, V. Ambrogi, E. Pompeo, and A. Baldi, “Immunohistochemistry-detected microscopic tumor spread affects outcome in en-bloc resection for T3-chest wall lung cancer,” *European Journal of Cardio-thoracic Surgery*, vol. 31, no. 6, pp. 1120–1124, 2007.

[7] J. Van Loon, C. Siedschlag, J. Stroom, H. Blauwgeers, R. J. Van Suylen, J. Knegjens, M. Rossi, A. Van Baardwijk, L. Boersma, H. Klomp, W. Vogel, S. Burgers, and K. Gilhuijs, “Microscopic disease extension in three dimensions for non-small-cell lung cancer: Development of a prediction model using pathology-validated positron emission tomography and computed tomography features,” *International Journal of Radiation Oncology Biology Physics*, vol. 82, no. 1, pp. 448–456, 2012.

[8] F. J. Salguero, J. S. A. Belderbos, M. M. G. Rossi, J. L. G. Blaauwgeers, J. Stroom, and J. J. Sonke, “Microscopic disease extensions as a risk factor for loco-regional recurrence of NSCLC after SBRT,” *Radiotherapy and Oncology*, vol. 109, no. 1, pp. 26–31, 2013.

[9] X. Meng, X. Sun, D. Mu, L. Xing, L. Ma, B. Zhang, S. Zhao, G. Yang, F. M. Kong, and J. Yu, "Noninvasive evaluation of microscopic tumor extensions using standardized uptake value and metabolic tumor volume in non-small-cell lung cancer," *International Journal of Radiation Oncology Biology Physics*, vol. 82, no. 2, pp. 960–966, 2012.

[10] Y. Chang, Eric L., Nagata, *Stereotactic Body Radiation Therapy*. Tokyo: Springer, 2005.

[11] A. Amini, N. Yeh, L. E. Gaspar, B. Kavanagh, and S. D. Karam, "Stereotactic Body Radiation Therapy (SBRT) for lung cancer patients previously treated with conventional radiotherapy: a review," *Radiation Oncology*, vol. 9, no. 1, p. 210, 2014.

[12] National Institute of Health, "Common Terminology Criteria for Adverse Events v4.0 (CTCAE).," tech. rep., National Cancer Institute, 2010.

[13] I. Kawrakow, D. W. O. Rogers, F. Tessier, and B. R. B. Walters, "The EGSnrc Code System : Monte Carlo Simulation of Electron and Photon Transport," *NRCC Report*, pp. 2001–2015, 2016.

[14] Darcy Mason, "pydicom documentation."

[15] C. Siedschlag, L. Boersma, J. Van Loon, M. Rossi, A. Van Baardwijk, K. Gilhuijs, and J. Stroom, "The impact of microscopic disease on the tumor control probability in non-small-cell lung cancer," *Radiotherapy and Oncology*, vol. 100, no. 3, pp. 344–350, 2011.

[16] International Commission on Radiation Units and Measurements, "ICRU Report 62. Prescribing, Recording, and Reporting Photon Beam Therapy (Supplement to ICRU Report 50)," *Journal of ICRU*, vol. 74, no. November, pp. Ix +52, 1999.

[17] D. Jones, "ICRU Report 50—Prescribing, Recording and Reporting Photon Beam Therapy," *Medical Physics*, vol. 21, no. 6, pp. 833–834, 1994.

[18] S. H. Benedict, K. M. Yenice, D. Followill, J. M. Galvin, W. Hinson, B. Kavanagh, P. Keall, M. Lovelock, S. Meeks, L. Papiez, T. Purdie, R. Sadagopan, M. C. Schell, B. Salter, D. J. Schlesinger, A. S. Shiu, T. Solberg, D. Y. Song, V. Stieber, R. Timmerman, W. A. Tomé, D. Verellen, L. Wang, and F.-F. Yin, "Stereotactic body radiation therapy: The report of AAPM Task Group 101," *Medical Physics*, vol. 37, no. 8, pp. 4078–4101, 2010.

[19] The International Commission on Radiation Units and Measurements, “Prescribing, recording, and reporting of stereotactic treatments with small photon beams,” *Journal of the ICRU*, vol. 14, no. 2, 2014.

[20] E. J. Hall and S. Willson, *Radiobiology for the Radiologist*. Philadelphia: Wolters Kluwer, 2012.

[21] J. F. Fowler, “21 years of Biologically Effective Dose,” *The British Journal of Radiology*, vol. 83, no. 991, pp. 554–568, 2010.

[22] M. Baumann, M. Krause, J. Overgaard, J. Debus, S. M. Bentzen, J. Daartz, C. Richter, D. Zips, and T. Bortfeld, “Radiation oncology in the era of precision medicine,” *Nature Reviews Cancer*, vol. 16, no. 4, pp. 8–11, 2016.

[23] C. Program, F. Quantities, I. Quality, and R. Ion-beam, “ICRU 83: Prescribing, Recording, and Reporting Photon-Beam Intensity-Modulated Radiation Therapy (IMRT),” *Journal of the ICRU*, vol. 10, no. 1, pp. NP–NP, 2010.

[24] Myonggeun Yoon, Sung Yong Park, Dongho Shin, and Se Byeong Lee, “A new homogeneity index based on statistical analysis of the dose–volume histogram,” *Journal of Applied Clinical Medical Physics*, vol. 8, no. 2, pp. 9–17, 2007.

[25] T. Kataria, K. Sharma, V. Subramani, K. P. Karrthick, and S. S. Bisht, “Homogeneity Index: An objective tool for assessment of conformal radiation treatments.,” *Journal of medical physics / Association of Medical Physicists of India*, vol. 37, pp. 207–13, oct 2012.

[26] I. El Naqa, R. Li, and M. J. Murphy, *Machine Learning in Radiation Oncology*. Springer, 2015.

[27] I. El Naqa, S. L. Kerns, J. Coates, Y. Luo, C. Speers, C. M. L. West, B. S. Rosenstein, and R. K. Ten Haken, “Radiogenomics and radiotherapy response modeling,” *Physics in Medicine and Biology*, 2017.

[28] M. W. J. Layard, “Estimation of the Medians for Dependant Variables,” *The Annals of Mathematical Statistics*, vol. 30, no. 1, pp. 192–197, 1959.

[29] O. Jean, “Multiple Comparisons Among Means,” *Journal of the American Statistical Association*, vol. 56, no. 293, pp. 52–64, 1961.

[30] A. Van Baardwijk, W. A. Tomé, W. Van Elmpt, S. M. Bentzen, B. Reymen, R. Wanders, R. Houben, M. Öllers, P. Lambin, and D. De Ruysscher, “Is high-dose stereotactic body radiotherapy (SBRT) for stage i non-small cell lung cancer (NSCLC) overkill? A systematic review,” *Radiotherapy and Oncology*, vol. 105, no. 2, pp. 145–149, 2012.

[31] K. C. Wink, A. van Baardwijk, E. G. Troost, and D. De Ruysscher, “Nodal recurrence after stereotactic body radiotherapy for early stage non-small cell lung cancer: Incidence and proposed risk factors,” *Cancer Treatment Reviews*, vol. 56, pp. 8–15, 2017.

[32] M. Guckenberger, M. Allgäuer, S. Appold, K. Dieckmann, I. Ernst, U. Ganswindt, R. Holy, U. Nestle, M. Nevinny-Stickel, S. Semrau, F. Sterzing, A. Wittig, and N. Andratschke, “Safety and efficacy of stereotactic body radiotherapy for stage i non-small-cell lung cancer in routine clinical practice: A patterns-of-care and outcome analysis,” *Journal of Thoracic Oncology*, vol. 8, no. 8, pp. 1050–1058, 2013.

[33] J.-J. Hung, W.-J. Jeng, W.-H. Hsu, T.-Y. Chou, B.-S. Huang, and Y.-C. Wu, “Predictors of death, local recurrence, and distant metastasis in completely resected pathological stage-I non-small-cell lung cancer.,” *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer*, vol. 7, no. 7, pp. 1115–1123, 2012.

[34] M. Riihimäki, A. Hemminki, M. Fallah, H. Thomsen, K. Sundquist, J. Sundquist, and K. Hemminki, “Metastatic sites and survival in lung cancer,” *Lung Cancer*, vol. 86, no. 1, pp. 78–84, 2014.

[35] I. Popp, A. Ligia Grosu, G. Niedermann, and D. Duda, “Immune modulation by hypofractionated stereotactic radiation therapy: Therapeutic implications,” *Radiotherapy and Oncology*, vol. 120, pp. 185–194, aug 2016.

[36] S. Webb and A. E. Nahum, “A model for calculating tumour control probability in radiotherapy including the effects of inhomogeneous distributions of dose and clonogenic cell density,” *Physics in Medicine and Biology*, vol. 38, no. 6, p. 653, 1993.

[37] A. Takeda, N. Yokosuka, T. Ohashi, E. Kunieda, H. Fujii, Y. Aoki, N. Sanuki, N. Koike, and Y. Ozawa, "The maximum standardized uptake value (SUVmax) on FDG-PET is a strong predictor of local recurrence for localized non-small-cell lung cancer after stereotactic body radiotherapy (SBRT)," *Radiotherapy and Oncology*, vol. 101, no. 2, pp. 291–297, 2011.

[38] J.-J. Hung, W.-J. Jeng, W.-H. Hsu, T.-Y. Chou, B.-S. Huang, and Y.-C. Wu, "Predictors of death, local recurrence, and distant metastasis in completely resected pathological stage-I non-small-cell lung cancer.," *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer*, vol. 7, no. 7, pp. 1115–1123, 2012.

[39] K. E. Rusthoven, B. D. Kavanagh, S. H. Burri, C. Chen, H. Cardenes, M. A. Chidel, T. J. Pugh, M. Kane, L. E. Gaspar, and T. E. Schefter, "Multi-institutional phase I/II trial of stereotactic body radiation therapy for lung metastases.," *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, vol. 27, pp. 1579–84, apr 2009.

[40] K. E. Rusthoven, B. D. Kavanagh, H. Cardenes, V. W. Stieber, S. H. Burri, S. J. Feigenberg, M. A. Chidel, T. J. Pugh, W. Franklin, M. Kane, L. E. Gaspar, and T. E. Schefter, "Multi-Institutional Phase I/II Trial of Stereotactic Body Radiation Therapy for Liver Metastases," *Journal of Clinical Oncology*, vol. 27, pp. 1572–1578, apr 2009.

[41] R. D. Timmerman, C. S. Bizekis, H. I. Pass, Y. Fong, D. E. Dupuy, L. A. Dawson, and D. Lu, "Local Surgical, Ablative, and Radiation Treatment of Metastases," *CA Cancer J Clin*, vol. 59, pp. 145–170, 2009.

# Chapter 5

# Comparing local control and distant metastasis in NSCLC patients between CyberKnife and conventional SBRT

**André Diamant**, Veng Jean Heng, Avishek Chatterjee, Sergio Faria, Houda Bahig, Edith Filion, Robert Doucet, Farzin Khosrow-Khavar, Issam El Naqa & Jan Seuntjens

Article published in: *Radiotherapy & Oncology, vol. 144, pp. 201-208, 2020*

## 5.1 Preface

Following the previous study (Chapter 4), the next sub-objective of the thesis was determining whether other treatment modalities also resulted in a similar observation. CyberKnife was a particularly interesting option, due to the robotic nature of the CyberKnife delivery, allowing for extreme non-coplanarity and robust motion tracking. We hypothesized that these differences, among other things, could have a drastic impact on the dose distribution outside the PTV, a hypothesis validated from the results. Ultimately, this study investigates whether radiotherapy when delivered by CyberKnife is comparable to conventional SBRT with respect to distant recurrence or local control.

## 5.2 Abstract

Previous literature suggests that the dose proximally outside the PTV could have an impact on the incidence of distant metastasis (DM) after SBRT in stage I NSCLC patients. We investigated this observation (along with local failure) in deliveries made by different treatment modalities: robotic mounted linac SBRT (CyberKnife) vs conventional SBRT (VMAT/CRT).

This study included 422 stage I NSCLC patients from 2 institutions who received SBRT: 217 treated conventionally and 205 with CyberKnife. The dose behavior outside the PTV of both sub-cohorts were compared by analyzing the mean dose in continuous shells extending 1, 2, 3, ..., 100 mm from the PTV. Kaplan-Meier analysis was performed between the two sub-cohorts with respect to DM-free survival and local progression-free survival. A multivariable Cox proportional hazards model was fitted to the combined cohort (n = 422) with respect to DM incidence and local failure.

The shell-averaged dose fall-off beyond the PTV was found to be significantly more modest in CyberKnife plans than in conventional SBRT plans. In a 30 mm shell around the PTV, the mean dose delivered with CyberKnife (38.1 Gy) is significantly larger than with VMAT/CRT (22.8 Gy, $p < 10^{-8}$). For 95% of CyberKnife plans, this region receives a mean dose larger than the 21 Gy threshold dose discovered in our previous study. In contrast, this occurs for only 75% of VMAT/CRT plans. The DM-free survival of the entire CyberKnife cohort is superior to that of the 25% of VMAT/CRT patients receiving less than the threshold dose (VMAT/CRT$_{<21\mathrm{Gy}}$), with a hazard ratio of 5.3 (95% CI: 3.0 - 9.3, $p < 10^{-8}$). The 2 year DM-free survival rates were 87% (95% CI: 81% - 91%) and 44% (95% CI: 28% - 58%) for CyberKnife and the below-threshold dose conventional cohorts, respectively. A multivariable analysis of the combined cohort resulted in the confirmation that threshold dose was a significant predictor of DM (HR = 0.28, 95% CI: 0.15 - 0.55, $p < 10^{-3}$) when adjusted for other clinical factors. CyberKnife was also found to be superior to the entire VMAT/CRT with respect to local control (HR = 3.44, CI: 1.6 - 7.3). The 2-year local progression-free survival rates for the CyberKnife cohort and the VMAT/CRT cohort were 96% (95% CI: 92% - 98%) and 88% (95% CI: 82% - 92%) respectively.

In standard-of-care CyberKnife treatments, dose distributions that aid distant control are achieved 95% of the time. Although similar doses could be physically achieved by conventional

SBRT, this is not always the case with current prescription practices, resulting in worse DM outcomes for 25% of conventional SBRT patients. Furthermore, CyberKnife was found to provide superior local control compared to VMAT/CRT.

## 5.3 Introduction

A major cause of death related to cancer is the distant recurrence of the primary cancer [1]. Metastatic cancer is typically related to a poor prognosis in part due to the unpredictable nature of the disease [2]. This often leads to an inability to explicitly define the ideal treatment. Our recent study suggested that microscopic disease extensions (MDEs) of a primary non-small cell lung cancer (NSCLC) tumor could be related to distant metastasis [3], although there has been little research done which could explicitly confirm this hypothesis [4–9].

The aforementioned study [3] investigated whether the dose to the region *outside* of the planning target volume (PTV) could have any influence on distant metastasis (DM) incidence using a cohort of 217 NSCLC patients treated with stereotactic body radiation therapy (SBRT). We hypothesized that due to the uncertainty in predicting where MDEs may be present, the conformal nature of SBRT could potentially increase the risk of microscopic cancer cells surviving after treatment, thus increasing the incidence of DM. The study concluded that there was a threshold dose of approximately 21 Gy (equivalent dose in 2 Gy fractions) to the region extending 30 mm outwards from the PTV. After 2 years, patients who received more than this dose (75% of the cohort) had a DM incidence rate of just 5%, while patients who received less than this dose (25% of the cohort) had a DM incidence rate of nearly 60%.

CyberKnife (Accuray Incorporated, Sunnyvale, California, USA) consists of a lightweight linear accelerator mounted on a robotic system, giving it the ability to deliver highly non-coplanar dose distributions. It further differs from conventional SBRT modalities with its real-time image guidance system, which allows for more accurate radiation delivery [10]. Additionally, the constraints in prescription for a CyberKnife plan can differ from conventional SBRT plans. While there are a number of studies investigating CyberKnife treatment outcomes [11–15], to the authors' knowledge, none focus specifically on distant control. Furthermore, none investigated the dose outside of the PTV, but rather focused on more conventional variables, such as tumor size, prescription dose or dosimetric variables within the PTV. The present study first investigates the dose distribution outside the PTV as planned for and subsequently delivered by the CyberKnife. This is compared to the cohort investigated in our previous study, all of whom had SBRT delivered through either Volumetric Modulated Arc Therapy (VMAT) or conformal radiation therapy (CRT). A multivariable analysis was conducted to adjust for clinical factors that may act as confounders. We then investigated whether the difference in these modalities could have an impact on distant metastasis incidence or local control, and thus whether CyberKnife could be a superior choice to VMAT/CRT under current clinical practice. This study represents a novel comparison between treatment modalities, while nearly doubling the number of patients previously included in our previous study.

## 5.4 Methods

### 5.4.1 Patient Selection

This study included 422 patients, all of whom received SBRT between July 2009 and February 2015 for stage I NSCLC. The dataset was gathered from two institutions (McGill University Health Centre (MUHC), Centre Hospitalier de l'Université de Montréal (CHUM)) and three modalities (VMAT, CRT, CyberKnife). For the purpose of this study, the dataset was split into two distinct cohorts. The first cohort consisted of 217 patients treated with either CRT or VMAT (identical to the previous study [3]), while the second cohort consisted of 205 patients treated with CyberKnife. Histology was used to confirm diagnosis of NSCLC in 72% of patients. The remainder did not have a biopsy due to either refusal or risk considerations and thus diagnosis was assumed from CT/PET scans. None of the patients received any form of chemotherapy or alternative cancer treatment. The two cohorts were not necessarily treated by the same physicians, however the same physician group was involved in the determination of outcomes. Retrospective analyses were performed in accordance with the relevant guidelines and regulations as approved by the Research Ethics Committee of the McGill University Health Centre (MUHC) (Protocol numbers: MP-37-2011-936, 10-263-GEN).

### 5.4.2 Follow-up and clinical factors

The primary clinical endpoints considered throughout this study were distant metastasis (DM) and local failure (LF). Regional failure (RF), locoregional failure (LRF) and overall survival

(OS) were considered as secondary clinical endpoints. Follow-up of each patient was performed 1-3 months after treatment completion and then every 3-6 months. Any patient who did not have at least 12 months of follow-up information in conjunction with no cancer recurrence was excluded from the study. Follow-up time was defined as time from treatment completion to time of last follow-up.

### 5.4.3 CT acquisition, tumor segmentation and dose prescription

Comprehensive details regarding the VMAT/CRT cohort can be found in the Methods section of the previous study [3]. In summary, the CRT patients were all from MUHC and were delivered with 5 to 11 fields using Novalis TX 6 MV photon beams (Varian Medical Systems, Palo Alto, CA, USA; Brainlab, AG, Munich, Germany). The VMAT patients were treated at CHUM using RapidArc delivery of Clinac 6 MV photon beams (Varian Medical Systems, Palo Alto, CA, USA). Prior to November 2013, 169 patients in the CyberKnife cohort were treated on the CyberKnife G4 system at CHUM. The CyberKnife VSI system at the same institution was used to treat the remaining 36 patients of this cohort. No significant distinction was observed between the commissioning measurements of the two CyberKnife models. CT images were acquired for CyberKnife patients in supine position on a 512×512 grid with 1 mm slice thickness, on which contours were delineated by a senior radiation oncologist. Depending on the institution, the PTV is defined to be a 3-5 mm extension to the gross tumor volume. For VMAT and CyberKnife patients, the same requirement is followed to choose the prescription isodose surface: 95% of the PTV must be covered by the prescription dose. Several lung SBRT protocols (RTOG 0813 [16], RTOG 0915 [17]) have further restrictions on the dose fall-off outside the PTV. In particular, upper limits are set on the ratio of the 50% prescription isodose volume to the PTV volume. For CyberKnife plans, it was difficult to fulfill the target coverage goal without exceeding these fall-off limits. The latter were therefore not imposed upon the CyberKnife plans in this study. When necessary, respecting dose constraints to organs at risk was prioritized over target coverage. Either XSight Lung (n = 64), XSight Spine (n = 51) or the Synchrony Respiratory Tracking system (n = 90) were used in CyberKnife treatments depending on the tumor's position, size and motion. Cone beam CT (CBCT) was employed for every fraction for all VMAT/CRT patients. Neither gating nor breath-hold techniques were used. In the Cyberknife cohort, 185 patients had their treatment planning dose initially calculated using Accuray's Ray-Tracing algorithm while the remaining 20 were calculated with Monte

Carlo. However for lung SBRT plans, Ray-Tracing doses have been shown to be substantially inaccurate, with overestimation of the mean PTV dose on the order of 10 - 20% when compared with the more accurate Monte Carlo technique [18–20]. All VMAT and CRT patients had their treatment planning dose calculated with the Anisotropic Analytical Algorithm (AAA). Several studies have found that the discrepancy between AAA and Monte Carlo is much milder, with differences in the mean PTV dose on the order of 1 - 4% [21–24]. To allow for an accurate comparison, CyberKnife dose distributions were recalculated with an independent EGSnrc Monte Carlo model (Ottawa, Canada, [25]), while keeping all beam parameters constant [26].

### 5.4.4 Region of Interest (ROI) creation

The algorithm used to evaluate the dose parameters to the region outside the PTV (of varying size) is identical to that of the previous study [3]. In summary, the algorithm consists of five primary steps, described in detail within the Supplementary Materials:

- Superimpose PTV point cloud onto dose grid, generate convex hull
- Grow the PTV point cloud in 3D space isotropically.
- From the grown point cloud, generate a convex hull.
- Exclusive OR (XOR) logic applied to both convex hulls, resulting in a shell-shaped region radially extending outwards from the PTV.
- Analysis of dose applied to this region.

The ROI specifically considered throughout this study was $ROI_{cont}$(x mm), representing a shell-shaped region including the entire volume up to $x$ mm (excluding the PTV itself) outside the PTV boundary. The volume was truncated if it crossed the lung boundary.

### 5.4.5 Parameters from the dose distribution

Prior to any analysis, all dose values were converted to the equivalent dose in 2 Gy fractions ($EQD2_{10}$) assuming an $\alpha/\beta$ of 10 in order to account for biological effects between the differing fractionation schemes [27–30]. $EQD2_{10}$ was specifically used for comparison purposes to the previous study [3]. The formula used for the calculation of $EQD2_{10}$ is as follows:

$$EQD2_{10} = N \times d \times \frac{d + \frac{\alpha}{\beta}}{2 + \frac{\alpha}{\beta}} \tag{5.1}$$

where $N$ is the number of fractions and $d$ is the dose per fraction.

As in the previous study [3], the arithmetic mean dose delivered to the ROIs was computed. A new parameter was calculated for this study; the fraction of the volume of the ROI which received at least $x$ Gy EQD$2_{10}$ dose. Henceforth, this parameter will be referred to as $V_x$, where $x$ represents the EQD$2_{10}$ dose of interest. Of particular interest to this study is $V_{TD}$, where $TD$ represents the threshold dose discovered in the previous study, 21 Gy [3].

### 5.4.6 Motion robustness

A unique feature of CyberKnife dose delivery is its target tracking system. Intrafraction motion is measured in real-time during the treatment and the beam positions are corrected accordingly. Thus, despite target motion due to patient breathing, the delivered dose distribution in the patient frame can be expected, in theory, to remain similar to the planning dose. In contrast, conventional SBRT delivery relies on the delineation of an internal target volume (ITV) that encompasses the GTV throughout all breathing phases. By optimizing the dose to the PTV, defined as an extension to the ITV, the target is thus ensured to be sufficiently covered. However, as no beam positioning correction is performed in this case, the overall delivered dose distribution inherently differs from the planning dose that is calculated on a static phantom. One may be concerned that this difference invalidates the comparison between CyberKnife and VMAT/CRT doses *beyond* the PTV. To investigate this possibility, the impact of intrafraction motion on the dose delivery was assessed using a subset of 18 VMAT plans with representative motion (0-12 mm in all directions). The intrafraction target motion is estimated from the difference between the inspiration and expiration CT images. For each plan, the treatment isocenter position was shifted by half of the maximum recorded displacement and the dose was recalculated using an independent Monte Carlo model, while keeping the patient phantom fixed[1]. This process was repeated for a displacement of equal magnitude in the opposite direction. This results in two shifted dose distributions. For each voxel of the dose grid, the arithmetic mean of the 2 shifted doses and the non-shifted dose was taken. This final

[1] 'Phantom' in this context is referring to voxelized representation of a CT image with inclusion of density and material information.

dose grid ('motion-averaged') represents an upper limit to the impact motion has on the dose distribution. The mean dose to the $ROI_{cont}$(x mm) was calculated for $x$ = 1 to $x$ = 100 mm using the motion-averaged dose distribution. For every point of the resulting motion-averaged fall-off curve, the difference with the non-shifted curve was computed, and averaged over the subset of plans. The magnitude of the difference induced in the VMAT/CRT fall-off curve when taking into account patient motion was then compared to the difference between the fall-off curve of VMAT/CRT and CyberKnife.

### 5.4.7 Statistical analysis

All statistical analysis was done using either Matlab 2018a software or Python 2.7. The mean dose fall-off for each patient was computed ($ROI_{cont}$(x mm)) from $x$ = 1 mm to $x$ = 100 mm. The fall-offs were then separated into the VMAT/CRT cohort and the CyberKnife cohort and averaged over all patients within the respective cohorts. This resulted in a comparison plot between the mean dose fall-off curve averaged over the VMAT/CRT patients and the mean dose fall-off curve averaged over the CyberKnife patients. 95% confidence intervals (of the mean) were calculated for each point along the curves. In an identical fashion, the average fall-off of $V_{TD}$ in both cohorts was computed and plotted.

To further visualize the comparison of the mean dose delivered to $ROI_{cont}$(30 mm) between the two cohorts, a histogram was generated with a bin width of 2 Gy. The value of $x$ was chosen to be 30 mm as suggested by the previous study to be the region most impacted by the delivered dose [3]. A Wilcoxon rank sum test was used to determine whether the distribution of mean dose to $ROI_{cont}$(30 mm) from either cohort differed in a statistically significant way (defined as $p < 0.05$).

Kaplan-Meier (KM) analysis of DM-free survival was performed to compare the CyberKnife cohort to the VMAT/CRT cohort. This analysis was repeated for the 25% of VMAT/CRT patients who received a mean dose to $ROI_{cont}$(30 mm) *less* than the threshold dose (VMAT/CRT$_{<21\text{Gy}}$). Hazard ratios (HR), 95% confidence intervals (95% CI) and $p$-values were generated from a univariable Cox proportional hazards regression. Both patient deaths and patients lost to follow-up were censored. As the competing risk of patient death was censored in the analysis, it is important to note that the cumulative incidence of DM as estimated by the KM curve may be overestimated. However the cause-specific hazard ratio between the two cohorts calculated from the Cox proportional hazards model remains valid in the presence of a competing risk [31]. This was repeated for local/regional failure and death as endpoints.

Finally, to adjust for potential confounders, a multivariable Cox proportional hazards model was fit to the combined cohort (n=422) using the following covariates: sex, age, staging, histology, PTV volume, prescription dose in $EQD2_{10}$, mean $EQD2_{10}$ to the PTV, treatment modality and a binary variable indicating whether the mean dose to the $ROI_{cont}$(30 mm) was larger than the threshold dose of 21 Gy. To account for histology being a non-binary categorical variable, dummy binary variables were created for each histology category while the unknown category, representing missing histology, was used as the reference and thus omitted. All continuous variables (age, PTV volume, prescription $EQD2_{10}$, mean $EQD2_{10}$ to the PTV) were modelled to follow a linear relationship with the logarithm of the hazard function. The analysis was performed separately using either DM, death or local/regional failure as clinical endpoints. The resulting hazard ratio, 95% confidence intervals and $p$-values were calculated for each covariate. Both the univariable and multivariable Cox regression were performed using the *lifelines* package (Python) [32].

## 5.5 Results

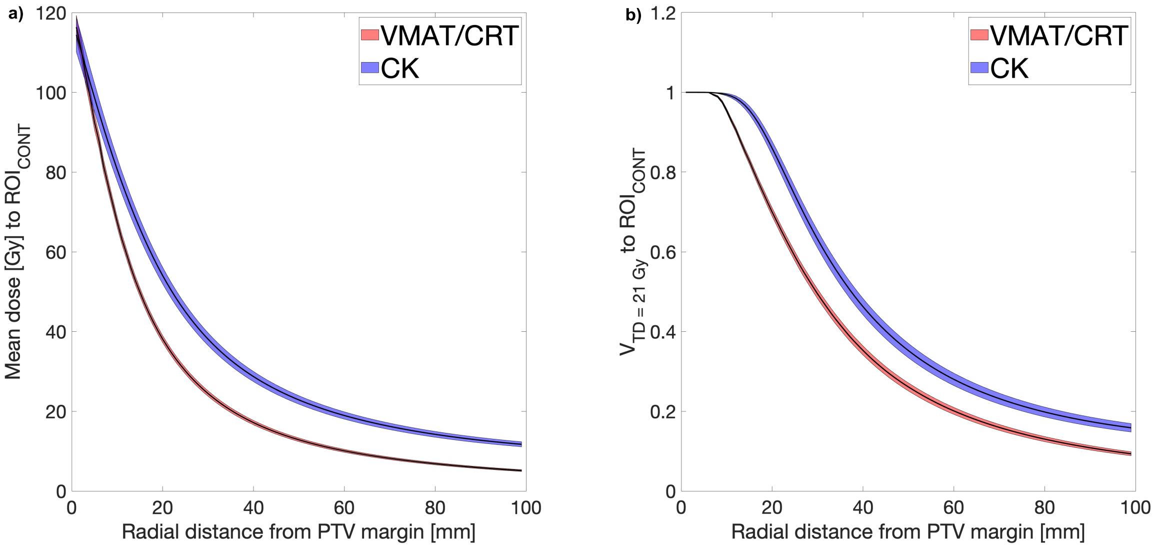
The median follow-up time was 24 months (range: 12 - 74 months). Distant metastasis was observed in 63 patients (15%). Local failure was observed in 37 patients (9%), while regional failure was observed in 28 patients (7%). Out of the 63 patients with DM, 12 (19%) also had local failure. Consistent with the previous study [3], the majority of DM sites were the contralateral lung (32%), brain (21%) and bone (17%). Detailed characteristics of both cohorts are shown in Table 5.1 (the VMAT/CRT cohort characteristics are identical to the previous study [3]).

The mean dose fall-off curves are shown in Figure 5.1a. The top curve (blue) represents the CyberKnife cohort while the bottom curve (red) represents the VMAT/CRT cohort. Figure 5.1b represents the fall-off of $V_{TD}$. Past $x$ = 10 mm, both sets of curves are statistically distinguishable at the 95% confidence level. One may note that the prescription $EQD2_{10}$ of VMAT/CRT patients was statistically significantly lower than that of CyberKnife patients: me-

Table 5.1 **Patients and tumor characteristics.** Percentages in square brackets represent proportion relative to the cohort size or median value. The standardized difference between the VMAT/CRT cohort vs. the CyberKnife cohort were calculated. The "not otherwise specified" histology is abbreviated as NOS.

| | VMAT/CRT cohort [3] | CyberKnife cohort | Percentage of total cohort [%] | Standardized difference |
|---|---|---|---|---|
| **Total** | 217 | 205 | 100% | |
| **Distant Metastasis Events** | 37 [17%] | 26 [13%] | 15% | 0.12 |
| **Local Failure Events** | 28 [13%] | 9 [4%] | 9% | 0.30 |
| **Regional Failure Events** | 17 [8%] | 11 [5%] | 7% | 0.10 |
| **Death Events** | 31 [14%] | 22 [11%] | 13% | 0.11 |
| **Institution** | | | | |
| MUHC | 96 | 0 | 23% | |
| CHUM | 121 | 205 | 77% | |
| **Tumor Stage** | | | | |
| T1 | 207 [95%] | 163 [80%] | 88% | 0.49 |
| T2 | 10 [5%] | 42 [20%] | 12% | -0.49 |
| **PTV volume range** [cm$^3$] | 5.7 - 305.4 [23.7] | 4.3 - 189.0 [22.5] | | 0.12 |
| **Age range** [years] | 46 - 93 [75] | 44 - 95 [76] | | -0.02 |
| **Follow-up range** [months] | 12 - 58 [24] | 12 - 74 [25] | | -0.14 |
| **Gender** | | | | |
| Male | 104 [48%] | 94 [46%] | 47% | 0.04 |
| Female | 113 [52%] | 111 [54%] | 53% | -0.04 |
| **Histology** | | | | |
| Unknown | 58 [27%] | 61 [30%] | 28% | -0.07 |
| NSCLC NOS | 20 [9%] | 17 [8%] | 9% | 0.03 |
| Adenocarcinoma | 96 [44%] | 66 [32%] | 38% | 0.25 |
| Squamous Cell Carcinoma | 41 [19%] | 53 [26%] | 23% | -0.17 |
| Large Cell Carcinoma | 2 [1%] | 8 [4%] | 2% | -0.20 |
| **Dose fractionation** | | | | |
| 48 Gy / 3 fr | 74 | 0 | 17% | |
| 48 Gy / 4 fr | 40 | 1 | 10% | |
| 50 Gy / 4 fr | 0 | 17 | 4% | |
| 50 Gy / 5 fr | 14 | 28 | 10% | |
| 60 Gy / 3 fr | 83 | 122 | 49% | |
| 60 Gy / 5 fr | 0 | 30 | 7% | |
| Other | 5 | 7 | 3% | |

dian prescription of 110 Gy vs. 150 Gy respectively (Wilcoxon rank sum test: $p$-value $= 0.006$, $\rho = 0.43$). However, after Monte Carlo recalculation of the CyberKnife plans, the mean-PTV dose of CyberKnife (median = 150 Gy) and VMAT/CRT patients (median = 137 Gy) were not found to be statistically distinguishable by a Wilcoxon rank sum test ($p$-value $= 0.65$, $\rho = 0.48$). No significant change to the fall-off curve past $x$ = 10 mm ($< 2\%$) was observed by taking into account the intrafraction motion and was therefore neglected. The impact of intrafraction motion on the fall-off curve is plotted for all $x$ in the Supplementary Material. The variability of the PTV margin (3 vs. 5 mm) was also considered. A 2 mm difference in the PTV margin translates to at most a 2 mm uncertainty in $x$ on the dose to the $ROI_{cont}$(x mm). Considering the magnitude of differences observed between fall-off curves between $x = 10$ mm and 100 mm, uncertainties on the order of 2 mm were not found to affect significance of the statistical results.



(a) (b)

Figure 5.1 **Dose fall-off comparison of the two cohorts considered within the study.** The black lines represent the mean value, while the shaded region represents the 95% confidence intervals. (a) represents mean dose [Gy] delivered to a region $x$ mm away from the boundary of the PTV while (b) represents the ratio of volumes that received a dose of at least 21 Gy (threshold dose). Both curves are statistically distinct past $x = 10$ mm.

A histogram representing the mean dose delivered to a continuous 30 mm shell outside the PTV ($ROI_{cont}$(30 mm)) is shown in Figure 5.2. The red bins represent the VMAT/CRT cohort with a median EQD2$_{10}$ of 22.8 Gy, while the blue bins represent the CyberKnife cohort with a median EQD2$_{10}$ of 38.1 Gy. The Wilcoxon rank sum test results in a significant difference between the two distributions with a $p$-value of $< 10^{-8}$ and $\rho = 0.13$. The vertical black line represents the previously determined threshold dose (21 Gy [3]), illustrating the small fraction of CyberKnife patients below said dose (5%), in comparison to the VMAT/CRT patients (25%).

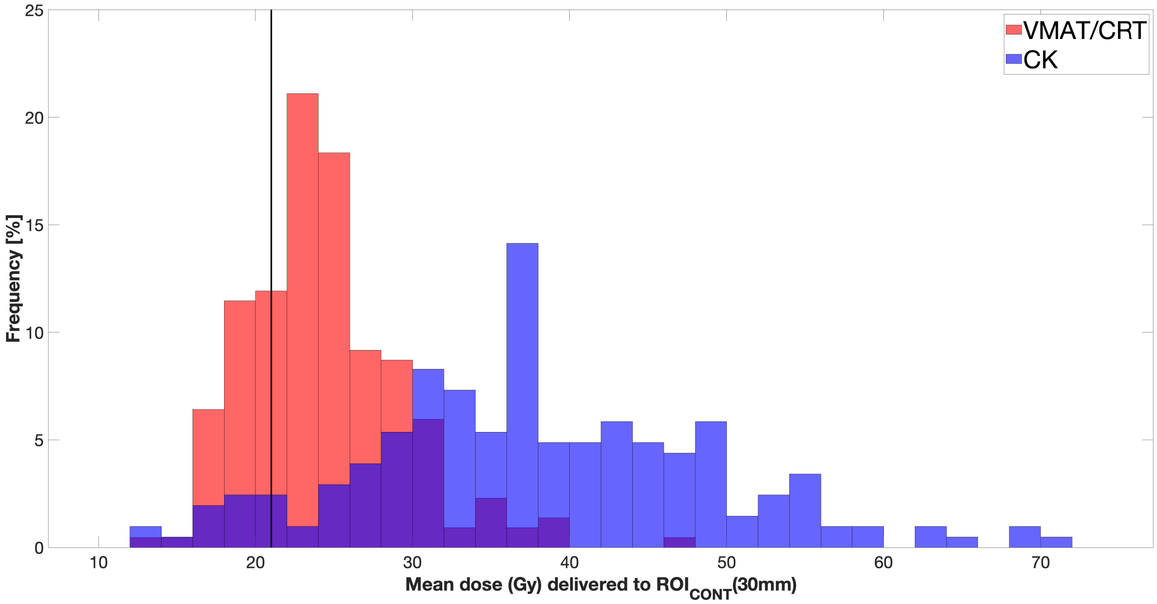
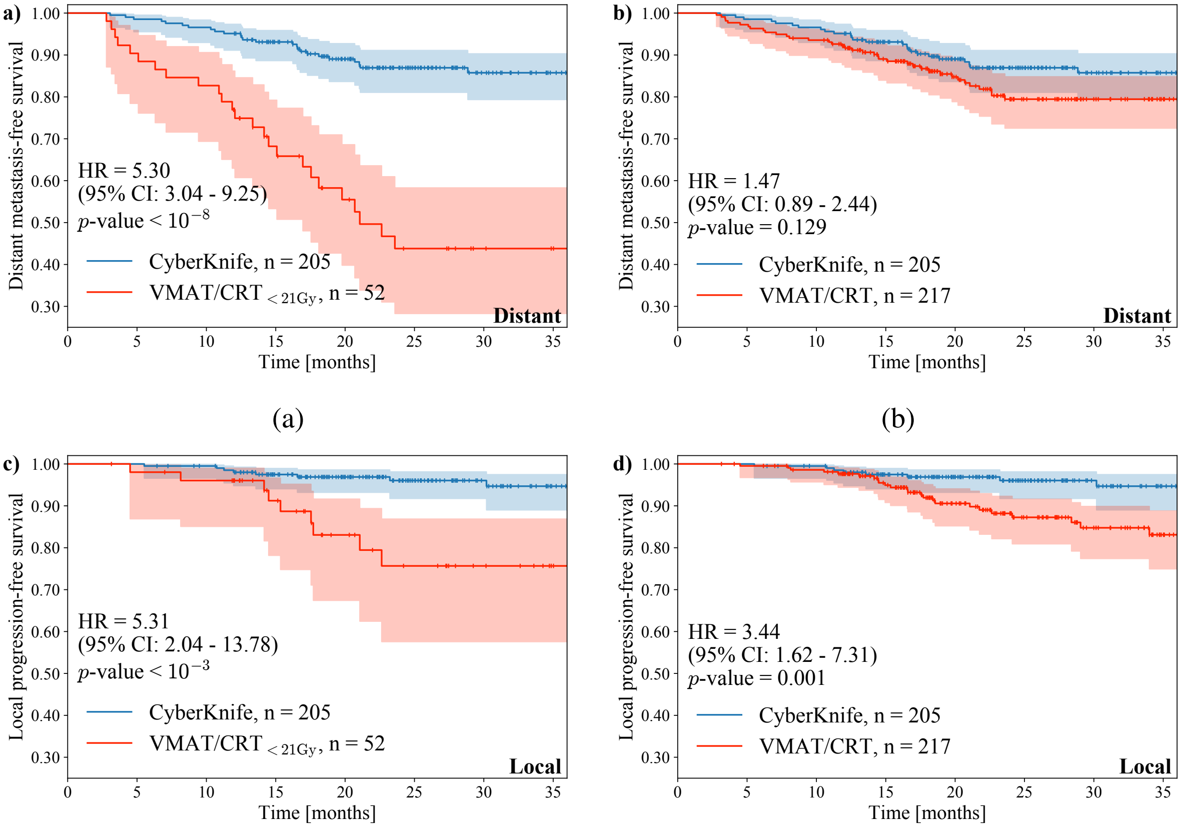


Figure 5.2 **Histogram of the mean dose (Gy) delivered to** $ROI_{CONT}$ **(30mm) across the two cohorts.** Vertical black line represents the threshold dose, 21 Gy. Purple shading indicates overlap of the two cohorts.

Kaplan-Meier analysis was performed to determine whether the CyberKnife delivery provided superior distant control rates in NSCLC patients. When comparing VMAT/CRT cohort (red) with the entire Cyberknife cohort (Figure 5.3b, there was a 47% increased risk of distant metastasis-free survival although with the confidence interval spanning the null: hazard ratio = 1.47 (95% CI: 0.89 - 2.44, *p*-value = 0.13). However, when comparing the entire CyberKnife cohort with the VMAT/CRT$_{<21\text{Gy}}$ subset in Figure 5.3a, patients in the VMAT/CRT$_{<21\text{Gy}}$ cohort were found to have a significantly higher hazard ratio of 5.30 (95% CI: 3.04 - 9.25, *p*-value $< 10^{-8}$). The 2 year DM-free survival rate is 87% (95% CI: 81% - 91%) for the CyberKnife cohort vs. 44% (95% CI: 28% - 58%) for the VMAT/CRT$_{<21\text{Gy}}$ subset. With respect to local progression-free survival, Figure 5.3d shows that patients treated with VMAT/CRT were found to have a higher risk of local failure, with a hazard ratio of 3.44 (95% CI: 1.62 - 7.31, *p*-value = 0.001). The 2 year local-progression free survival rates were 96% (95% CI: 92% - 98%) and 88% (95% CI: 82% - 92%) for CyberKnife and VMAT/CRT respectively. Patients in the VMAT/CRT$_{<21\text{Gy}}$ subset were similarly found to have even worse local outcomes, with a hazard ratio of 5.31 (95% CI: 2.04 - 13.8, $p$-value $< 10^{-3}$) when compared with the entire CyberKnife cohort (Figure 5.3c). We also performed KM analysis for regional failure, loco-regional failure and overall survival (KM curves in the Supplementary Material).



(a) (b)

(c) (d)

Figure 5.3 **Kaplan-Meier progression-free survival curves.** The shaded regions correspond to the 95% confidence band of their respective survival curves. Crosses represent censored datapoints. (a) and (b) represent distant metastasis-free survival curves, while (c) and (d) represent local progression-free survival curves. The top curves (blue) represent the entire CyberKnife cohort while the bottom curves (red) represent either the entire VMAT/CRT cohort (b,d) or the VMAT/CRT$_{<21\text{Gy}}$ cohort (a,c).

A multivariable Cox proportional hazards regression was performed on the combined cohort of CyberKnife and VMAT/CRT patients (n=422) using DM or local failure as clinical endpoints. The proportional hazards assumption was verified to be valid. After adjusting for other clinical factors, exceeding the threshold dose to the $ROI_{cont}$(30 mm) was found to be associated with a significantly lower risk of DM with a hazard ratio of 0.28 (95% CI: 0.15 - 0.55, $p$-value $< 10^{-3}$). With respect to DM, the hazard ratio was moderately elevated when comparing VMAT/CRT with Cyberknife (HR: 1.23, 95% CI: 0.70 - 2.16) although with a wide confidence interval spanning the null. With respect to LF, the modality choice of VMAT/CRT was associated with a significantly higher hazard ratio: 3.12 (95% CI: 1.42 - 6.85, $p$-value = 0.004). Exceeding the threshold dose was associated with a decreased risk of LF (HR: 0.68, 95% CI: 0.29 - 1.55) although with a wide confidence interval spanning the null. The hazard ratios for all other covariates are tabulated in the Supplementary Materials. The result of this analysis using RF, LRF and death as endpoints are also included in the Supplementary Materials.

As sensitivity analysis, continuous variables were modelled with restricted cubic splines with 3 knots to account for the non-linear relationship with the outcome [33]. Exceeding the threshold dose to the $ROI_{cont}$(30 mm) was still found to be associated with significantly reduced risk of DM, with a hazard ratio of 0.25 (95% CI: 0.13 - 0.47, $p$-value $< 10^{-4}$). VMAT/CRT treatments as compared to CK treatments were also found to be associated with a higher risk of LF, albeit with a lower HR: 2.31 (95% CI: 1.02 - 5.21, $p$-value = 0.04). As the prescription dose, a continuous variable, was significantly associated with LF (and not DM) when log-linearly modelled, this larger change in magnitude of LF-related HR as compared to DM is expected.

Adenocarcinomas have been shown to be more multi-focal, may have associated ground-glass opacities and are substantially harder to contour [34]. To determine whether this could be a confounder, a restricted analysis (n = 260) was performed by omitting patients with adenocarcinoma. This was done for both the univariable KM analysis and the multivariable Cox model. No difference in conclusions regarding the hazard ratios of the two above-mentioned covariates were observed. The restricted KM survival curves and a table of the hazard ratios for all covariates is included in the Supplementary Material.

A concern may be that the difference in prescription dose between the VMAT/CRT and Cyberknife cohort could be a confounder. Although the prescription $EQD2_{10}$ was explicitly accounted for in the multivariable analysis, a restricted analysis (n = 205) was also performed with only patients receiving a prescription of 60 Gy in 3 fractions in either cohort. No changes in the conclusions were found (detailed results in the Supplementary Material).

## 5.6 Discussion

Often the primary objective when employing SBRT in the treatment of lung cancer is local control [35]. This leads to a lack of literature specifically investigating distant control and its potential risk factors. The previous study by Diamant *et al.* [3] investigated whether the dose delivered to the region outside the PTV could have an impact on distant metastasis. We found that the dose to the region defined as $ROI_{cont}$(x mm) was strongly (inversely) correlated with the occurrence of distant metastasis. In particular, separating the patient cohort by a mean threshold dose of approximately 21 Gy to $ROI_{cont}$(30 mm) resulted in a hazard ratio of 24.2 (95% CI: 10.7 - 54.4) in favor of the above-threshold branch. It was hypothesized that this observation could be due to the incomplete eradication of microscopic spread (suggested to be present within a range of 8-26 mm outside the GTV [7]) or immuno-response stimulation ('abscopal effect'[36]).

In the present study, the shallower dose fall-off of CyberKnife treatment plans as compared to conventional SBRT plans results in a significantly smaller proportion (5% vs. 25%) of CyberKnife plans receiving less than the threshold dose in the $ROI_{cont}$(30 mm). Supporting the hypothesis of our previous study, the entire CyberKnife cohort was found to significantly outperform VMAT/CRT$_{<21\text{Gy}}$ cohort with respect to distant control. No meaningful statistical analysis of the lower-than-threshold branch of the CyberKnife cohort was possible due to its small sample size (n = 13, 6% of cohort). Notably, this dosimetric characteristic of CyberKnife plans was not specifically intended during treatment planning, but is rather inherent to its high non-coplanarity and lack of strict adherence to RTOG guidelines. This result implies that with current lung SBRT prescriptions not taking the threshold dose in consideration, CyberKnife treatments provide statistically superior DM-free outcomes compared to at least 25% of VMAT/CRT plans. As the threshold dose is indeed achieved in the remaining 75% of VMAT/CRT patients, no statistically significant improvement is observed when comparing the *entire* VMAT/CRT cohort to the CyberKnife cohort (HR = 1.47, 95% CI: 0.89 - 2.44, $p$-value $= 0.13$ as seen in Figure 5.3b). For conventional SBRT delivered at other institutions, a similar analysis needs to be performed to establish whether a non-negligible fraction of patients also receive less than the threshold dose. While there are numerous speculative biological explanations for these results [3], we stress that biologically oriented studies are required to understand further. This study presents a statistically confirmed observation, hopefully leading to further research.

The present study did *not* find an association in the CyberKnife cohort between incidence rate of DM and the numerical mean dose delivered to $ROI_{cont}$(x mm) for any value of $x$. This is in stark contrast with our previous study [3], where a strong correlation was found for all values of $x$. Our hypothesis is that this is due to the known sigmoidal behaviour of tumor control relative to delivered dose [28, 37]. We speculate that MDEs within the region outside the PTV would follow a similar model, considering they are also cancer cells. Since 95% of the CyberKnife cohort receive above the threshold dose (in other words, the plateau of the sigmoid model), we do not find any correlation between delivered dose and distant metastasis incidence.

A thorough multivariable analysis was performed to account for potential confounders. The only factor that was significantly associated with a difference in the risk of DM is whether or not the patient's mean dose to the $ROI_{cont}$*(30 mm)* reached the threshold. This result suggests that it is not the treatment modality that affects DM incidence but rather the dose to a 30 mm shell surrounding the PTV. This result further supports the finding that the proposed threshold dose is meaningful. Consistent with our previous study [3], no association was found between prescription/mean PTV dose and DM. We also note the larger number of T2 tumors in the CK cohort, compared to the VMAT/CRT cohort (20% to 5%, standardized difference = 0.49). As the multivariable analysis did not find staging to be associated with the outcome (DM), it is unlikely to be a confounder.

The previous study [3] did not find a correlation between the incidence rate of local failure and the mean dose delivered to $ROI_{cont}$(x mm) for any value of $x$. This study *did* find a significant difference between the two cohorts investigated (Figure 5.3d). After adjusting for other clinical factors, including prescription $EQD2_{10}$, VMAT/CRT patients were still found to have a significantly higher risk of local failure when compared to CyberKnife patients. There are a number of potential factors that could have caused this association. As shown and discussed throughout this report, the differences between VMAT/CRT and CyberKnife dose distributions seem to have an impact. We note that the mean dose to the $ROI_{cont}$(30 mm), used as a continuous variable rather than a threshold, was found to be associated with local control in a univariable analysis (HR = 0.94, 95% CI: 0.90 - 0.98, $p$-value = 0.002). However, when adjusted for other clinical factors including treatment modality (VMAT/CRT vs. CyberKnife), this metric was not found to be statistically significantly associated with local failure (HR = 0.96, 95% CI: 0.90 - 1.02, *p*-value = 0.17). The multivariable analysis was repeated while varying the threshold of the mean dose to the $ROI_{cont}$(30 mm). Although not statistically significant, hazard ratios in the range of 0.28 to 0.41 were found for threshold doses of 30 to 34 Gy (see Supplementary Material). Future work should be done to determine an intuitive, predictive metric that is appropriately representative of the dosimetric differences between conventional SBRT techniques and CyberKnife. Furthermore, the motion tracking employed by a CyberKnife unit may have an impact on local control rates in our representative cohort. Several steps were taken to reduce the impact of motion during VMAT/CRT treatments, including the use of the internal target volume based on 4DCT and daily cone beam CT. However these measures cannot compensate for motion to the extent provided by real-time tracking as performed by CyberKnife. We note that this study's 2-year local control rates (96% (95% CI: 92% - 98% and 88%, 95% CI: 82% - 92%), for CyberKnife and VMAT/CRT respectively) were comparable to those found in a previous systematic review by Soldà *et al.* (88% (95% CI: 78% - 94%) and 91% (95% CI: 89% - 93%)) [38].

The slower dose fall-off for CyberKnife compared to our standard-of-care conventional SBRT resulted in significantly lower incidence rates of DM for CyberKnife patients than VMAT/CRT$_{<21\text{Gy}}$ patients. However, as slower dose fall-offs are still achievable when using VMAT or CRT, the former distinction could potentially disappear if a threshold mean dose was introduced into the radiotherapy prescription. This is demonstrated by the non-significant hazard ratio of the treatment modality when adjusted for the threshold dose and other clinical factors. However, we note that dose distributions are spatially complex, heterogeneous and highly dependent on treatment modality. A secondary mean dose prescription to a region outside the PTV does not address this spatial heterogeneity across individual treatment plans. Further research should be done in order to determine the ideal prescription parameter that results in the most consistent dose distribution across treatment plans. Prior to clinical implementation, these results warrant external validation using additional independent cohorts, ideally through a prospective trial. Furthermore, this study is purely observational in nature. More biologically oriented investigations should be done to determine the validity of the hypothesized causality presented within this report. Nevertheless, as CyberKnife treatment plans routinely deliver above-threshold dose to regions outside the PTV, the proposed clinical recommendations should not cause further complications than the ones currently observed in patients treated with CyberKnife. This was explicitly shown in our previous study [3] which found no correlation between radiation pneumonitis and the dose to $ROI_{cont}$(30 mm). Thus we recommend the introduction of a secondary dose prescription of approximately 21 Gy to a region extending 30 mm outwards from the PTV particularly when using conventional SBRT modalities.

**Conflicts of Interest** None.

**Acknowledgements** André Diamant acknowledges majority support from grant FRQNT-205163 and partial support from the grants NSERC CREATE 432290, CIHR FDN-143257 and MOP-114910. Veng Jean Heng acknowledges support from the grants NSERC CREATE 432290 and NSERC RGPIN-2014-06475. None of the funding sources had any role in the creation or execution of the study. The authors acknowledge Piotr Pater and Tanner Connell for their contributions explaining and confirming the historical standard of care.

# References

[1] Canadian Cancer Society, “What is non-small cell lung cancer?,” 2015.

[2] N. C. Institute, “Metastatic Cancer Fact Sheet - National Cancer Institute,” 2016.

[3] A. Diamant, A. Chatterjee, S. Faria, I. Naqa, H. Bahig, E. Filion, C. Robinson, H. Al-Halabi, and J. Seuntjens, “Can dose outside the PTV influence the risk of distant metastases in stage I lung cancer patients treated with stereotactic body radiotherapy (SBRT)?,” *Radiotherapy and Oncology*, 2018.

[4] L. Kim, C. Wang, A. Khan, and M. Pierce, “CTV: The Third Front,” *International Journal of Radiation Oncology*Biology*Physics*, vol. 95, no. 2, pp. 800–801, 2016.

[5] M. Van Herk, “Errors and Margins in Radiotherapy,” *Seminars in Radiation Oncology*, vol. 14, no. 1, pp. 52–64, 2004.

[6] T. C. Mineo, V. Ambrogi, E. Pompeo, and A. Baldi, “Immunohistochemistry-detected microscopic tumor spread affects outcome in en-bloc resection for T3-chest wall lung cancer,” *European Journal of Cardio-thoracic Surgery*, vol. 31, no. 6, pp. 1120–1124, 2007.

[7] J. Van Loon, C. Siedschlag, J. Stroom, H. Blauwgeers, R. J. Van Suylen, J. Knegjens, M. Rossi, A. Van Baardwijk, L. Boersma, H. Klomp, W. Vogel, S. Burgers, and K. Gilhuijs, "Microscopic disease extension in three dimensions for non-small-cell lung cancer: Development of a prediction model using pathology-validated positron emission tomography and computed tomography features," *International Journal of Radiation Oncology Biology Physics*, vol. 82, no. 1, pp. 448–456, 2012.

[8] F. J. Salguero, J. S. A. Belderbos, M. M. G. Rossi, J. L. G. Blaauwgeers, J. Stroom, and J. J. Sonke, "Microscopic disease extensions as a risk factor for loco-regional recurrence of NSCLC after SBRT," *Radiotherapy and Oncology*, vol. 109, no. 1, pp. 26–31, 2013.

[9] X. Meng, X. Sun, D. Mu, L. Xing, L. Ma, B. Zhang, S. Zhao, G. Yang, F. M. Kong, and J. Yu, "Noninvasive evaluation of microscopic tumor extensions using standardized uptake value and metabolic tumor volume in non-small-cell lung cancer," *International Journal of Radiation Oncology Biology Physics*, vol. 82, no. 2, pp. 960–966, 2012.

[10] W. Kilby, J. R. Dooley, G. Kuduvalli, S. Sayeh, and C. R. Maurer, "The CyberKnife® Robotic Radiosurgery System in 2010," *Technology in Cancer Research & Treatment*, vol. 9, pp. 433–452, oct 2010.

[11] K. Tamari, O. Suzuki, N. Hashimoto, N. Kagawa, M. Fujiwara, I. Sumida, Y. Seo, F. Isohashi, Y. Yoshioka, T. Yoshimine, and K. Ogawa, "Treatment outcomes using CyberKnife for brain metastases from lung cancer," *Journal of Radiation Research*, vol. 56, pp. 151–158, jan 2015.

[12] W. T. Brown, X. Wu, F. Fayad, J. F. Fowler, B. E. Amendola, S. García, H. Han, A. de la Zerda, E. Bossart, Z. Huang, and J. G. Schwade, "CyberKnife Radiosurgery for Stage I Lung Cancer: Results at 36 Months," tech. rep., University of Miami, 2007.

[13] Z. Wang, Q.-T. Kong, J. Li, X.-H. Wu, B. Li, Z.-T. Shen, X.-X. Zhu, and Y. Song, "Clinical outcomes of cyberknife stereotactic radiosurgery for lung metastases.," *Journal of thoracic disease*, vol. 7, pp. 407–12, mar 2015.

[14] D. Coon, A. S. Gokhale, S. A. Burton, D. E. Heron, C. Ozhasoglu, and N. Christie, "Fractionated Stereotactic Body Radiation Therapy in the Treatment of Primary, Recurrent, and Metastatic Lung Tumors: The Role of Positron Emission Tomography/Computed

Tomography–Based Treatment Planning," *Clinical Lung Cancer*, vol. 9, pp. 217–221, jul 2008.

[15] B. T. Collins, K. Erickson, C. A. Reichner, S. P. Collins, G. J. Gagnon, S. Dieterich, D. A. McRae, Y. Zhang, S. Yousefi, E. Levy, T. Chang, C. Jamis-Dow, F. Banovac, and E. D. Anderson, "Radical stereotactic radiosurgery with real-time tumor motion tracking in the treatment of small peripheral lung tumors.," *Radiation oncology (London, England)*, vol. 2, p. 39, oct 2007.

[16] A. Bezjak, R. Paulus, L. E. Gaspar, R. D. Timmerman, W. L. Straube, W. F. Ryan, Y. I. Garces, A. T. Pu, A. K. Singh, G. M. Videtic, R. C. McGarry, P. Iyengar, J. R. Pantarotto, J. J. Urbanic, A. Y. Sun, M. E. Daly, I. S. Grills, P. Sperduto, D. P. Normolle, J. D. Bradley, and H. Choy, "Safety and Efficacy of a Five-Fraction Stereotactic Body Radiotherapy Schedule for Centrally Located Non-Small-Cell Lung Cancer: NRG Oncology/RTOG 0813 Trial.," *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, vol. 37, pp. 1316–1325, may 2019.

[17] G. M. Videtic, C. Hu, A. K. Singh, J. Y. Chang, W. Parker, K. R. Olivier, S. E. Schild, R. Komaki, J. J. Urbanic, H. Choy, and H. Choy, "A Randomized Phase 2 Study Comparing 2 Stereotactic Body Radiation Therapy Schedules for Medically Inoperable Patients With Stage I Peripheral Non-Small Cell Lung Cancer: NRG Oncology RTOG 0915 (NCCTG N0927)," *International Journal of Radiation Oncology*Biology*Physics*, vol. 93, pp. 757–764, nov 2015.

[18] E. E. Wilcox, G. M. Daskalov, H. Lincoln, R. C. Shumway, B. M. Kaplan, and J. M. Colasanto, "Comparison of Planned Dose Distributions Calculated by Monte Carlo and Ray-Trace Algorithms for the Treatment of Lung Tumors With CyberKnife: A Preliminary Study in 33 Patients," *International Journal of Radiation Oncology*Biology*Physics*, vol. 77, pp. 277–284, may 2010.

[19] V. W. Wu, K.-w. Tam, and S.-m. Tong, "Evaluation of the influence of tumor location and size on the difference of dose calculation between Ray Tracing algorithm and Fast Monte Carlo algorithm in stereotactic body radiotherapy of non-small cell lung cancer using CyberKnife," *Journal of Applied Clinical Medical Physics*, vol. 14, pp. 68–78, sep 2013.

[20] S. E. Braunstein, S. A. Dionisio, M. W. Lometti, D. S. Pinnaduwage, C. F. Chuang, S. S. Yom, A. R. Gottschalk, and M. Descovich, “Evaluation of ray tracing and Monte Carlo algorithms in dose calculation and clinical outcomes for robotic stereotactic body radiotherapy of lung cancers.,” *Journal of radiosurgery and SBRT*, vol. 3, no. 1, pp. 67–79, 2014.

[21] E. Gete, T. Teke, and W. Kwa, “Evaluation of the AAA treatment planning algorithm for SBRT lung treatment: Comparison with monte carlo and homogeneous pencil beam dose calculations,” *Journal of Medical Imaging and Radiation Sciences*, vol. 43, pp. 26–33, mar 2012.

[22] J. J. Ojala, M. K. Kapanen, S. J. Hyödynmaa, T. K. Wigren, and M. A. Pitkänen, “Performance of dose calculation algorithms from three generations in lung SBRT: Comparison with full Monte Carlo-based dose distributions,” *Journal of Applied Clinical Medical Physics*, vol. 15, no. 2, pp. 4–18, 2014.

[23] Y. Tsuruta, M. Nakata, M. Nakamura, Y. Matsuo, K. Higashimura, H. Monzen, T. Mizowaki, and M. Hiraoka, “Dosimetric comparison of Acuros XB, AAA, and XVMC in stereotactic body radiotherapy for lung cancer,” *Medical Physics*, vol. 41, no. 8, 2014.

[24] B. Huang, L. Wu, P. Lin, and C. Chen, “Dose calculation of Acuros XB and Anisotropic Analytical Algorithm in lung stereotactic body radiotherapy treatment with flattening filter free beams and the potential role of calculation grid size,” *Radiation Oncology*, vol. 10, feb 2015.

[25] I. Kawrakow, D. W. O. Rogers, F. Tessier, and B. R. B. Walters, “The EGSnrc Code System : Monte Carlo Simulation of Electron and Photon Transport,” *NRCC Report*, pp. 2001–2015, 2016.

[26] V. J. Heng, M.-A. Renaud, R. Doucet, A. Diamant, H. Bahig, and J. Seuntjens, “Large-scale dosimetric assessment of Monte Carlo recalculated doses for lung robotic stereotactic body radiation therapy,” tech. rep., McGill University, 2019.

[27] Y. Chang, Eric L., Nagata, *Stereotactic Body Radiation Therapy*. Tokyo: Springer, 2005.

[28] E. J. Hall and S. Willson, *Radiobiology for the Radiologist*. Philadelphia: Wolters Kluwer, 2012.

[29] J. F. Fowler, “21 years of Biologically Effective Dose,” *The British Journal of Radiology*, vol. 83, no. 991, pp. 554–568, 2010.

[30] S. M. Bentzen, W. Dörr, R. Gahbauer, R. W. Howell, M. C. Joiner, B. Jones, D. T. Jones, A. J. Van Der Kogel, A. Wambersie, and G. Whitmore, “Bioeffect modeling and equieffective dose concepts in radiation oncology-Terminology, quantities and units,” *Radiotherapy and Oncology*, vol. 105, no. 2, pp. 266–268, 2012.

[31] B. Lau, S. R. Cole, and S. J. Gange, “Competing risk regression models for epidemiologic data,” *American Journal of Epidemiology*, vol. 170, pp. 244–256, jul 2009.

[32] C. Davidson-Pilon, J. Kalderstam, P. Zivich, B. Kuhn, A. Fiore-Gartland, L. Moneda, Gabriel, D. WIlson, A. Parij, K. Stark, S. Anton, L. Besson, Jona, H. Gadgil, D. Golland, S. Hussey, R. Kumar, J. Noorbakhsh, A. Klintberg, E. Ochoa, D. Albrecht, Dhuynh, D. Medvinsky, D. Zgonjanin, D. S. Katz, D. Chen, C. Ahern, C. Fournier, Arturo, and A. F. Rendeiro, “CamDavidsonPilon/lifelines: v0.22.3 (late),” aug 2019.

[33] J. Gauthier, Q. V. Wu, and T. A. Gooley, “Cubic splines to model relationships between continuous variables and outcomes: a guide for clinicians,” *Bone Marrow Transplantation*, oct 2019.

[34] I. S. Grills, D. L. Fitch, N. S. Goldstein, D. Yan, G. W. Chmielewski, R. J. Welsh, and L. L. Kestin, “Clinicopathologic Analysis of Microscopic Extension in Lung Adenocarcinoma: Defining Clinical Target Volume for Radiotherapy,” *International Journal of Radiation Oncology Biology Physics*, vol. 69, pp. 334–341, oct 2007.

[35] A. Amini, N. Yeh, L. E. Gaspar, B. Kavanagh, and S. D. Karam, “Stereotactic Body Radiation Therapy (SBRT) for lung cancer patients previously treated with conventional radiotherapy: a review,” *Radiation Oncology*, vol. 9, no. 1, p. 210, 2014.

[36] I. Popp, A. Ligia Grosu, G. Niedermann, and D. Duda, “Immune modulation by hypofractionated stereotactic radiation therapy: Therapeutic implications,” *Radiotherapy and Oncology*, vol. 120, pp. 185–194, aug 2016.

[37] M. K. Martel, R. K. Ten Haken, M. B. Hazuka, M. L. Kessler, M. Strawderman, a. T. Turrisi, T. S. Lawrence, B. a. Fraass, and a. S. Lichter, “Estimation of tumor control

probability model parameters from 3-D dose distributions of non-small cell lung cancer patients.,” *Lung cancer (Amsterdam, Netherlands)*, vol. 24, no. 1, pp. 31–37, 1999.

[38] F. Soldà, M. Lodge, S. Ashley, A. Whitington, P. Goldstraw, and M. Brada, “Stereotactic radiotherapy (SABR) for the treatment of primary non-small cell lung cancer; Systematic review and comparison with a surgical cohort,” *Radiotherapy and Oncology*, vol. 109, pp. 1–7, oct 2013.

# Chapter 6

# Deep learning in head & neck cancer outcome prediction

**André Diamant**, Avishek Chatterjee, Martin Vallières, George Shenouda and Jan Seuntjens

Article published in: *Scientific Reports, vol. 9, no.2764, 2019.*

## 6.1 Preface

The previous two studies (Chapters 4 and 5) used a simple statistical model on dose metrics to perform outcome prediction on NSCLC patients. Midway through my PhD, we made the decision to pivot into using a more advanced outcome prediction model. At the time, deep learning was growing in popularity but had not been applied to prognosis prediction based on pre-treatment data. For a number of reasons (discussed throughout the studies), our hypothesized deep learning model would use raw image data (both CT and PET) to perform outcome prediction on head & neck cancer patients. The study presented within this chapter was the first results obtained, solely using the CT images.

## 6.2 Abstract

Traditional radiomics involves the extraction of quantitative texture features from medical images in an attempt to determine correlations with clinical endpoints. We hypothesize that convolutional neural networks (CNNs) could enhance the performance of traditional radiomics, by detecting image patterns that may not be covered by a traditional radiomic framework. We test this hypothesis by training a CNN to predict treatment outcomes of patients with head and neck squamous cell carcinoma, based solely on their pre-treatment computed tomography image. The training (194 patients) and validation sets (106 patients), which are mutually independent and include 4 institutions, come from The Cancer Imaging Archive. When compared to a traditional radiomic framework applied to the same patient cohort, our method results in a AUC of 0.88 in predicting distant metastasis. When combining our model with the previous model, the AUC improves to 0.92. Our framework yields models that are shown to explicitly recognize traditional radiomic features, be directly visualized and perform accurate outcome prediction.

## 6.3 Introduction

Radiation therapy is often used (74%[1]) to treat head and neck (H&N) cancers, a group of neoplasms originating from the squamous cells that line the mucosal surfaces of the oral cavity, paranasal sinuses, pharynx or larynx. Although loco-regional control of most H&N cancers is reasonably good (≈90%)[2], long-term survival can be quite poor (5-year survival rates as low as 50%)[2], in large part due to the development of distant metastasis or second primary cancers [3, 4]. Thus, the development of a model capable of identifying potential high-risk patients prior to treatment is critical. With such a model, a better-informed decision could be made regarding patient risk stratification. A high-risk patient could be assigned a more aggressive treatment regimen, potentially improving their outcome. Similarly, a low-risk patient could receive a more conservative treatment, delivering less radiation in order to reduce the chance of harmful side effects, such as hormonal disorders, tismus, xerostomia or dental disease [5]. The primary focus of this work is to build a model that is capable of discerning high-risk H&N cancer patients prior to their treatment using solely their computed tomography (CT) image.

Machine learning has played an increasingly prominent role over the past few decades in nearly every aspect of the STEM (science, technology, engineering and medicine) fields [6, 7].

Recently, deep learning, a sub-field of machine learning, has risen to the forefront of the artificial intelligence community [8]. One of the most popular deep learning tools is the convolutional neural network (CNN), a type of algorithm inspired by the biological neural networks within the animal visual cortex. CNNs consist of sequential *layers* which contain increasingly complex representations of data, eventually resulting in a classification of the input data [9, 10]. In particular, they are very effective at analyzing images and have achieved enormous success in numerous computer vision tasks, such as object detection, semantic segmentation, object classification and CADx (computer-aided diagnosis) [11–19].

Radiomics is the study of “image biomarkers" - the characterization of tumor phenotypes via the extraction of data from all types of medical images [20]. In the past few years, it has been extensively deployed for outcome prediction, among other applications [21–28]. For image-based outcome prediction, there are three common approaches. The first is the use of handcrafted features [29] which are directly extracted from the medical images. Often, these features are then fed into a machine learning algorithm for outcome prediction (e.g., random forest, support vector machine) [21–28]. We refer to this as “traditional radiomics". The second approach uses the outputs of the deeper layers in a CNN (often the final or penultimate fully connected layer) as “deep features". Similar to the first approach, these “deep features" are then fed into a secondary machine learning algorithm for outcome prediction [30–32, 18, 17]. The third approach employs transfer learning to fine-tune the weights of a pre-existing network to predict outcomes [9, 10]. Our methodology represents a novel fourth approach in that we use a single end-to-end CNN trained *de novo* (with no secondary machine learning algorithms) to predict oncological outcomes. To our knowledge, this is something that has not been successfully attempted in this context. This study will be specifically benchmarked against a previous study on the same data by Vallières *et al.* [25] which correlated a number of radiomic features from pre-treatment pre-segmented CT images with the outcome of H&N cancer patients. We use a novel deep CNN framework on the same cohort of patients, improving on a number of metrics, both quantitative and qualitative; detailed comparisons are made throughout this report. In the benchmark study, the most predictive combination of radiomic features related with distant metastasis involved $LRHGE_{GLRLM}$ (long run high grey level emphasis of the grey level run length matrix), $ZSV_{GLSZM}$ (zone size variance of the grey level size zone matrix) and $ZSN_{GLSZM}$ (zone size non-uniformity of the grey level size zone matrix) [25, 29]. We show that our network is capable of directly recognizing these radiomic features without having any prior information regarding their mathematical definition. While CNNs can be used for image segmentation [11, 12, 16], our methodology functions on pre-segmented tumor volumes, both to stay consistent with the benchmark study [25] and to simplify the task at hand.

One of our primary hypotheses is that a carefully trained CNN could learn the ability to recognize radiomic features. Typically, when attempting to apply a CNN framework to an unexplored image dataset that is of limited size, one uses a methodology called *transfer learning*. In transfer learning, a network that has already been trained and evaluated on another (much-larger) dataset is used as a starting point, and subsequently fine-tuned for the dataset of interest [9, 10]. This often results in excellent performance [33]. As an example, ImageNet [34] is a database of over 1 million RGB images which programmers compete on in an attempt to accurately classify the images into 1 of over 20,000 categories (e.g., dog, cat, plane, car, bench). Transfer learning was used to teach a top-performing network (Google's Inception-v3 [35]) to classify dogs into one of eleven breeds (e.g., bulldog, dachshund) with 96% accuracy [36] . The success is largely because of the *similarity* in the features that distinguish objects and dog breeds, features such as sharp edges or color gradients. In accordance with our goal of training a CNN that can recognize radiomic features, employing a transfer learning approach is possibly less successful. Furthermore, if an adequately sized dataset is available for the relevant classification task, it may not be necessary to use transfer learning as a methodology. Since medical images look substantially different from the everyday world to the human eye, and we have a dataset of 300 patients at our disposal, we decided to explore training a network *de novo* using gray-scale CT images. A comparison to a more traditional transfer learning approach is included in the Supplementary Information to quantitatively evaluate this hypothesis.

The novel contributions of this work are three-fold. Firstly, we develop a deep CNN-based framework capable of *accurately predicting H&N cancer treatment outcomes based solely on a patient's pre-treatment CT image*. Secondly, the framework is an externally validated medical gray-scale end-to-end CNN built *de novo*, rather than using transfer learning. Finally, the CNN is shown to *explicitly recognize previously engineered radiomic features with proven predictive power[25] on a benchmark study*, and is shown to complement their performance in a number of qualitative and quantitative ways which will be discussed throughout this report.

## 6.4 Results

### 6.4.1 Benchmark study comparison

The testing set of 106 patients used in the benchmark study [25] is used as an independent validation set in this study. The training set used was identical to that of the benchmark study. The results obtained from our CNN framework are shown alongside the results of the compared study [25] in Tables 6.1 and 6.2. The training and evaluation was done on the *central* tumor slice, which was defined as the slice with the maximum number of tumor pixels within a patient's entire set of CT images. This is in contrast to the benchmark study, where the model was trained and evaluated on the *entire* tumor volume. To evaluate the model's robustness with respect to the precise choice of tumor slice, the same network was also trained and evaluated on the slice directly above (superior) and below (inferior) the central slice. In summary, the most powerful network resulted in an area under the receiver operating characteristic curve (AUC) of 0.88 when predicting distant metastasis, comparable to the benchmark result. The most improved network, trained to predict loco-regional failure, had an AUC of 0.65, a substantial improvement over the prior study which was unable to find *any* predictive radiomic features. An AUC of 0.70 was found when predicting overall survival, comparable to the benchmark result. These improvements will have to be further verified by using an additional independent testing set on which the CNN is applied without any change to the hyper-parameters. The precise choice of the evaluation slice did not have a significant impact on the results, as shown in Table 6.1. It is noted that this study calculated specificity and sensitivity based on an optimized threshold, while the benchmark study[25] performed imbalance adjustments and used a threshold of 0.5. Additionally, using the same logistic regression methodology described in the benchmark study [25],we combined the final output score of our DM CNN model with the three aforementioned features used in the DM model of the benchmark study. The DeLong test was used to assess whether the combined model resulted in a statistically significant change in the AUC [37]. The new four-feature model had an AUC of 0.92 in the validation set ($p$-value of 0.04 when compared to the benchmark model, $p$-value of 0.12 when compared to the CNN model). This combination approach could not be implemented for the other outcomes, due to the traditional radiomics model's inability to find strong individual features.

Table 6.1 **Validation set AUC results compared to Vallières *et al.'s*[25] testing set results on the same patient cohort.** Robustness was evaluated by training and evaluating networks on the center slice, the inferior slice and the superior slice. Final column represents a combined model utilizing the CNN score and the traditional features from Vallières *et al.* [25]. The combined model was only implemented for DM, due to the traditional radiomics model's inability to find strong individual features for the other two outcomes.

| | **AUC (Area under the curve)** | | | | **Combined model AUC** |
|---|---|---|---|---|---|
| | *Central slice* | *Superior slice* | *Inferior slice* | *Vallières et al.*[25] | |
| **Distant metastasis (DM)** | 0.88 | 0.88 | 0.88 | 0.86 | 0.92 |
| **Loco-regional failure (LRF)** | 0.65 | 0.63 | 0.64 | 0.50 | - |
| **Overall survival (OS)** | 0.70 | 0.68 | 0.67 | 0.65 | - |

Table 6.2 **Validation set results compared to Vallières *et al.'s*[25] testing set results on the same patient cohort.** Balanced accuracy is defined as the average of the specificity and sensitivity. It is noted that this study calculated specificity and sensitivity based on thresholds optimized in the training set, while the benchmark study[25] performed imbalance adjustments during training and then used a single probability threshold of 0.5 in the testing phase. DM: Distant metastasis; LRF: Loco-regional failure; OS: Overall survival.

| | **Specificity** | | **Sensitivity** | | **Balanced Accuracy** | |
|---|---|---|---|---|---|---|
| | *Present study* | *Vallières et al.[25]* | *Present study* | *Vallières et al.[25]* | *Present study* | *Vallières et al.[25]* |
| **DM** | 0.89 | 0.77 | 0.86 | 0.79 | 88% | 77% |
| **LRF** | 0.67 | 0.61 | 0.65 | 0.39 | 66% | 58% |
| **OS** | 0.67 | 0.67 | 0.68 | 0.55 | 68% | 62% |

### 6.4.2 Cross validation

To better assess the stability of our results, we performed 5-fold cross validation on the entire set of 300 DM images. No changes to the hyper-parameters were made between any of the folds and thus remained identical to the hyper-parameters used in the comparison presented above. The mean AUC was found to be 0.85 (range: 0.80 to 0.88). It is noted that the 5-fold cross validation results should not be directly compared to the results of the benchmark study [25] due to the differing data partitioning scheme.

### 6.4.3 Visualization of results

There are a number of visualization tools that we can use to better understand the behaviour of the CNN, many of which are facilitated by the Keras-*vis* toolbox [38]. All visualization examples in this section represent the highest performing network (i.e., trained on the central tumor slice to predict distant metastasis). Figure 6.1 represents a montage of four patient CTs, two of whom developed DM (top), and two who did not (bottom). These particular CT images are chosen to represent the diversity in features perceivable by the human eye (e.g., shape, first-order textures). The leftmost column is a zoomed-in view of the 512 × 512 pixel CT image that enters the model, representing the pre-processing done (which merely amounts to setting any pixels outside of the gross tumor volume to zero). The middle column shows gradient class activation maps (Grad-CAMs [39]) on the penultimate convolutional block, which depict what areas of the image the CNN found most relevant for outcome prediction. The heat map represents how important each region of the image is to the given classification. This information can potentially be used by clinicians to make further hypotheses regarding the nature of the tumor. The final column depicts a merger of the Grad-CAM and the CT image. This is the image we recommend is used when attempting to understand the network's behavior on a particular input image.

Another method of visualizing our network is through an activation map. Shown in Figure 6.2, an activation map represents a procedurally generated image that would result in a distant metatasis classification of maximal probability (a score of 1). We stress that this image was generated on the fully trained network and thus does not represent an individual CT image within the dataset. The image appears quite disorderly at first, but there are some interesting aspects that we can discern from it. Firstly, the image appears mostly homogeneous on a large scale, meaning no region of the image favors one pattern over another. This is an indication that our network is approaching location invariance, due to tumor locations being highly variable regardless of outcome within the training set. Secondly, when focusing on a small portion of the image, the image appears heterogeneous both in shape and intensity. This is indicative of heterogeneous tumors being more aggressive and likely to be assigned a poor outcome, an observation consistent with the published literature [40–43]. The corresponding minimal activation map (i.e., a score of 0) is shown and discussed in the Supplementary Note/Fig. A1. Although Figure 6.2 does not directly explain any specific patient prediction, it gives insight into the trained network and the patterns it is associating with a specific classification.

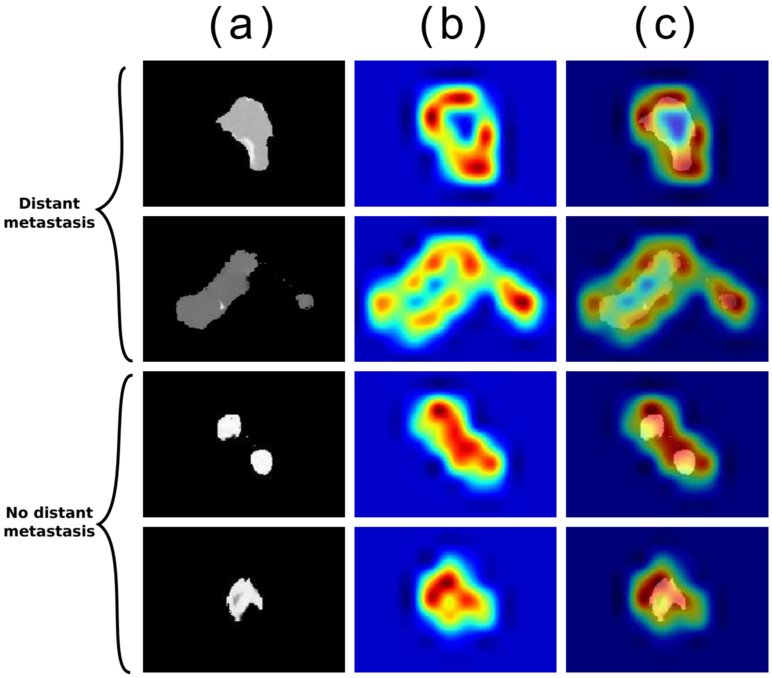


Figure 6.1 **Montage of tumors and gradient class activation maps (Grad-CAM):** First two rows represent patients who developed distant metastasis (DM). Last two rows represent patients who did not develop DM. **(a):** Raw image input into the model (zoomed in for visualization purposes). Note that tumor segmentation is performed prior to being input into the model. **(b):** Gradient class activation map (Grad-CAM [39]) of the penultimate convolutional block, red represents a region more significant to the designated classification. **(c):** Image merge of the first two columns.

### 6.4.4 Filters within the CNN explicitly recognize radiomic features

To determine whether the CNN trained *de novo* could be recognizing radiomic features, we visualized a montage of *filters*. Each convolutional block within the network functions by convolving a variable number of learned *filters* with the input data [9]. Figure 6.3 represents 4 of the 128 filters that make up the final convolutional block. Similar to Figure 6.2, Figure 6.3 is not representative of any specific CT image but rather the final trained network. Each square represents the procedurally generated input image that would maximize the mean output of a specific filter, thus informing us at an abstract level what sort of image each filter is interested in

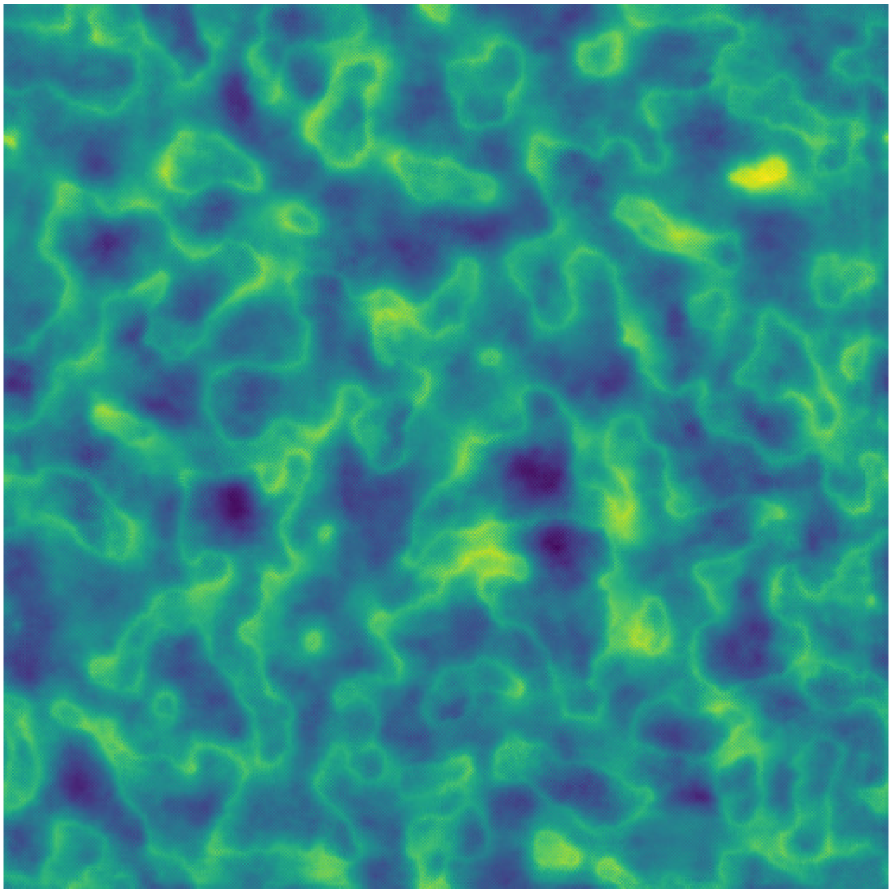


Figure 6.2 **Maximal activation map depicting a procedurally generated image that results in a classification of maximal probability.** Represents a procedurally generated image input that would result in a maximal classification score of 1 (i.e. distant metastasis). Of particular interest is the large scale homogeneity and the small scale heterogeneity. Color map chosen solely for visualization purposes.

when making a decision. It is evident that each of these filters is maximally activated by various *textures*, rather than a particular shape or object as is common in more typical convolutional neural networks. In order to determine whether any of the filters were specifically activated by previously engineered radiomic features, we extracted 94 radiomic features (as described in the Image Biomarker Standardization Initiative (IBSI [29])) from each filter's maximal activation map (Figure 6.3). An example of this analysis when performed on 64 of the filters (chosen among those whose maximal activation maps converged) is shown in Figure 6.4. The $y$-axis represents the normalized value of a specific radiomic feature that Vallières *et al.*[25] found to be predictive (blue: $ZSV_{GLSZM}$, red: $ZSN_{GLSZM}$, yellow: $LRHGE_{GLRLM}$). These features were calculated using the exact same extraction parameters as the benchmark study. The letters indicate the corresponding square in Figure 6.3. Of particular interest are the numerous blue peaks (**a**, **c** and **d**). These filters are strongly activated by an input region with a high value of the radiomic feature $ZSV_{GLSZM}$, precisely the feature that Vallières *et al.*[25] found to be most predictive. Many of these filters are also strongly activated by extreme (high or low) values of $ZSN_{GLSZM}$ and $LRHGE_{GLRLM}$ (red and yellow respectively). In particular, many filters represent various permutations of the three features. As an example, (**a**) is activated by all three radiomic features, (**b**) is mostly activated by red, (**c**) is mostly activated by blue and yellow, while (**d**) is mostly activated by blue. In essence, these filters are recognizing and combining various radiomic features to help classify a particular input image.

The 64 filters' maximal activation maps are displayed in Supplementary Fig. A2. The radiomic analysis for these 64 filters and all 94 radiomic features (as described in the Image Biomarker Standardization Initiative (IBSI [29]) and extracted using 128 gray levels and a scale of 1 mm) is displayed in Supplementary Fig. A3.

## 6.5 Discussion

These results show great potential in using convolutional neural networks trained *de novo* on medical gray-scale images to predict oncological treatment outcomes. The average adult human is capable of looking at an object or person and immediately classifying it properly (type of object/name) with virtually 100% accuracy. This is largely due to the types of features that our brains have developed to subconsciously look for and associate with a particular object or person (e.g. sharp edges/color gradients combining into shapes). These concepts are precisely what networks trained on ImageNet have learned to process. A major benefit to training a network *de novo* is that it can learn abstract concepts unique to the dataset of interest such as specific radiomic texture features. This is explicitly shown in Figures 6.3 and 6.4. The radiomic analysis performed (Figure 6.4) shows that many of these filters are able to process and distinguish radiomic features *without explicitly being told the definition of any feature*. The primary example of this are the numerous blue peaks in Figure 6.4. Effectively, these filters have learned the ability to see an image from a perspective that is interested in the value of the radiomic feature $ZSV_{GLS`ZM}$. Furthermore, this feature is one of the radiomic features that Vallières *et al.* [25] found to be predictive. Similarly, many of the filters are strongly attuned to

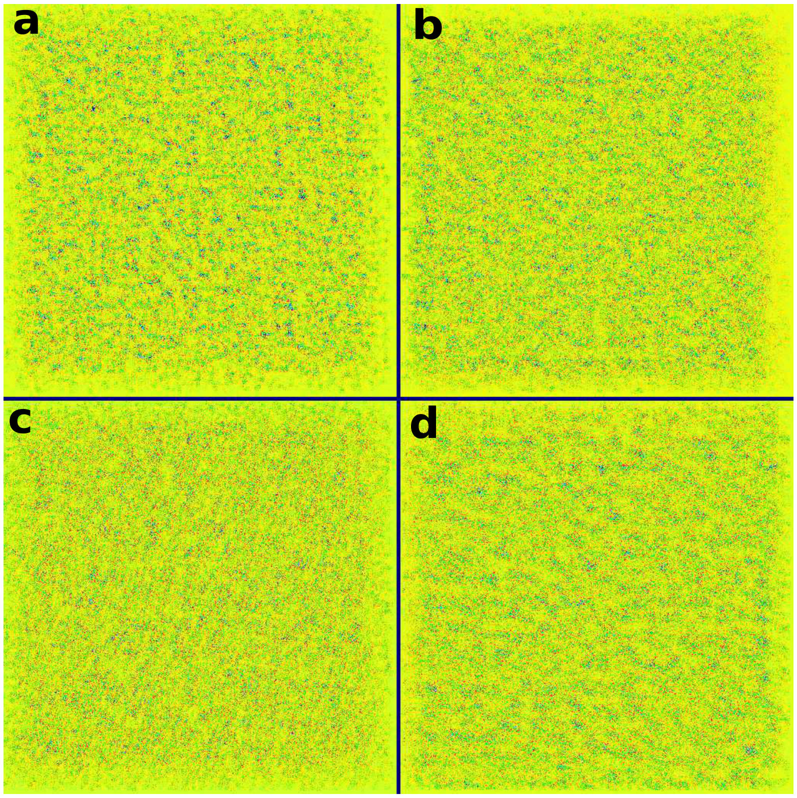


Figure 6.3 **Maximal activation maps of four filters within the final convolutional layer.** Represents procedurally generated images that would each result in a particular filter being maximally activated. While humans are capable of distinguishing between these four images, we are currently unable to directly interpret them. Our framework is capable of analyzing the type of data that these images represent. The lettering scheme is relevant to Figure 6.4. Color map chosen solely for visualization purposes. The maximal activation map was generated as a 512 × 512 image to spatially represent the input CT shape.

the features $LRHGE_{GLRLM}$ and $ZSN_{GLSZM}$, the two other features that Vallières *et al.* [25] found predictive. The ability of our network to directly recognize radiomic features without being told their definition is powerful for a number of reasons, one of which being it may remove the need to specifically engineer new features. The relationship between the filters of our network and the most predictive radiomic features of the compared study also helps build confidence in the network's output. As seen in Table 6.1, a combination approach does increase the AUC from 0.88 to 0.92. This indicates that although our network does recognize the features to some extent, it does not directly represent them. In other words, there is still some orthogonality between the quantitative value of a radiomic feature and the filter's impact on our network's

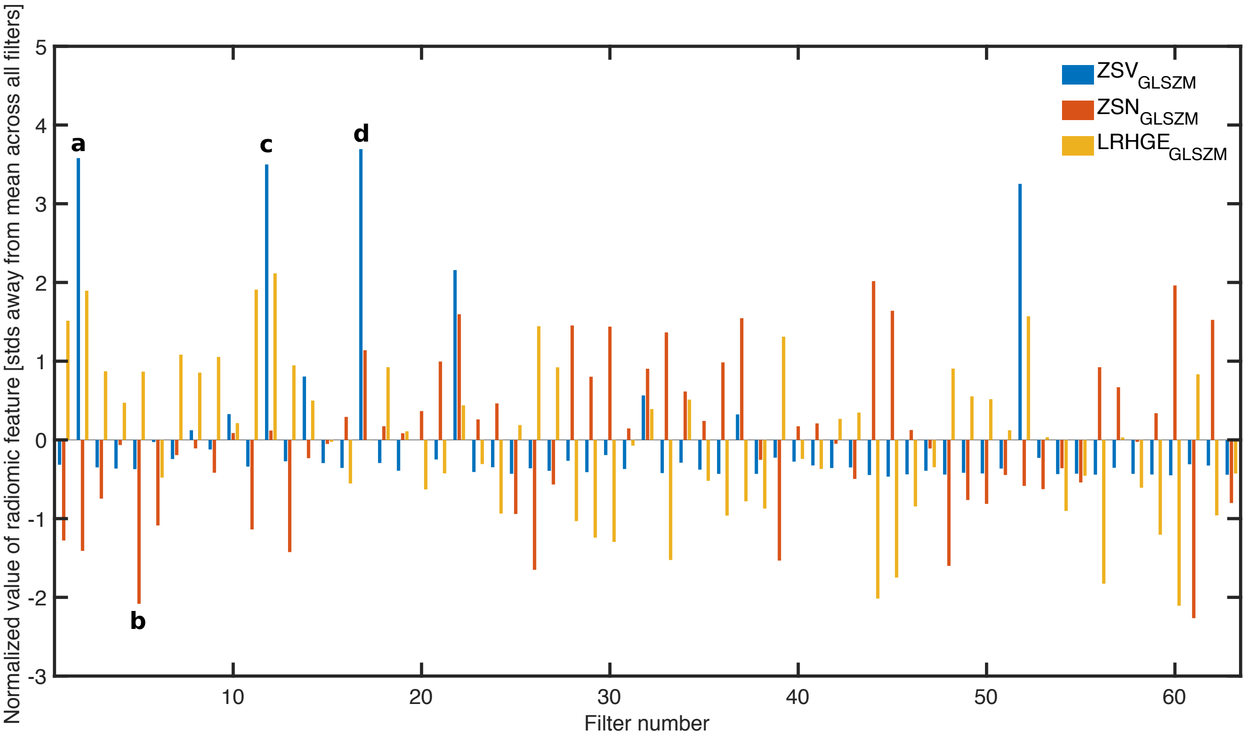


Figure 6.4 **Normalized value of three radiomic features of interest across 64 convolutional filters within the final convolutional layer.** $x$-axis represents which filter within the third convolutional block. $y$-axis represents the value of the radiomic feature, normalized across all filters. Blue bars represent $ZSV_{GLSZM}$, red bars represent $ZSN_{GLSZM}$ and yellow bars represent $LRHGE_{GLRLM}$. Of particular interest are the numerous blue peaks (**a**, **c** and **d**). These filters are strongly activated by an input region with a high value of the radiomic feature $ZSV_{GLSZM}$, precisely the feature that Vallières *et al.*[25] found to be most predictive. Many of these filters are also strongly activated by extreme (high or low) values of $ZSN_{GLSZM}$ and $LRHGE_{GLRLM}$ (red and yellow respectively). In particular, many filters represent various permutations of the three features. As an example, (**a**) is activated by all three radiomic features, (**b**) is mostly activated by red, (**c**) is mostly activated by blue and yellow, while (**d**) is mostly activated by blue. The lettering scheme corresponds to the maximal activation maps shown in Figure 6.3.

output score. Future work could be done to further investigate the difference between the two representations.

The visualization tools we used in this work begin to overcome one of the primary obstacles in outcome analysis using a machine learning model: interpretability. The benchmark study found the most predictive combination of radiomic features related with distant metastasis to be $LRHGE_{GLRLM}$, $ZSV_{GLSZM}$ and $ZSN_{GLSZM}$. While these features are mathematically well-defined, it is difficult for humans to visualize them, let alone develop an intuition for them. In comparison, the visualization tools our framework uses, particularly the class activation maps (Figure 6.1), are more interpretable. It is noted that the interpretations discussed thus far do not exhaust all the information our method can extract. Through future research and collaboration, it is our belief that our framework can lead to further hypotheses. The visualization tools grant the ability to not only have more confidence in the output, but also provide a tool to the medical community that may help discover unknown aspects of these gray-scale images.

One major advantage of this framework is the lack of feature engineering, in stark contrast to traditional radiomic frameworks. In particular, the benchmark study [25] required the extraction of 55 pre-defined radiomic features, 40 of which were extracted using combinations of three parameters (isotropic voxel size, quantization algorithm and number of gray levels). In total, this resulted in the extraction of 1615 radiomic features. Prior to the extraction, an elaborate and complex feature selection process was required to identify the “potentially useful" features [25]. Elaborate procedures for the selection of potentially useful features have also been developed for radiomic analyses of other cancer types [44]. The approach developed in this report eschews this problem by giving the algorithm the full set of un-altered pixel data of the tumor and allowing the algorithm to tell the user what is important rather than the user explicitly telling the algorithm what is important. This is one of the primary motivations behind our exploration of a end-to-end CNN, without any feature or machine learning algorithm selection. Similarly, as the handcrafted features were extracted from pre-segmented tumors, maintaining this segmentation allows us to better evaluate the hypothesized connection between the CNN's behavior and the radiomic features.

There are improvements that could be made to this framework to potentially increase prediction performance and generalizablity. As an example of a potential pre-processing step, the CT images could be cropped to include only the field-of-view surrounding the tumor itself. This could improve the learning capabilities of the algorithm particularly by improving the location invariance of the model. However, this would be (albeit slightly) complicating the framework, directly opposing one of its primary advantages. There are two sources of information that could be added to our framework. The first would be to fully incorporate 3-dimensional information. In contrast to the radiomic study [25] which used the entire tumor volume, our framework only considers the central slice (with robustness estimated by training on one slice superior/inferior). It is emphasized that this study was capable of surpassing the predictive power of the benchmark study solely using a single 2-dimensional image. By incorporating the entire tumor, the performance could potentially be further improved. However, convolutional neural networks which incorporate 3-dimensional image information are architecturally complex and computationally expensive. Another information source is each patient's positron emission tomography (PET) image. Vallières *et al.* [25] found additional predictive power in each patient's PET image, so incorporating this information into our framework should improve performance. Potential image manipulation needs aside, this would be computationally simpler than incorporating volumetric information. The CT + PET image could be introduced into the network in a 2-channel fashion: input data would be 512 × 512 × 2 pixels, rather than 512 × 512 × 1 pixels. This is akin to the 3-channel RGB input that many traditional CNNs use. Additionally, our framework uses solely the pre-segmented gross tumor volume as an input rather than including the surrounding tissue. Ideally one would include the surrounding tissue and build a network capable of incorporating any information within. This would also remove the need for location invariance, as there could be additional information contained in the tumor's precise location within the anatomy. Finally, while we performed a relatively simple combination approach (logistic regression), future work could study the different methods that could be used to better combine traditional radiomics and CNN information. An in-depth study regarding transfer learning and whether a transfer-learnt network is capable of recognizing the same radiomic features could also be performed. These improvements were not included to align with the goal of assessing our primary hypothesis, keeping the initial framework as simple as possible and to reduce training time.

This report showed the power and potential of using a deep convolutional neural network built *de novo* to perform outcome predictions on the pre-treatment CT image of head and neck cancer patients. Often transfer learning is used to train a CNN on a new dataset due to the perception that thousands, if not millions of images are required to build an accurate model. Our framework shows that a training set of 200 medical gray-scale images may be sufficient to train a network *de novo*, with proper data augmentation. The model was shown to have the ability to explicitly recognize radiomic features and further improve on the performance of a traditional radiomics framework. Performance gains aside, our framework overcomes many of the typical issues when building a traditional radiomics-based model. Specifically, our model is capable of being interpreted in a more intuitive fashion and completely eschews the need for feature engineering. We believe our framework could serve as the base of a gray-scale image analysis tool capable of being adapted to other imaging modalities (e.g., PET, MRI) or other cancer sites (e.g., liver, lung, breast).

## 6.6 Methods

### 6.6.1 Patient cohorts

Extensive details regarding the patient cohort used throughout this report are publicly available on The Cancer Imaging Archive (TCIA) [25, 45] repository. Eligible patients were taken from four separate institutions (Hôpital général juif (HGJ), Centre hospitalier universitaire de Sherbooke (CHUS), Hôpital Maisonneuve-Rosemont (HMR) and Centre hospitalier de l'Université de Montréal (CHUM)). The majority received chemotherapy adjuvant to radiotherapy (92%) while the remainder solely received radiation (8%). All patients underwent joint FDG-PET/CT scans, however this study only made use of the CT image. The training set was defined as the patients from HGJ and CHUS, while the validation set was defined as the patients from HMR and CHUM. This was the same distribution used in the compared study [25], specifically notable due to the validation set only containing patients from independent institutions. Any patients with metastases or recurrent H&N cancer at presentation were excluded, along with any patients receiving palliative care. The median age of patients across the total cohort was 63 years (range: 18-88). The median follow-up period across all patients was 43 months (range: 6-112). Any patients that did not develop cancer recurrence and had a follow-up period of less than 24 months were discarded. The outcome distribution for both cohorts is shown in Table 6.3. It is noted that 2 patients from the training cohort were lost to data corruption.

### 6.6.2 Convolutional neural network architecture

Our CNN contains four main operations: **convolution, non-linearity, pooling and classification**. These four operations are facilitated by *layers*, which are the building blocks of the overall framework. The **convolution** operation is ultimately what learns and subsequently extracts features from the input data. The layer includes a variable number of *convolutional filters*, each of which acts as a sliding window (of a small size, e.g., 5 $\times$ 5 pixels) applying

Table 6.3 **Patient/outcome distribution [25, 45]**

| | Training cohort | Validation cohort |
|---|---|---|
| **Total** | 194 | 106 |
| **Outcome** | | |
| Distant metastasis (DM) | 26 (13%) | 14 (13%) |
| Loco-regional failure (LRF) | 29 (15%) | 16 (15%) |
| Death | 32 (16%) | 24 (23%) |
| **Institution** | | |
| Hôpital général juif (HGJ) | 92 (47%) | - |
| Centre hospitalier universitaire de Sherbooke (CHUS) | 102 (53%) | - |
| Hôpital Maisonneuve-Rosemont (HMR) | - | 41 (39%) |
| Centre hospitalier de l'Université de Montréal (CHUM) | - | 65 (61%) |
| **Tumor type** | | |
| Oropharynx | 129 (66%) | 77 (73%) |
| Hypopharynx | 5 (3%) | 7 (7%) |
| Nasopharynx | 20 (10%) | 8 (7%) |
| Larynx | 36 (19%) | 9 (8%) |
| Unknown | 4 (2%) | 5 (5%) |

a convolution over the input data. By learning a number of different filters (e.g., 64), the network is able to incorporate a large variety of features. The more filters we choose to learn, the more image features the network is able to ultimately extract and recognize in unseen images. The **non-linearity** operation is needed to accurately model the type of real-world data we are interested in. Many CNNs have adopted the use of a rectified linear unit (ReLU), which simply replaces all negative input values (from the preceding convolutional layer) by 0. Instead, we use a parametrized rectified linear unit (PReLU), which has largely the same effect but allows a small amount of the negative input values to propagate through the network by multiplying the negative portion of the input domain by a learnt non-zero slope [46]. Next, the **pooling** operation serves to progressively reduce the spatial size of the input information. This is important for computational efficiency, to ensure that the model can be generalized and most importantly, to introduce location invariance. In our model we utilize "max-pooling", an operation that replaces every $4 \times 4$ region of input data with the maximum value among them. Finally, the **classification** operation takes all the high-level features from the previous representations and combines them using a sigmoid activation function to determine which class the input represents.

Henceforth, we will refer to *convolutional blocks*, which contain the convolution operation, the non-linearity operation and the pooling operation stacked one after another. By stacking convolutional blocks, the network is able to progressively learn complex image features that humans are not used to processing.

Our CNN architecture is shown in Figure 6.5 and was largely chosen for its simplicity, a characteristic that increases the model's ability to generalize and reduces the risk of over-fitting. Of particular note is the usage of a 512 × 512 pixel input layer, allowing any standard CT image to directly be analyzed by the network with minimal pre-processing. The input layer is followed by 3 consecutive convolutional blocks. Each block consists of a convolutional layer, a max-pooling layer and a PReLu layer. It is noted that the PReLu layers are not explicitly depicted in Figure 6.5. The convolutional layers used a filter size of 5 × 5 pixels, 3 × 3 pixels and 3 × 3 pixels, respectively. These layers formed the foundation of the network, each subsequent layer uncovering more complex features. The first block containing a larger filter allowed the later layers to have larger effective fields of view, combining a larger number of input pixels to determine important features. Notably, the max-pooling layers grant the network some degree of location invariance, a crucial attribute given the fact that the location of the tumor within the CT should not impact the outcome. The 3 convolutional blocks were followed by two consecutive fully connected layers and a final PReLu. Finally, a drop-out layer was included after the last fully connected layer, directly prior to the classification layer. Drop-out played a major role in reducing over-fitting by removing half of the information every single stochastic gradient descent iteration. Each iteration, the output of half of the nodes in the final fully connected layer were set to 0. This teaches the network that it must be capable of functioning even with a substantial amount of missing information, effectively forcing it to not rely too heavily on a single piece of information [9].

### 6.6.3 Implementation details

Our framework was built on Python3 using the Keras library operating on the well-optimized tensor manipulation library Tensorflow[47, 48]. The final outcome probability (in the classification layer) was computed using a sigmoid classifier. Each convolutional block used a PReLU as an activation function. The network weights were optimized using a stochastic gradient descent algorithm with a fixed learning rate of 0.001 and a momentum of 0.5. The mini-batch size was 32 and the objective function used was binary cross-entropy. Image augmentation was

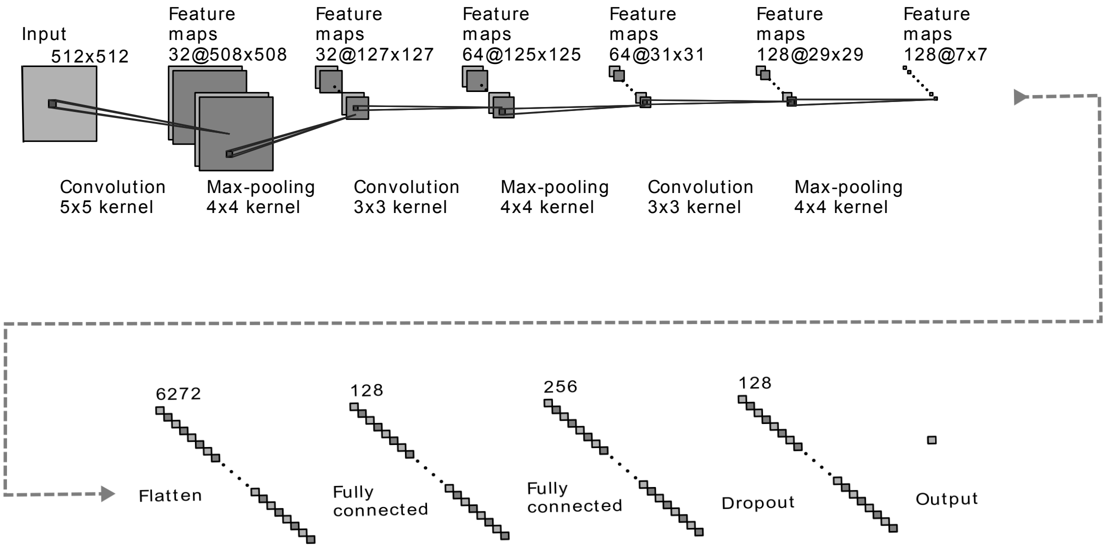


Figure 6.5 **Depiction of our convolutional neural network's architecture.** Text below the graphic represents the operation between layers. Text above the graphic represents the number of feature maps or nodes within the layer. The CNN consists of three consecutive *convolutional blocks*, each of which contain a convolutional layer (of varying filter size), a max-pooling layer (4 × 4 kernel) and a parametric rectified linear unit (not shown). Following this, the output is flattened and proceeds through two fully connected layers, a parametric rectified linear unit (not shown) and a dropout layer prior to being classified via a sigmoid activation function.

performed to increase generalization and reduce the training bias that the network is inherently subjected to [9]. Prior to training, each image was randomly flipped (horizontally and/or vertically), rotated a random amount (0-20°), and shifted a random fraction (0 to 0.4 times the total width of the image) in a random direction. This resulted in the total training dataset (and thus a single epoch) consisting of 4000 images (each tumor is augmented roughly 20 times). Our algorithm was trained and evaluated on a pair of NVIDIA GTX 1080TI graphic processing units to exploit their computational speed. Total training time for one network required approximately 5 hours (100 epochs). The time required to predict outcomes on the validation cohort is approximately 100 milliseconds. More details regarding the implementation and the specific range of parameters tested can be found in the Supplementary Methods.

**Data availability** The dataset analyzed throughout this study is publicly available on The Cancer Imaging Archive (TCIA) repository (doi: 10.7937/K9/TCIA.2017.8oje5q00) [25, 45, 49].

**Code availability** Code to build, compile, train and evaluate each model along with all visualization scripts are publicly available on Github.

**Acknowledgements** We acknowledge majority financial support from grant FRQNT-205163 and partial support from the grants NSERC CREATE 432290, CIHR FDN-143257 and MOP-114910. None of the funding sources had any role in the creation or execution of the study.

**Competing Interests** The authors declare no competing interests related to this work.

# References

[1] R. Atun, D. A. Jaffray, M. B. Barton, F. Bray, M. Baumann, B. Vikram, T. P. Hanna, F. M. Knaul, Y. Lievens, T. Y. M. Lui, M. Milosevic, B. O'Sullivan, D. L. Rodin, E. Rosenblatt, J. Van Dyk, M. L. Yap, E. Zubizarreta, and M. Gospodarowicz, "Expanding global access to radiotherapy.," *The Lancet. Oncology*, vol. 16, pp. 1153–86, sep 2015.

[2] S.-A. Yeh, "Radiotherapy for head and neck cancer.," *Seminars in plastic surgery*, vol. 24, pp. 127–36, may 2010.

[3] S. S. Baxi, L. C. Pinheiro, S. M. Patil, D. G. Pfister, K. C. Oeffinger, and E. B. Elkin, "Causes of death in long-term survivors of head and neck cancer," *Cancer*, vol. 120, pp. 1507–1513, may 2014.

[4] A. Ferlito, A. R. Shaha, C. E. Silver, A. Rinaldo, and V. Mondin, "Incidence and Sites of Distant Metastases from Head and Neck Cancer," *ORL*, vol. 63, no. 4, pp. 202–207, 2001.

[5] E. d. S. Tolentino, B. S. Centurion, L. H. C. Ferreira, A. P. de Souza, J. H. Damante, and I. R. F. Rubira-Bullen, "Oral adverse effects of head and neck radiotherapy: literature review and suggestion of a clinical oral care guideline for irradiated patients.," *Journal of applied oral science : revista FOB*, vol. 19, pp. 448–54, oct 2011.

[6] R. C. Deo, “Machine Learning in Medicine,” *Circulation*, vol. 132, pp. 1920–1930, nov 2015.

[7] M. I. Jordan and T. M. Mitchell, “Machine learning: Trends, perspectives, and prospects.,” *Science (New York, N.Y.)*, vol. 349, pp. 255–60, jul 2015.

[8] J. Schmidhuber, “Deep learning in neural networks: An overview,” *Neural Networks*, vol. 61, pp. 85–117, jan 2015.

[9] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[10] N. Buduma, *Fundamentals of Deep Learning*. O’Reilly, 2015.

[11] Dinggang Shen, G. Wu, H.-I. Suk, D. Shen, G. Wu, and H.-I. Suk, “Deep Learning in Medical Image Analysis,” *Annual Review of Biomedical Engineering*, vol. 19, no. 1, pp. 221–248, 2017.

[12] T. Brosch, L. Y. W. Tang, Y. Youngjin Yoo, D. K. B. Li, A. Traboulsee, and R. Tam, “Deep 3D Convolutional Encoder Networks With Shortcuts for Multiscale Feature Integration Applied to Multiple Sclerosis Lesion Segmentation.,” *IEEE transactions on medical imaging*, vol. 35, pp. 1229–1239, may 2016.

[13] Q. Qi Dou, H. Hao Chen, L. Lequan Yu, L. Lei Zhao, J. Jing Qin, D. Defeng Wang, V. C. Mok, L. Lin Shi, and P.-A. Pheng-Ann Heng, “Automatic Detection of Cerebral Microbleeds From MR Images via 3D Convolutional Neural Networks.,” *IEEE transactions on medical imaging*, vol. 35, pp. 1182–1195, may 2016.

[14] G. van Tulder and M. de Bruijne, “Combining Generative and Discriminative Representation Learning for Lung CT Analysis With Convolutional Restricted Boltzmann Machines.,” *IEEE transactions on medical imaging*, vol. 35, pp. 1262–1272, may 2016.

[15] J. Kleesiek, G. Urban, A. Hubert, D. Schwarz, K. Maier-Hein, M. Bendszus, and A. Biller, “Deep MRI brain extraction: A 3D convolutional neural network for skull stripping,” *NeuroImage*, vol. 129, pp. 460–469, apr 2016.

[16] W. Zhang, R. Li, H. Deng, L. Wang, W. Lin, S. Ji, and D. Shen, “Deep convolutional neural networks for multi-modality isointense infant brain image segmentation,” *NeuroImage*, vol. 108, pp. 214–224, mar 2015.

[17] B. Q. Huynh, H. Li, and M. L. Giger, “Digital mammographic tumor classification using transfer learning from deep convolutional neural networks,” *Journal of Medical Imaging*, vol. 3, p. 034501, aug 2016.

[18] N. Antropova, B. Q. Huynh, and M. L. Giger, “A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets,” *Med. Phys.*, vol. 44, no. 10, 2017.

[19] S. Trebeschi, J. J. M. Van Griethuysen, D. M. J. Lambregts, M. J. Lahaye, C. Parmar, F. C. H. Bakers, N. H. G. M. Peters, R. G. H. Beets-Tan, and H. J. W. L. Aerts, “Deep Learning for Fully-Automated Localization and Segmentation of Rectal Cancer on Multiparametric MR OPEN Background Work,” *Scientific reports*, vol. 7, no. 1, p. 5301, 2017.

[20] R. J. Gillies, P. E. Kinahan, and H. Hricak, “Radiomics: Images Are More than Pictures, They Are Data,” *Radiology*, vol. 278, pp. 563–577, feb 2016.

[21] H. J. W. L. Aerts, E. R. Velazquez, R. T. H. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebers, M. M. Rietbergen, C. R. Leemans, A. Dekker, J. Quackenbush, R. J. Gillies, and P. Lambin, “Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach,” *Nature Communications*, vol. 5, p. 4006, dec 2014.

[22] M. Vallières, C. R. Freeman, S. R. Skamene, and I. El Naqa, “A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities.,” *Physics in medicine and biology*, vol. 60, no. 14, pp. 5471–96, 2015.

[23] J. E. van Timmeren, R. T. Leijenaar, W. van Elmpt, B. Reymen, C. Oberije, R. Monshouwer, J. Bussink, C. Brink, O. Hansen, and P. Lambin, “Survival prediction of non-small cell lung cancer patients using radiomics analyses of cone-beam CT images,” *Radiotherapy and Oncology*, vol. 123, no. 3, pp. 363–369, 2017.

[24] T. P. Coroller, P. Grossmann, Y. Hou, E. Rios, R. T. H. Leijenaar, G. Hermann, P. Lambin, B. Haibe-kains, R. H. Mak, and H. J. W. L. Aerts, “CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma,” *Radiotherapy and Oncology*, vol. 114, no. 3, pp. 345–350, 2015.

[25] M. Vallières, E. Kay-Rivest, L. J. Perrin, X. Liem, C. Furstoss, H. J. Aerts, N. Khaouam, P. F. Nguyen-Tan, C. S. Wang, K. Sultanem, J. Seuntjens, and I. El Naqa, "Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer," *Scientific Reports*, vol. 7, no. 1, pp. 1–33, 2017.

[26] C. Parmar, P. Grossmann, D. Rietveld, M. M. Rietbergen, P. Lambin, and H. J. W. L. Aerts, "Radiomic Machine-Learning Classifiers for Prognostic Biomarkers of Head and Neck Cancer," *Frontiers in Oncology*, vol. 5, p. 272, dec 2015.

[27] C. Parmar, P. Grossmann, J. Bussink, P. Lambin, and H. J. W. L. Aerts, "Machine Learning methods for Quantitative Radiomic Biomarkers," *Scientific Reports*, vol. 5, p. 13087, oct 2015.

[28] C. Parmar, R. T. H. Leijenaar, P. Grossmann, E. Rios Velazquez, J. Bussink, D. Rietveld, M. M. Rietbergen, B. Haibe-Kains, P. Lambin, and H. J. W. L. Aerts, "Radiomic feature clusters and Prognostic Signatures specific for Lung and Head & Neck cancer," *Nature Publishing Group*, 2015.

[29] A. Zwanenburg, S. Leger, M. Vallières, and S. Löck, "Image biomarker standardisation initiative," *Arxiv*, 2016.

[30] R. Paul, S. H. Hawkins, Y. Balagurunathan, M. B. Schabath, R. J. Gillies, L. O. Hall, and D. B. Goldgof, "Deep Feature Transfer Learning in Combination with Traditional Features Predicts Survival Among Patients with Lung Adenocarcinoma.," *Tomography (Ann Arbor, Mich.)*, vol. 2, pp. 388–395, dec 2016.

[31] R. Paul, S. H. Hawkins, M. B. Schabath, R. J. Gillies, L. O. Hall, and D. B. Goldgof, "Predicting malignant nodules by fusing deep features with classical radiomics features," *Journal of Medical Imaging*, vol. 5, p. 1, mar 2018.

[32] R. Paul, S. H. Hawkins, L. O. Hall, D. B. Goldgof, and R. J. Gillies, "Combining deep neural network and traditional image features to improve survival prediction accuracy for lung cancer patients from diagnostic CT," in *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 002570–002575, IEEE, oct 2016.

[33] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, pp. 1345–1359, oct 2010.

[34] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge,” *International Journal of Computer Vision*, vol. 115, pp. 211–252, dec 2015.

[35] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the Inception Architecture for Computer Vision,” *Arxiv*, dec 2015.

[36] P. Devikar, “Transfer Learning for Image Classification of various dog breeds,” *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, vol. 5, no. 12, pp. 2278–1323, 2016.

[37] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, “Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach.,” *Biometrics*, vol. 44, pp. 837–45, sep 1988.

[38] R. Kotikalapudi, “keras-vis,” 2017.

[39] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization,” *Arxiv*, oct 2016.

[40] A. Marusyk and K. Polyak, “Tumor heterogeneity: causes and consequences.,” *Biochimica et biophysica acta*, vol. 1805, pp. 105–17, jan 2010.

[41] R. Fisher, L. Pusztai, and C. Swanton, “Cancer heterogeneity: implications for targeted therapeutics.,” *British journal of cancer*, vol. 108, pp. 479–85, feb 2013.

[42] L. Gay, A.-M. Baker, and T. A. Graham, “Tumour Cell Heterogeneity.,” *F1000Research*, vol. 5, 2016.

[43] D. R. Caswell and C. Swanton, “The role of tumour heterogeneity and clonal cooperativity in metastasis, immune evasion and clinical outcome,” *BMC Medicine*, vol. 15, p. 133, dec 2017.

[44] A. Chatterjee, M. Vallieres, A. Dohan, I. R. Levesque, Y. Ueno, V. Bist, S. Saif, C. Reinhold, and J. Seuntjens, “An empirical approach for avoiding false discoveries when

applying high-dimensional radiomics to small datasets," *IEEE Transactions on Radiation and Plasma Medical Sciences*, pp. 1–1, 2018.

[45] M. Vallières, E. Kay-Rivest, L. J. Perrin, X. Liem, C. Furstoss, N. Khaouam, P. F. Nguyen-Tan, C.-S. Wang, and K. Sultanem, "Data from Head-Neck-PET-CT," *The Cancer Imaging Archive*, 2017.

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," *Arxiv*, 2015.

[47] F. Chollet, "Keras," 2015.

[48] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, and G. Research, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems."

[49] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, and F. Prior, "The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository," *Journal of Digital Imaging*, vol. 26, pp. 1045–1057, dec 2013.

# Chapter 7

# Multi-modal deep learning framework for head & neck cancer outcome prediction

**André Diamant**, Avishek Chatterjee, Martin Vallières, Monica Serban, Yujing Zou, Reza Forghani, George Shenouda & Jan Seuntjens

Manuscript under preparation for submission

## 7.1 Preface

This study builds on the previous work (Chapter 6) by integrating more readily available information into the outcome prediction. The logical step to take after the previous study was to incorporate the PET image and subsequently the clinical information. The current chapter describes the improvements gained from including these other types of pre-treatment data. These results in specific indicate the strength of outcome prediction models, particularly when harnessing pre-treatment data and the combination of multiple forms of input data.

## 7.2 Abstract

As patients receive cancer care, a tremendous amount of data is generated, much of which is often not used to its full potential. In particular, with few exceptions all patients undergo CT imaging and a substantial fraction receive PET imaging. We hypothesized that a novel

deep learning framework could incorporate the pre-treatment PET/CT imaging data of head & neck cancer patients along with relevant clinical variables to perform outcome prediction that exceeds the ability of the current state-of-the-art models. We tested this hypothesis using a training (194 patients) and validation set (106 patients) which were mutually independent and included 4 institutions. The architecture involved a novel training methodology whereupon 2 image input *branches* (PET/CT) were trained independently prior to being merged with a clinical *branch*. For predicting overall survival, our method achieved an AUC of 0.86, a substantial increase ($p$-value = $10^{-3}$) over the previous work (AUC = 0.70), which incorporated CT data only. Furthermore, we evaluated the impact that missing data has on our model to determine its robustness. Our framework yields models that easily combine multiple forms of input information and perform accurate outcome prediction.

## 7.3 Introduction

Radiation therapy is widely used to treat cancer patients, often in conjunction with chemotherapy, surgery, targeted therapy or other alternatives. Over 50% of all patients receive radiation therapy at some point during their cancer treatment [1]. For some cancers, radiation therapy is even more prominent (e.g., 87% of breast cancer patients). Advances in radiation therapy over the past decades have led to survival rates that exceed what was achievable prior to its usage. However, the prognosis of any individual patient is still often poorly determined. This report will focus on patients afflicted with head and neck (H&N) cancer, a group of neoplasms originating from the squamous cells that line the mucosal surfaces of the oral cavity, paranasal sinuses, pharynx or larynx. Long-term survival is often quite poor within these patients (as low as 35% depending on the tumor type) [2–4]. Thus, a model that can accurately discern between high-risk patients and low-risk patients is crucial, so that personalized treatments may eventually become a reality.

The *vast majority* of oncology patients receive some amount of medical imaging prior to treatment, in order to diagnose, localize and facilitate the treatment of their cancer. Two of the most common imaging modalities used in this context are computed tomography (CT) and positron emission tomography (PET). With few exceptions (blood cancers), the CT scan is used for all [5] cancer patients, while the PET scan is used for a substantial fraction. Although the majority of H&N cancer patients have this data available, it is seldom used to its full potential, particularly in outcome prediction. There is a wealth of literature investigating potential predictive image features (*radiomics*) of both modalities individually, but rarely are they combined and investigated together within a deep learning context [6–13]. There is little literature, if any, investigating the usage of both CT and PET input data as a direct input to deep learning techniques, such as convolutional neural networks (CNN) [14]. This study develops a deep learning framework which combines PET and CT image data with possibly relevant clinical information, such as tumor type, age, chemotherapy (or lack thereof) and tumor staging (TMN system [15]) to predict various clinical endpoints.

Within the past decade, deep learning algorithms have been increasingly used for healthcare applications and have achieved great success in a number of tasks, including object detection, semantic segmentation, object classification and CADx (computer-aided diagnosis) [16–26]. Our group showed that a convolutional neural network (CNN) could be applied directly to a patient's CT imaging data and perform outcome prediction of comparable accuracy to a benchmark study on the same data-set [27]. A particularly novel aspect of that study was that the network was trained *de novo* rather than using a transfer learning methodology. This is in stark contrast to the vast majority of published literature. The current study improves on this methodology by introducing both PET images and clinical information into the framework. Although there have been deep learning frameworks developed to combine multiple medical images, to the author's knowledge, none allow for unregistered images or incorporate clinical information. Moreover, few focus on outcome prediction and rather most are concentrated on organ segmentation by combination of multiple MRI modalities [28, 29, 23, 30, 31].

The novel contributions of this work are three-fold. Firstly, we develop a *novel multi-branch deep learning framework*, capable of incorporating multiple forms of input data to perform accurate outcome prediction. Secondly, we propose a *novel training methodology, where each branch of the framework is trained independently prior to combination resulting in a performance improvement*. Finally, *the framework is capable of functioning even if one of the forms of input data is missing*, a commonly occurring situation within outcome prediction and thus a desirable feature within any predictive model.

## 7.4 Results

### 7.4.1 Benchmark study comparison

This study used the same training and validation partitioning scheme as the benchmark study [27]. This allows for direct comparison between the predictive metrics. The results obtained from our multi-branch framework are shown alongside the results of the benchmark study in Table 7.1. *Method A* represents the novel methodology introduced in this paper where the image branches are trained individually prior to introducing the clinical data and the combining multi-layer perceptron (MLP, see Methods for more detail). *Method B* represents training the entire framework from a random initialization, with all branches participating simultaneously. Table 7.1 represents the results obtained when using all available information: PET images, CT images and clinical data. As shown, our framework improves performance for all three outcomes. Using the DeLong test [32], improvement is statistically significant for loco-regional failure (LRF, $p$-value of $10^{-3}$) and overall survival (OS, $p$-value of $10^{-3}$). Our proposed training methodology (Method A) improves the performance when compared to Method B, although not in a statistically significant fashion. The most substantial improvement was for overall survival prediction (from an AUC of 0.70 to 0.86) and thus the subsequent evaluations/analyses are discussed with respect to the OS model.

### 7.4.2 Cross validation

To evaluate the stability and robustness of our results, we performed 5-fold cross validation on the entirety of the patient cohort (all 298 patients). Thus, each network was trained on 238 patients and tested on 60 patients. No hyper-parameter tuning was performed prior to, or between any of the folds. The mean AUC in predicting overall survival across the 5 folds (within the respective validation sets) was 0.83 (range: 0.77 to 0.87). This numerical quantity should not be directly compared to the other results, considering the differing partitioning methodologies.

### 7.4.3 Evaluation of framework with missing data

Figure 7.1 represents the network's behavior if one form of input data is entirely missing (implementation described further in Methods, using *Method A* for training). It is seen that

Table 7.1 **Validation set results compared to Diamant *et al.'s* [27] results on the same patient cohort.** Method A represents the novel methodology introduced in this paper of training each CNN branch individually, prior to introducing the full framework. Method B represents training the entire framework from random initialization. Both methodologies are more comprehensively described in the Methods section. Square brackets represent $p$-values when compared to Diamant *et al.*[27, 32]. Bolded values represent a statistically significant increase.

| | **Distant metastasis (DM)** | **Loco-regional failure (LRF)** | **Overall survival (OS)** |
|---|---|---|---|
| **AUC** | | | |
| *Method A* | 0.93 [0.27] | **0.78 [$10^{-3}$]** | **0.86 [$10^{-3}$]** |
| *Method B* | 0.85 | 0.70 | 0.78 |
| *Diamant et al. [27]* | 0.88 | 0.65 | 0.70 |
| **Specificity** | | | |
| *Present study (A)* | 0.90 | 0.79 | 0.83 |
| *Diamant et al. [27]* | 0.89 | 0.67 | 0.67 |
| **Sensitivity** | | | |
| *Present study (A)* | 0.93 | 0.75 | 0.79 |
| *Diamant et al. [27]* | 0.86 | 0.65 | 0.68 |
| **Balanced Accuracy** | | | |
| *Present study (A)* | 92% | 77% | 81% |
| *Diamant et al. [27]* | 88% | 66% | 68% |

the framework continues to perform in all such cases, albeit with reduced performance. This evaluation also gives intuition into the importance of both PET and clinical data. It is evident that the removal of PET (shown by purple bars) has a substantial impact on the sensitivity of the model, while in contrast the removal of the clinical data (shown by blue bars) has a substantial impact on the specificity of the model. This is further evidence that the two forms of information are complementary in nature. Figure 7.1 represents the behavior of the model when trained with *all* data present, but evaluated with some data missing. Similar analysis was performed to evaluate how the model(s) would behave if one (or more) forms of data were missing from training. This behavior is shown in Figure 7.2. An additional visualization of these results is shown in Figure 7.3. Solid lines represent the presence of both PET and clinical data, dashed lines represent the absence of PET data and dotted lines represent the absence of clinical data. A clear 'clustering' is seen in the ROC behavior, highlighting the importance of both the PET and clinical data (and conversely, the relative insignificance of the CT data).

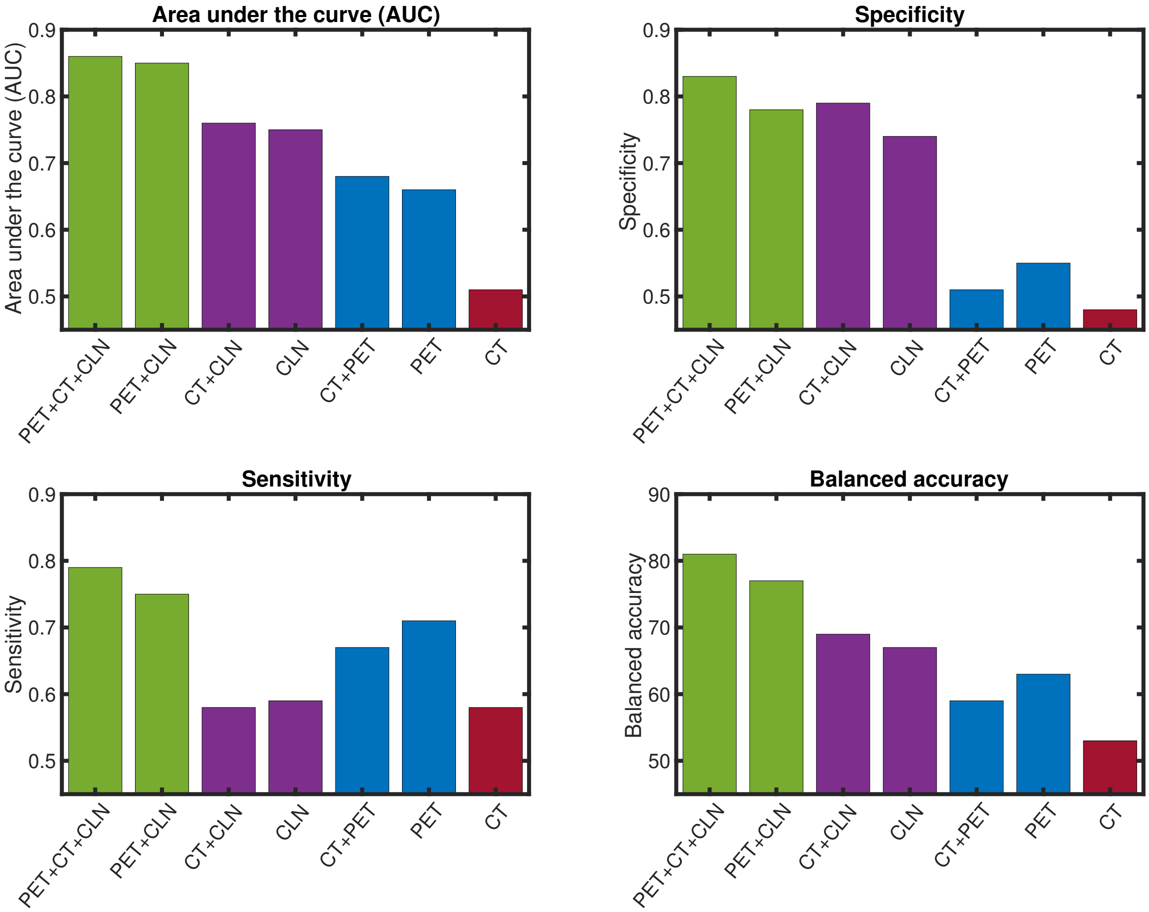


Figure 7.1 **Results when evaluating the model's overall survival prediction performance on the validation set if one branch of input data is entirely missing** (implementation described in Methods). This is a quantitative representation of how the model is capable of functioning even if individual patients are missing a piece of input data, albeit while losing predictive power. Balanced accuracy is described as the average of the specificity and sensitivity. Of particular note is the reduction in sensitivity with discarded PET data (purple) and the reduction in specificity with discarded clinical data (blue).

### 7.4.4 Interpretation of clinical branch

To further understand the behavior of the clinical branch, we created a number of artificial cohorts where the clinical variables were constrained to take a certain value (while all other information remained the same). This was done for tumor stage (all patients set to T4, all patients set to T0), node stage (all N3, all N0) and tumor type (all hypopharyngeal cancer).

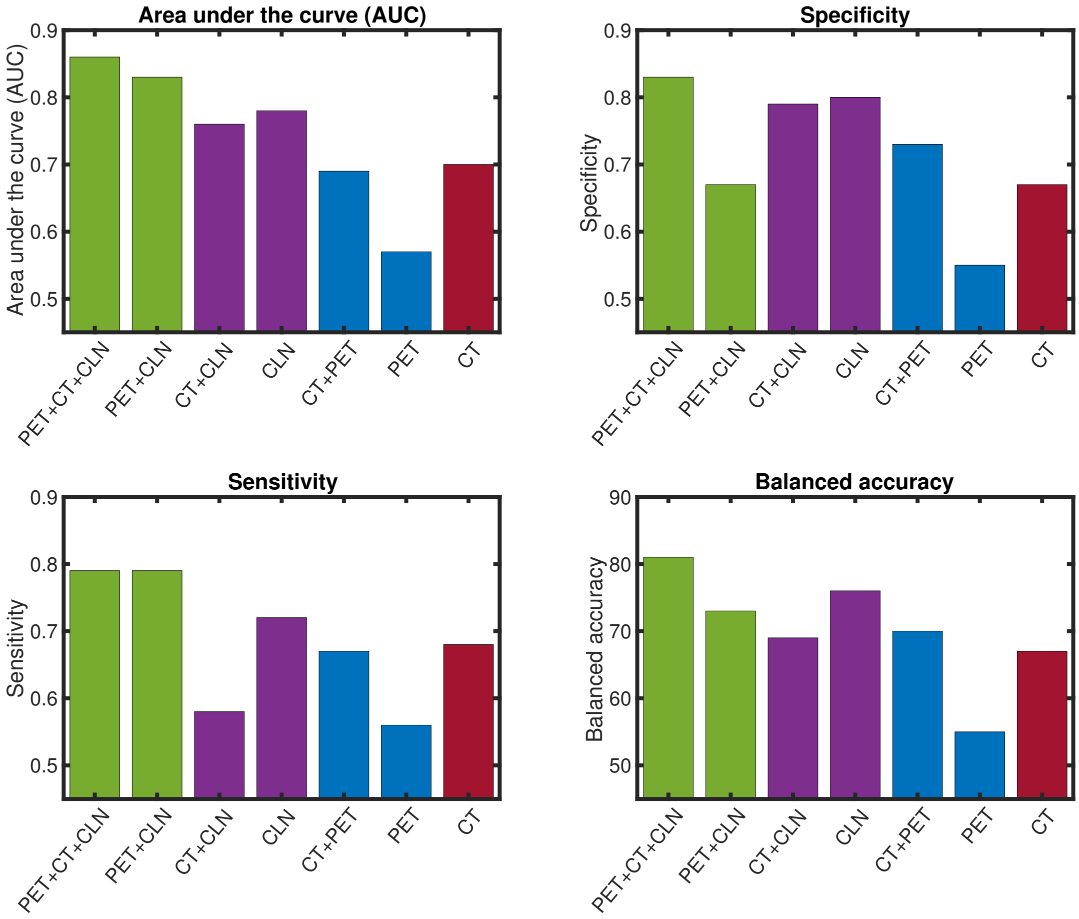


Figure 7.2 **Results when evaluating overall survival prediction performance on the validation set whereupon each model was trained with solely the information depicted on the *x*-axis.** Balanced accuracy is described as the average of the specificity and sensitivity.

Numerical results can be found in the Supplementary Information, however all artificial cohorts behaved in a fashion that is consistent with what the literature would suggest [2–4]. If all patients were changed to be stage T4 or N3, or have hypopharyngeal cancer, the number of predicted high risk patients drastically increased. Similarly, if all patients were changed to be stage T0 or N0, the number of predicted high risk patients dramatically decreased.

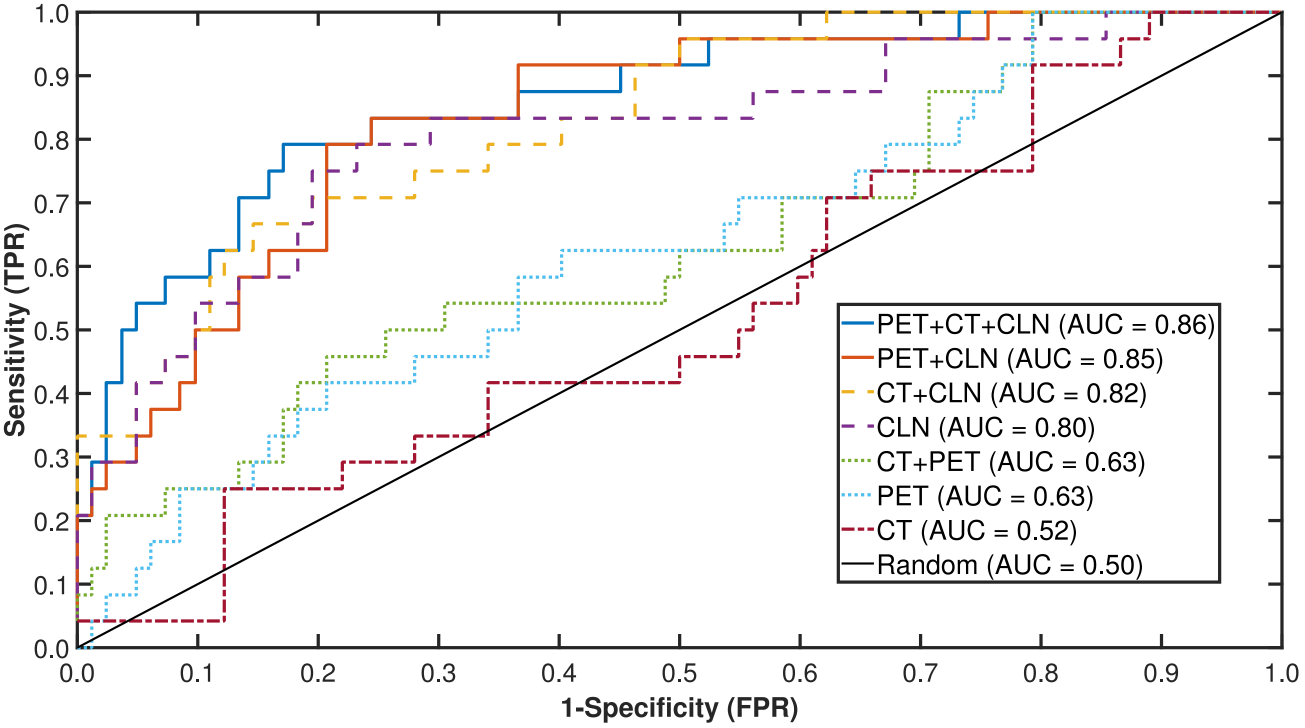


Figure 7.3 **ROC results when evaluating the model's overall survival prediction performance on the validation set if one branch of input is entirely missing** (implementation described in Methods). Solid lines represent the presence of both PET and clinical data, dashed lines represent the absence of PET data and dotted lines represent the absence of clinical data.

## 7.5 Discussion

The results presented in this report show great promise when combining multiple forms of imaging and clinical data to perform outcome prediction. Additionally, the framework developed is easily adaptable to forms of data not included here, such as other imaging modalities (e.g. MRI, ultrasound) or other discrete variables (e.g. radiomics, genomics, biomarker concentrations). Table 7.1 shows statistically significant performance improvements in local control and overall survival. Furthermore, the novel training methodology introduced in this report again results in an improvement over simply training the entire architecture from a random initialization.

While the full framework represents the potential of using all available information, certain patients may not have all the input data types the model was trained on. This problem will only be exasperated as more forms of data are added into such a framework (e.g. genomics). It is for this reason that we have deliberately built our framework (and subsequently evaluated it) in such a fashion that allows for one (or many) forms of data to be missing. Naturally, the performance will decrease if any particular input is missing (Figure 7.1). However, we have shown that the framework does not inherently cease to function. This methodology also allows us to gain intuition of the importance of a particular type of input data. For example, discarding PET data (purple bars) results in a significant drop in sensitivity, while discarding clinical data (blue bars) results in a significant drop in specificity. Effectively, the PET information seems to be particularly relevant in identifying patients with high-risk tumors (resulting in many false negatives if removed) while the clinical information helps in recognizing patients with lower-risk tumors (resulting in many false positives if removed). We note that the performance of a model trained solely on CT images (Figure 7.2) is superior to that of one trained on all 3 forms of data and evaluated on only CT (Figure 7.1). This may indicate that a superior approach to allowing the framework to function with missing data would be to have an ensemble of models, each of which was trained on every possible permutation of data inputs. However, a substantially larger cohort, ideally including patients of varying data availability, would be required to concretely make such conclusions. In contrast, the performance of a model trained solely on PET images (Figure 2) is *inferior* to that of one trained on all 3 forms of data and evaluated on only PET (Figure 1). A possible explanation for this result is that when the model is trained under the blanket influence of all the other forms of data, the model could learn certain predictive correlations that would be unknown without said influence. In other words, the optimization algorithm is able to get closer to the minimum along the ‘PET-axis" when under the influence of other forms of data.

Another advantage to our framework is the lack of a requirement for the input images to be spatially registered. The majority of algorithms that involve multiple images, for outcome prediction or otherwise, require the images to be spatially registered. Not only can this task be time-consuming, challenging or impossible, it also adds an unavoidable source of error. Our framework was specifically developed to not require registration. A desirable consequence of this choice is that the framework also does not necessarily require the PET/CT images to be of the same anatomical location. In addition to the benefits just described, this also improves the robustness of the model by allowing flexibility in precise location of all 3 dimensions of the images used. Currently, both images are taken from the same general anatomical location (i.e.

the GTV), however our framework allows for two images of differing location. This could be used for outcome prediction incorporating information from multiple organs. A more detailed analysis of the impact that tumor type has on the performance of the model was not possible due to the relatively small number of patients. With a larger data-set, this impact could be quantified and further investigated.

This work represents advancements towards the goal of truly personalized treatment. By incorporating more information that is readily available from the clinic, we were able to significantly improve the predictive performance of current state-of-the-art outcome prediction models. Although this model only incorporated as intervention the use of chemotherapy as a binary input, other intervention information could be also introduced. First, this could be simple discrete variables, such as: type of radiation treatment (e.g. VMAT, CyberKnife, tomotherapy), prescription dose, and number of fractions. However, using our framework, one could also implement the dose grid of a patient's radiation treatment plan. By combining the PET/CT images, clinical information and dose grid of a patient in an outcome prediction context, one could introduce a 'quality score' of a treatment plan. This could lead to generative modeling where a truly personalized (and optimized) radiation treatment plan can be created based on patient outcomes rather than an anatomy-based atlas or physical / radiobiological models.

## 7.6 Methods

### 7.6.1 Patient cohorts

Extensive details regarding the patient cohort used in this report are publicly available on The Cancer Imaging Archive (TCIA) [10, 33] repository. Eligible patients were taken from four separate institutions (Hôpital général juif (HGJ), Centre hospitalier universitaire de Sherbooke (CHUS), Hôpital Maisonneuve-Rosemont (HMR) and Centre hospitalier de l'Université de Montréal (CHUM)). The majority received chemotherapy adjuvant to radiotherapy (92%) while the remainder solely received radiation (8%). All patients underwent joint FDG-PET/CT scans, both of which were used in this study. The training set was defined as the patients from HGJ and CHUS, while the validation set was defined as the patients from HMR and CHUM. This was the same distribution used in the two previous studies [10, 27], specifically notable due to the two sets containing patients from independent institutions. Any patients with metastases or recurrent H&N cancer at presentation were excluded, along with any patients receiving palliative care. The median age of patients across the total cohort was 63 years (range: 18-90). The median follow-up period across all patients was 43 months (range: 6-112). Any patients that did not develop cancer recurrence/pass away and had a follow-up period of less than 24 months were discarded.

The clinical variables included in the framework are: age, T-stage, N-stage, tumor type and whether or not chemotherapy was administered. Although HPV-status was available for some patients (40%) it was not incorporated due to the amount of missing data which would have prevented the network from accurately training. The overall TMN stage was not incorporated as the majority of patients (68%) were Stage IV and thus the distinction between T-stage and N-stage was more informative. Similarly, *all* of the patients were M0 and thus M-stage was not included. The outcome distribution for both cohorts is shown in Table 7.2. It is noted that 2 patients from the training cohort were lost to data corruption.

### 7.6.2 Multi-branch network architecture

Our multi-branch framework currently includes 3 separate *branches*. The terminology ‘branch’ is used to indicate distinct forms of input data (e.g. PET, CT or clinical). The terminology is also used to describe the portion of the entire framework which is trained independently prior to merging the branches (described as *Method A*). The framework is based around a combination of convolutional networks (which make up the *branches*) aligned horizontally, followed by a multi-layer perceptron (MLP). As an input, the MLP takes the flattened output of each branch and concatenates them into a single vector. Conceptually, this is depicted in Figures 7.4 and 7.5.

This overarching framework served as a basis for all hyper-parameter tuning. In other words, the concatenation of 3 individual *branches* into a single MLP stayed consistent throughout the report. Furthermore, informed by prior literature [27], the number of convolutional layers within each CNN stayed constant. The following subsections describe each branch individually, a detailed figure of the architecture (including exact sizes for each layer) is shown in the Supplementary Methods.

Table 7.2 **Patient/outcome distribution [10, 33].** Percentage values are with respect to each cohort.

| | Training cohort | Validation cohort |
|---|---|---|
| **Total** | 194 | 106 |
| **Outcome** | | |
| Distant metastasis (DM) | 26 (13%) | 14 (13%) |
| Loco-regional failure (LRF) | 29 (15%) | 16 (15%) |
| Death | 32 (16%) | 24 (23%) |
| **Institution** | | |
| Hôpital général juif (HGJ) | 92 (47%) | - |
| Centre hospitalier universitaire de Sherbooke (CHUS) | 102 (53%) | - |
| Hôpital Maisonneuve-Rosemont (HMR) | - | 41 (39%) |
| Centre hospitalier de l'Université de Montréal (CHUM) | - | 65 (61%) |
| **Age** | | |
| Range | 18 - 88 | 44 - 90 |
| Mean | $62 \pm 8$ | $67 \pm 7$ |
| **Tumor type** | | |
| Oropharynx | 129 (66%) | 77 (73%) |
| Hypopharynx | 5 (3%) | 7 (7%) |
| Nasopharynx | 20 (10%) | 8 (7%) |
| Laryngeal | 36 (19%) | 9 (8%) |
| Unknown | 4 (2%) | 5 (5%) |
| **T-stage** | | |
| T1 | 29 (15%) | 10 (9%) |
| T2 | 65 (34%) | 45 (43%) |
| T3 | 66 (34%) | 28 (26%) |
| T4 | 30 (15%) | 17 (16%) |
| Unknown | 4 (2%) | 6 (6%) |
| **N-stage** | | |
| N0 | 51 (26%) | 9 (8%) |
| N1 | 29 (15%) | 12 (11%) |
| N2 | 108 (56%) | 72 (69%) |
| N3 | 6 (3%) | 13 (12%) |
| **Therapy type** | | |
| Chemotherapy + Radiation therapy | 156 (80%) | 95 (90%) |
| Radiation therapy | 38 (20%) | 11 (10%) |

### Convolutional branches (PET/CT)

First, we describe the convolutional neural networks (henceforth referred to as the *PET branch* and *CT branch*). Each CNN branch uses an architecture similar to that described in our

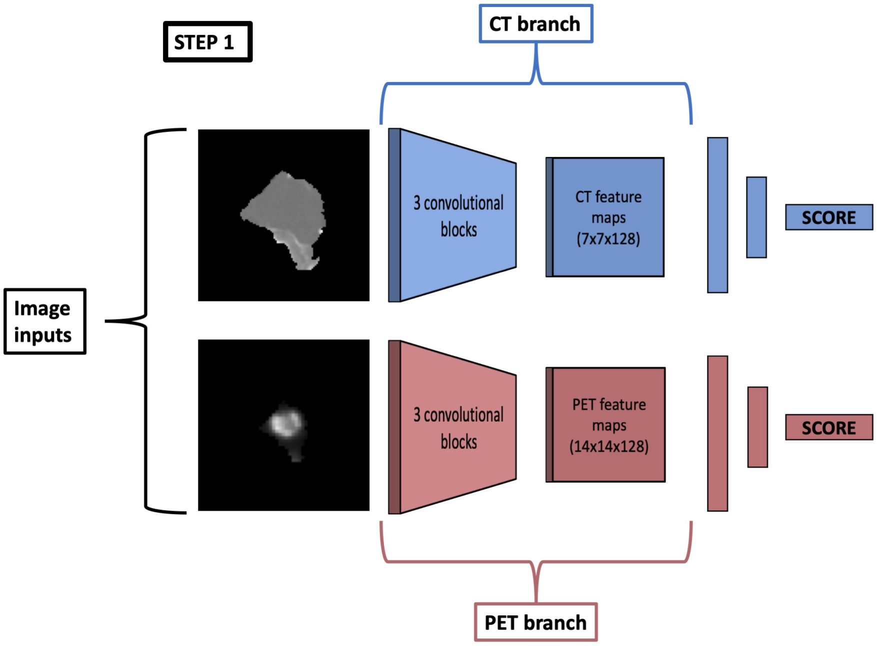


Figure 7.4 **Step 1 of *Method A*. Represents the individual (and separate) training of the PET branch and the CT branch.** This involves a (subsequently discarded) MLP, consisting of two fully connected layers and a dropout layer. Specifics regarding architecture are found in the Supplementary Information.

previous work [27]. This consists of sequential *convolutional blocks*, each of which contains a convolutional layer, followed by an activation function, followed by a pooling layer. The architectures for both the PET and CT branches contain 3 convolutional blocks, whereupon the convolutional layers use filter sizes of 5 × 5 pixels, 3 × 3 pixels and 3 × 3 pixels, respectively. The activation function used was the parametric linear unit (PReLU), empirically determined to be superior to other common activation functions. The max-pooling layers either used a 2 × 2 pixel or 4 × 4 pixel kernel for PET/CT respectively. The discrepancy was due to the difference in image input size (128 × 128 pixels vs. 512 × 512 pixel for PET/CT) and to ensure that the compressed representation at the end of the branches were of comparable size. Both image modalities used the central slice of the tumor, defined as the slice with the maximum number of GTV pixels (same as the previous study).

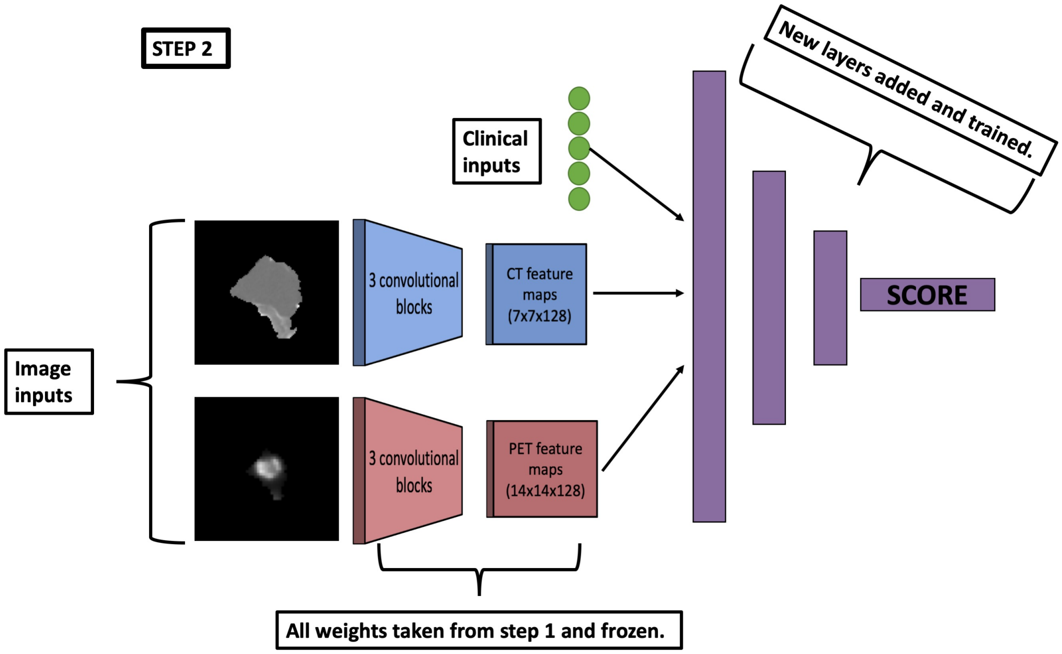

STEP 2
Clinical inputs
New layers added and trained.
Image inputs
3 convolutional blocks
CT feature maps (7x7x128)
SCORE
3 convolutional blocks
PET feature maps (14x14x128)
All weights taken from step 1 and frozen.

Figure 7.5 **Step 2 of *Method A*. Represents the incorporation of learned weights from Step 1, followed by the concatenation of the output feature maps and a *new* MLP being added and subsequently trained.** The new MLP consists of a 256-node fully connected layer, followed by a 128-node fully connected layer followed by a 0.5 dropout layer.

#### Discrete branch (clinical)

The discrete branch is made up of 5 input values. Age was simply entered as a continuous value while T-stage, N-stage and tumor type were encoded in order to facilitate their implementation. Whether or not a patient received chemotherapy/targeted therapy neoadjuvant/concomitant to radiotherapy was encoded as a binary variable. No hyper-parameter tuning was performed for this branch.

### 7.6.3 Training methods

Two training methods were developed and applied to the framework described. Method A represents training the two image branches independently prior to their concatenation. The independent training involves the usage of a 2-layer MLP stacked on top of the convolutional layers . After the branch has been trained individually, that MLP is discarded (Figure 7.4) and the convolutional layers are used as an input for the next step. After both image branches have been trained individually, an MLP is added to the concatenation of both branches and trained, *while the convolutional layer weights remain frozen* (Figure 7.5). Finally, the MLP weights are frozen while the convolutional layer weights (previously frozen) are allowed to vary with a small learning rate ($10^{-5}$). Method B is where the entire architecture is initialized randomly prior to being trained in one step.

### 7.6.4 Evaluation of framework with missing data

It was necessary to develop some methodology to evaluate the framework's performance with missing data as the network is not capable of generating an output with *literally* no input to any given branch. This was trivial for the encoded values of tumor type, T-stage or N-stage as the network was trained with some values of "Unknown" and thus when removing clinical information, they were set to "Unknown". For age, we used the mean age of the entire training set as an input for every patient (63 years). For chemotherapy, the node was set to 'chemotherapy + radiation therapy", as this was the case for the majority of patients (92%) that the network was trained on. For PET/CT, an image of noise was generated and used as an input. To maintain the same range of pixel values, the noise images were generated with the same mean as their respective training cohort.

### 7.6.5 Implementation details

Our framework was built on Python3 using the Keras functional API operating on the tensor manipulation library Tensorflow 2.0 [34, 35]. The final outcome probability (in the combining MLP) was computed via a sigmoid classifier. All network weights were trained using the stochastic optimization algorithm Adam [36]. The mini batch size was 32 and as all outcomes were binary, cross-entropy was used as an objective function. Image augmentation was performed on the PET/CT input data as it has been shown to increase generalization and reduce training bias, particularly on smaller data-sets [14]. No augmentation was performed on the clinical data. Prior to training, each image was randomly flipped (horizontally, vertically or both), rotated a random amount (from 0 - 20°), and shifted a random fraction (from 0 to 0.2 times the total width/height of the image) in a random direction. This resulted in the total training data-set (and thus a single epoch) consisting of 4000 images (each patient's images are augmented roughly 20 times). It is noted that while the PET/CT input data is unregistered, their augmentation was paired. In other words, for each input sample, both the PET and CT images were flipped/rotated/shifted in an identical fashion. If this was not done, the algorithm would be attempting to learn from two ‘different’ augmented tumors. All computation performed throughout this report was done on a pair of NVIDIA GTX 1080TI graphic processing units. Total training time for one network required approximately 3 hours (150 epochs). More details regarding the implementation and the specific range of parameters tested can be found in the Supplementary Methods.

**Data availability** The data-set analyzed throughout this study is publicly available on The Cancer Imaging Archive (TCIA) repository [10, 33, 37].

**Code availability** Code to build, compile, train and evaluate each model will be publicly available on Github.

**Acknowledgements** We acknowledge majority financial support from grant FRQNT-205163 and partial support from the grants NSERC CREATE 432290, CIHR FDN-143257 and MOP-114910. None of the funding sources had any role in the creation or execution of the study.

**Author Contributions** A.D., A.C. and J.S. conceived the initial project: combining multiple imaging modalities and clinical data in a deep learning framework. A.D. developed the framework including the concept of training each branch individually, developed/trained the models, and wrote the manuscript. A.C., J.S., M.V., R.F. and G.S. provided expert knowledge and consultation throughout the course of the research. M.S. and Y.Z. helped with data collection and curation (on the eventual external testing set). All authors edited the manuscript.

**Competing Interests** The authors declare no competing interests related to this work.

## References

[1] R. Atun, D. A. Jaffray, M. B. Barton, F. Bray, M. Baumann, B. Vikram, T. P. Hanna, F. M. Knaul, Y. Lievens, T. Y. M. Lui, M. Milosevic, B. O’Sullivan, D. L. Rodin, E. Rosenblatt,

J. Van Dyk, M. L. Yap, E. Zubizarreta, and M. Gospodarowicz, “Expanding global access to radiotherapy.,” *The Lancet. Oncology*, vol. 16, pp. 1153–86, sep 2015.

[2] S.-A. Yeh, “Radiotherapy for head and neck cancer.,” *Seminars in plastic surgery*, vol. 24, pp. 127–36, may 2010.

[3] S. S. Baxi, L. C. Pinheiro, S. M. Patil, D. G. Pfister, K. C. Oeffinger, and E. B. Elkin, “Causes of death in long-term survivors of head and neck cancer,” *Cancer*, vol. 120, pp. 1507–1513, may 2014.

[4] A. Ferlito, A. R. Shaha, C. E. Silver, A. Rinaldo, and V. Mondin, “Incidence and Sites of Distant Metastases from Head and Neck Cancer,” *ORL*, vol. 63, no. 4, pp. 202–207, 2001.

[5] Canadian Institute for Health Information, “Medical Imaging Report,” tech. rep., Camadian Institute for Health Information, 2012.

[6] H. J. W. L. Aerts, E. R. Velazquez, R. T. H. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebers, M. M. Rietbergen, C. R. Leemans, A. Dekker, J. Quackenbush, R. J. Gillies, and P. Lambin, “Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach,” *Nature Communications*, vol. 5, p. 4006, dec 2014.

[7] M. Vallières, C. R. Freeman, S. R. Skamene, and I. El Naqa, “A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities.,” *Physics in medicine and biology*, vol. 60, no. 14, pp. 5471–96, 2015.

[8] J. E. van Timmeren, R. T. Leijenaar, W. van Elmpt, B. Reymen, C. Oberije, R. Monshouwer, J. Bussink, C. Brink, O. Hansen, and P. Lambin, “Survival prediction of non-small cell lung cancer patients using radiomics analyses of cone-beam CT images,” *Radiotherapy and Oncology*, vol. 123, no. 3, pp. 363–369, 2017.

[9] T. P. Coroller, P. Grossmann, Y. Hou, E. Rios, R. T. H. Leijenaar, G. Hermann, P. Lambin, B. Haibe-kains, R. H. Mak, and H. J. W. L. Aerts, “CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma,” *Radiotherapy and Oncology*, vol. 114, no. 3, pp. 345–350, 2015.

[10] M. Vallières, E. Kay-Rivest, L. J. Perrin, X. Liem, C. Furstoss, H. J. Aerts, N. Khaouam, P. F. Nguyen-Tan, C. S. Wang, K. Sultanem, J. Seuntjens, and I. El Naqa, "Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer," *Scientific Reports*, vol. 7, no. 1, pp. 1–33, 2017.

[11] C. Parmar, P. Grossmann, D. Rietveld, M. M. Rietbergen, P. Lambin, and H. J. W. L. Aerts, "Radiomic Machine-Learning Classifiers for Prognostic Biomarkers of Head and Neck Cancer," *Frontiers in Oncology*, vol. 5, p. 272, dec 2015.

[12] C. Parmar, P. Grossmann, J. Bussink, P. Lambin, and H. J. W. L. Aerts, "Machine Learning methods for Quantitative Radiomic Biomarkers," *Scientific Reports*, vol. 5, p. 13087, oct 2015.

[13] C. Parmar, R. T. H. Leijenaar, P. Grossmann, E. Rios Velazquez, J. Bussink, D. Rietveld, M. M. Rietbergen, B. Haibe-Kains, P. Lambin, and H. J. W. L. Aerts, "Radiomic feature clusters and Prognostic Signatures specific for Lung and Head & Neck cancer," *Nature Publishing Group*, 2015.

[14] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[15] American Joint Committee on Cancer, "AJCC - Cancer Staging Manual," *JAMA*, vol. 304, pp. 1726–1727, 2010.

[16] R. C. Deo, "Machine Learning in Medicine," *Circulation*, vol. 132, pp. 1920–1930, nov 2015.

[17] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects.," *Science (New York, N.Y.)*, vol. 349, pp. 255–60, jul 2015.

[18] Dinggang Shen, G. Wu, H.-I. Suk, D. Shen, G. Wu, and H.-I. Suk, "Deep Learning in Medical Image Analysis," *Annual Review of Biomedical Engineering*, vol. 19, no. 1, pp. 221–248, 2017.

[19] T. Brosch, L. Y. W. Tang, Y. Youngjin Yoo, D. K. B. Li, A. Traboulsee, and R. Tam, "Deep 3D Convolutional Encoder Networks With Shortcuts for Multiscale Feature Integration Applied to Multiple Sclerosis Lesion Segmentation.," *IEEE transactions on medical imaging*, vol. 35, pp. 1229–1239, may 2016.

[20] Q. Qi Dou, H. Hao Chen, L. Lequan Yu, L. Lei Zhao, J. Jing Qin, D. Defeng Wang, V. C. Mok, L. Lin Shi, and P.-A. Pheng-Ann Heng, "Automatic Detection of Cerebral Microbleeds From MR Images via 3D Convolutional Neural Networks.," *IEEE transactions on medical imaging*, vol. 35, pp. 1182–1195, may 2016.

[21] G. van Tulder and M. de Bruijne, "Combining Generative and Discriminative Representation Learning for Lung CT Analysis With Convolutional Restricted Boltzmann Machines.," *IEEE transactions on medical imaging*, vol. 35, pp. 1262–1272, may 2016.

[22] J. Kleesiek, G. Urban, A. Hubert, D. Schwarz, K. Maier-Hein, M. Bendszus, and A. Biller, "Deep MRI brain extraction: A 3D convolutional neural network for skull stripping," *NeuroImage*, vol. 129, pp. 460–469, apr 2016.

[23] W. Zhang, R. Li, H. Deng, L. Wang, W. Lin, S. Ji, and D. Shen, "Deep convolutional neural networks for multi-modality isointense infant brain image segmentation," *NeuroImage*, vol. 108, pp. 214–224, mar 2015.

[24] B. Q. Huynh, H. Li, and M. L. Giger, "Digital mammographic tumor classification using transfer learning from deep convolutional neural networks," *Journal of Medical Imaging*, vol. 3, p. 034501, aug 2016.

[25] N. Antropova, B. Q. Huynh, and M. L. Giger, "A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets," *Med. Phys.*, vol. 44, no. 10, 2017.

[26] S. Trebeschi, J. J. M. Van Griethuysen, D. M. J. Lambregts, M. J. Lahaye, C. Parmar, F. C. H. Bakers, N. H. G. M. Peters, R. G. H. Beets-Tan, and H. J. W. L. Aerts, "Deep Learning for Fully-Automated Localization and Segmentation of Rectal Cancer on Multiparametric MR OPEN Background Work," *Scientific reports*, vol. 7, no. 1, p. 5301, 2017.

[27] A. Diamant, A. Chatterjee, M. Vallières, G. Shenouda, and J. Seuntjens, "Deep learning in head & neck cancer outcome prediction," *Scientific reports*, no. August 2018, pp. 1–10, 2019.

[28] J. Dolz, K. Gopinath, J. Yuan, H. Lombaert, C. Desrosiers, and I. B. Ayed, “HyperDense-Net: A hyper-densely connected CNN for multi-modal image segmentation,” tech. rep., Ecole de technologie Superieure, 2018.

[29] M. Aygün, Y. H. Şahin, and G. Ünal, “Multi Modal Convolutional Neural Networks for Brain Tumor Segmentation,” *Arxiv*, 2018.

[30] A. M. DSouza, L. Chen, Y. Wu, A. Z. Abidin, C. Xu, and A. Wismüller, “MRI tumor segmentation with densely connected 3D CNN,” in *Medical Imaging 2018: Image Processing* (E. D. Angelini and B. A. Landman, eds.), vol. 10574, p. 50, SPIE, mar 2018.

[31] Y. Xu, A. Hosny, R. Zeleznik, C. Parmar, T. Coroller, I. Franco, R. H. Mak, and H. J. Aerts, “Deep Learning Predicts Lung Cancer Treatment Response from Serial Medical Imaging,” *Clinical Cancer Research*, 2019.

[32] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, “Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach.,” *Biometrics*, vol. 44, pp. 837–45, sep 1988.

[33] M. Vallières, E. Kay-Rivest, L. J. Perrin, X. Liem, C. Furstoss, N. Khaouam, P. F. Nguyen-Tan, C.-S. Wang, and K. Sultanem, “Data from Head-Neck-PET-CT,” *The Cancer Imaging Archive*, 2017.

[34] F. Chollet, “Keras,” 2015.

[35] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, and G. Research, “TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems.”

[36] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” *Arxiv*, dec 2014.

[37] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, and F. Prior, “The Cancer Imaging Archive (TCIA):

Maintaining and Operating a Public Information Repository," *Journal of Digital Imaging*, vol. 26, pp. 1045–1057, dec 2013.

# Chapter 8

# Summary and future directions

This chapter summarizes the work presented throughout this thesis, highlighting the novelty of each work and concludes by discussing how outcome prediction could evolve within radiation oncology eventually resulting in generative modeling within a personalized treatment context.

## 8.1 Summary

The entirety of this thesis focused on using pre-treatment data to predict a cancer patient's prognosis after they receive radiation therapy. There are numerous types of methodologies to predict outcomes and numerous types of medical data. The two primary objectives incorporated different methodologies (statistical correlations/deep learning), different types of data (dose distributions/medical images) applied to different types of cancer patients (non-small cell lung cancers/head & neck cancers). In order to properly understand the presented work, it was first necessary to detail the statistical tests & metrics used (Chapter 2) along with the basic theory behind deep learning in addition to the current deep learning applications within the field of medical physics (Chapter 3). Then, the thesis described the found correlation between dose metrics outside the PTV and distant metastasis rates for NSCLC patients who received conventional (Chapter 4) or robotic-mounted (Chapter 5) radiotherapy. The latter half of the thesis described the usage of medical images (both CT & PET) as input into a deep learning framework applied to head & neck cancer patients (Chapters 6 & 7).

### 8.1.1 Dose outside the PTV as a predictor of DM in NSCLC SBRT patients

Often, dosimetric outcome prediction studies involve analysis of the dose to conventionally contoured structures, such as the PTV or the OARs. In contrast, the first goal of this thesis investigated the dose *outside* the PTV and whether it could predict DM in NSCLC SBRT patients. The initial hypothesis was that microscopic cancer cells could be present outside of the PTV and thus the dose to that region could be of importance. An isotropic PTV growing algorithm was built in order to facilitate the analysis of this region. Two types of regions were investigated; continuous shells of thickness $x$ and discrete 1 mm thick rings $x$ mm away from the edge of the PTV. Both the mean and median dose to these regions were computed.

Both chapters represented the first time that the region directly outside the PTV was investigated in a formulaic fashion. They also resulted in the discovery of an association which through prospective trials could lead to a clinical change in practice. Chapter 4 considered solely NSCLC patients who received conventional treatment modalities, specifically conformal radiotherapy (CRT) or volumetric modulated arc radiotherapy (VMAT). A relationship was found between the mean dose to a continuous region of thickness 30 mm and the incidence of distant metastasis. Of particular note was the discovery of a 'threshold dose'. Patients who received higher than said threshold dose had substantially reduced DM rates compared to those who received below. Chapter 5 explored whether a similar correlation was found for a unique robotic-mounted treatment modality (CyberKnife). Interestingly, no explicit correlation was found between the dose metrics and any region analyzed. However, nearly all of the CyberKnife patients were found to be above the previously determined threshold dose. Given the sigmoidal behaviour of tumor response [1, 2], the lack of correlation was consistent with what we would expect. Despite the lack of explicit correlation in the dose metrics, CyberKnife was still found to perform significantly better with respect to distant control *as well as* local control. We hypothesized that this could be due to numerous factors, including the non-coplanarity and the robust motion tracking employed exclusively by the CyberKnife.

This work demonstrated that the spatial information within the dose distribution is of relevance to outcome prediction research. Although the majority of studies (and modern treatment planning practice) solely consider dose prescriptions/constraints or relatively generalized

NTCP/TCP models, these results suggest that more complex spatial personalization of treatment plans may result in superior outcomes. Notably, this research suggests that attention ought to be taken to not just the region suspected to be cancerous, but also the most proximal regions.

### 8.1.2 Deep learning framework for outcome prediction in H&N patients

The latter half of this thesis investigated a different outcome prediction methodology applied to a different cohort of cancer patients. In this case, a deep learning framework was built (and subsequently developed) to predict the prognosis of head & neck cancer patients. Chapter 6 presents the first architecture that was built, one that only incorporates the CT. This, along with the concept of training *de novo*, were deliberate decisions in an attempt to start simple. This work represented the first end-to-end convolutional neural network that could perform outcome prediction on H&N cancer patients using solely a pre-treatment CT of the GTV with performance comparable to the current state-of-the-art. This architecture was also shown to recognize images with similar radiomic features to those found to be predictive in the previous study [3], without being explicitly told their definition. Visualization tools were also developed and used to further interpret the quantitative results.

Chapter 7 further developed the deep learning architecture by incorporating both the PET image and the clinical information. This work represented the novel application (and refinement) of a deep learning algorithm to outcome prediction within oncology. Specifically, the multi-modal framework and the explicit ability to function with missing information present findings which are crucial when building an outcome prediction model capable of functioning within a healthcare environment. It was shown that significant increases in performance were gained, particularly when predicting local recurrence or overall survival. This study also incorporated a novel training methodology where each modality of input data was trained independently prior to their combination. Also shown was the ability to interpret the model's dependence on each type of data, along with its ability to function even when a data modality was missing. An external dataset is currently being curated for further testing, however this portion of the thesis showed great promise in the future of outcome prediction within the field.

## 8.2 Future directions

The final portion of this thesis presents a speculative discussion regarding the future of outcome prediction within radiation oncology.

### 8.2.1 Outcome prediction within radiation oncology

As presented throughout this thesis and discussed in the preceding section, the vast majority (if not all) of ‘outcome prediction’ models have not been clinically utilized yet. It is my firm belief that within the next few years, substantial steps will be made towards leveraging the full power of these models. There are a number of initiatives within (and outside of) radiation oncology pursuing this aim. MEDomics’ vision is to ‘provide research scientists with integrated, end-to-end, open-source multi-omics computation tools for state-of-the-art outcome prediction modeling in oncology’ [4]. To that end, their consortium is developing an open-source computation platform and novel algorithms to detect/understand/predict a number of clinical endpoints in the treatment of cancers. CERR (a *Computational Environment for Radiotherapy Research*) was developed to provide a tool which could facilitate the combination of various types of information found within radiation oncology. The motivation was to make reproducing research within radiation oncology treatment planning far easier, as often it’s difficult to even review results from other researchers due to the lack of standardization. More specifically, the format used in CERR was designed to be a ‘compact, self-describing object containing all the treatment plan archive data’ [5].

In order for a model to be even considered as a clinical decision support tool, above all else it must be *trusted*. For a model to be trusted, it requires extensive validation on a very large body of data. The majority of mainstream outcome prediction medical models were built on hundreds of thousands of data-points. For example, Esteva et al. [6] used 129,450 images to perform binary classification on images of suspected skin cancer. They showed that a CNN could achieve performance on par with 21 board-certified dermatologists. Within a hospital setting, often acquiring anywhere close to one thousand data-points, let alone hundreds of thousands is a struggle. This is one of the primary reasons why outcome prediction models are not currently used in the radiation oncology clinic, but many steps are being taken (e.g. clinical data repositories or ‘data lakes’) in an attempt to rectify this problem. However, it is still unclear as to how many data-points are required for an end-to-end outcome prediction model to be clinically validated

& subsequently accepted. Perhaps with numerous types of information (images, pathology, genomic sequencing, etc...) one thousand patients, albeit with complete data, could be sufficient.

Ideally, the field (and to a broader extent, healthcare) will reach a point where a predictive validated model exists for each cancer histology & treatment modality. Then, each patient could be assigned a computationally determined pre-treatment 'risk score', indicative of their most likely prognosis. This could then inform treatment options, such as whether escalating the dose could reduce recurrence probability or whether a less aggressive prescription could reduce negative side-effects. By incorporating models within all of oncology, the care team in conjunction with the patient could also determine whether an alternative treatment or some combination of treatments (e.g. chemotherapy, targeted therapy, surgery, etc...) could be ultimately beneficial. The first iteration of models to be employed in the clinic will likely be focused on targeting outliers; patients whose predicted treatment response is significantly distinct from the more general population and perhaps an additional medical opinion or treatment alterations are required.

Finally, the full potential of generative modeling (and thus, deep learning) will not be realized until it is possible to automatically create a treatment plan for each patient, optimized for that patient's particular prognosis on that patient's data. This could be done by training a model (on a large sample size) to perform outcome prediction based on an input treatment plan (including dose distribution, chemotherapy dose, etc...), medical images, clinical variables, and more... After the model is trained (as a classification algorithm), the output classification score could become an optimization metric, whereupon the treatment plan becomes the new output. For new patients, the purpose of the model is to generate a treatment plan, given the influence of the other input data (medical images, etc...) while optimizing the classification score. Prior to generating an 'outcomes-optimized' treatment plan, it is necessary to fully understand how the patient's information (be it imaging, dose-volume or otherwise) relates to the outcome. The work presented within this thesis does not explicitly include generative modeling; however, represents steps towards this vision, one where a patient's treatment is *truly* personalized and explicitly optimized for their unique cancer.

# References

[1] E. J. Hall and S. Willson, *Radiobiology for the Radiologist*. Philadelphia: Wolters Kluwer, 2012.

[2] M. K. Martel, R. K. Ten Haken, M. B. Hazuka, M. L. Kessler, M. Strawderman, a. T. Turrisi, T. S. Lawrence, B. a. Fraass, and a. S. Lichter, "Estimation of tumor control probability model parameters from 3-D dose distributions of non-small cell lung cancer patients.," *Lung cancer (Amsterdam, Netherlands)*, vol. 24, no. 1, pp. 31–37, 1999.

[3] M. Vallières, E. Kay-Rivest, L. J. Perrin, X. Liem, C. Furstoss, H. J. Aerts, N. Khaouam, P. F. Nguyen-Tan, C. S. Wang, K. Sultanem, J. Seuntjens, and I. El Naqa, "Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer," *Scientific Reports*, vol. 7, no. 1, pp. 1–33, 2017.

[4] "MEDomics Consortium."

[5] J. O. Deasy, A. I. Blanco, and V. H. Clark, "CERR: A computational environment for radiotherapy research," *Medical Physics*, vol. 30, pp. 979–985, may 2003.

[6] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, pp. 115–118, feb 2017.

# Appendix A

# Supplementary Information - Chapter 4

## Region of Interest Creation Algorithm

First, the PTV point cloud (a list of voxel positions contained directly on the PTV contour) was used to create a 3D volume and superimposed upon the dose grid (contained within the DICOM dose file). A convex hull encompassing the PTV, defined as the smallest convex set that contains the PTV, was generated using this point cloud. Second, the PTV point cloud was "grown" by a variable amount. The growing was facilitated by determining the centroid of the PTV and generating a direction vector between the centroid and every point. The direction vectors were scaled by a variable amount. Third, the grown point cloud was used to generate a convex hull of the isotropically grown PTV. This convex hull was superimposed with the lung contour mask, ensuring that the region of interest did not encroach on the contralateral lung and only included voxels within the ipsilateral lung. Next, XOR logic was used to create a mask which only considered points within the grown region but not within the PTV itself. This new ring-shaped region is hereafter referred to as the continuous cumulative region of interest of width x ($ROI_{cont}$(x mm)). x was varied in a range of 1 mm and 100 mm in 1 mm increments and represents the number of millimeters each direction vector was grown by. This maximal range was chosen as to be significantly beyond the range which is suggested to contain microscopic spread [6,7,13]. Additionally, regions representing a differential region of thickness 1 mm were created. This region was also varied in relative distance from the PTV boundary from 1 mm to 100 mm in 1 mm increments and is hereafter referred to as $ROI_{diff}$(x mm). The ROI masks were then applied to the dose grid, from which all voxels contained within each ROI were analyzed.

# Statistical Analysis w.r.t. RP

The summary of statistical analysis w.r.t RP is shown in Table A.1.

Table A.1 **Summary of statistical analysis w.r.t RP**

| | AUC | OR [95% CI ] | *p*-value |
|---|---|---|---|
| **Differential 1 mm thick region at distance x** [$ROI_{diff}(16\ mm)$] | | | |
| Mean dose received by $ROI_{diff}(16\ mm)$ | 0.50 | 1.16(0.55, 2.62) | 0.72 |
| Median dose received by $ROI_{diff}(16\ mm)$ | 0.50 | 1.18(0.55, 2.63) | 0.72 |
| **Continuous 30 mm spherical region** [$ROI_{cont}(30\ mm)$] | | | |
| Mean dose received by $ROI_{cont}(30\ mm)$ | 0.50 | 1.23(0.55, 2.37) | 0.61 |
| Median dose received by $ROI_{cont}(30\ mm)$ | 0.50 | 1.11(0.40, 3.08) | 0.85 |
| **PTV volume** | 0.62 | 2.50(1.17,5.32) | 0.02 |
| **Homogeneity Index** | 0.50 | / | / |
| **Median dose received by PTV** | 0.55 | / | / |
| **Mean dose received by PTV** | 0.55 | / | / |

The summary of statistical analysis w.r.t local control is shown in Table A.2.

Table A.2 **Summary of statistical analysis w.r.t local control**

| | AUC | OR [95% CI ] | *p*-value |
|---|---|---|---|
| **Differential 1 mm thick region at distance x** [$ROI_{diff}(16\ mm)$] | | | |
| Mean dose received by $ROI_{diff}(16\ mm)$ | 0.50 | 1.48(0.71, 3.79) | 0.30 |
| Median dose received by $ROI_{diff}(16\ mm)$ | 0.50 | 1.47(0.70, 3.78) | 0.31 |
| **Continuous 30 mm spherical region** [$ROI_{cont}(30\ mm)$] | | | |
| Mean dose received by $ROI_{cont}(30\ mm)$ | 0.51 | 1.64(0.81, 3.42) | 0.20 |
| Median dose received by $ROI_{cont}(30\ mm)$ | 0.51 | 1.65(0.80, 3.43) | 0.20 |
| **PTV volume** | 0.60 | 1.48(0.76,2.85) | 0.25 |
| **Homogeneity Index** | 0.50 | / | / |
| **Median dose received by PTV** | 0.51 | / | / |
| **Mean dose received by PTV** | 0.50 | / | / |

# Sub-cohort statistical analysis

The summary of statistical analysis w.r.t distant metastasis for each sub-cohort is shown in Tables A.3, A.4, A.5, A.6, A.7 and A.8.

Table A.3 **Summary of statistical analysis (sub-cohort A verified by biopsy)**

| | AUC[a] | OR[b] [95% CI[c]] | $p$-value |
|---|---|---|---|
| **Differential 1 mm thick region at distance x** [$ROI_{diff}(16\ mm)$] | | | |
| Mean dose received by $ROI_{diff}(16\ mm)$ | 0.79 | 0.12(0.04,0.34) | $< 10^{-5}$ |
| Median dose received by $ROI_{diff}(16\ mm)$ | 0.77 | 0.16(0.06,0.41) | $< 10^{-5}$ |
| **Continuous 30 mm spherical region** [$ROI_{cont}(30\ mm)$] | | | |
| Mean dose received by $ROI_{cont}(30\ mm)$ | 0.80 | 0.10(0.03,0.31) | $< 10^{-5}$ |
| Median dose received by $ROI_{cont}(30\ mm)$ | 0.77 | 0.15(0.06,0.39) | $< 10^{-5}$ |
| **PTV volume** | 0.74 | 0.22(0.09,0.53) | $10^{-4}$ |
| **Homogeneity Index** | 0.51 | / | / |
| **Median dose received by PTV** | 0.50 | / | / |
| **Mean dose received by PTV** | 0.50 | / | / |

[a] with respect to distant metastasis
[b] with respect to distant metastasis
[c] Confidence Interval

Table A.4 **Summary of statistical analysis (sub-cohort B restricted between first and third quartile of follow-up period)**

| | AUC[a] | OR[b] [95% CI[c]] | $p$-value |
|---|---|---|---|
| **Differential 1 mm thick region at distance x** [$ROI_{diff}(16\ mm)$] | | | |
| Mean dose received by $ROI_{diff}(16\ mm)$ | 0.84 | 0.05(0.01,0.22) | $< 10^{-5}$ |
| Median dose received by $ROI_{diff}(16\ mm)$ | 0.77 | 0.14(0.05,0.38) | $< 10^{-5}$ |
| **Continuous 30 mm spherical region** [$ROI_{cont}(30\ mm)$] | | | |
| Mean dose received by $ROI_{cont}(30\ mm)$ | 0.83 | 0.05(0.01,0.22) | $< 10^{-5}$ |
| Median dose received by $ROI_{cont}(30\ mm)$ | 0.83 | 0.06(0.01,0.25) | $< 10^{-5}$ |
| **PTV volume** | 0.72 | 0.24(0.10,0.62) | $10^{-4}$ |
| **Homogeneity Index** | 0.54 | / | / |
| **Median dose received by PTV** | 0.55 | / | / |
| **Mean dose received by PTV** | 0.54 | / | / |

[a] with respect to distant metastasis
[b] with respect to distant metastasis
[c] Confidence Interval

Table A.5 **Summary of statistical analysis (sub-cohort C restricted between first and third quartile of PTV volume)**

| | AUC[a] | OR[b] [95% CI[c]] | $p$-value |
|---|---|---|---|
| **Differential 1 mm thick region at distance x** [$ROI_{diff}(16\ mm)$] | | | |
| Mean dose received by $ROI_{diff}(16\ mm)$ | 0.76 | 0.15(0.04,0.53) | $10^{-4}$ |
| Median dose received by $ROI_{diff}(16\ mm)$ | 0.74 | 0.20(0.06,0.65) | $10^{-3}$ |
| **Continuous 30 mm spherical region** [$ROI_{cont}(30\ mm)$] | | | |
| Mean dose received by $ROI_{cont}(30\ mm)$ | 0.74 | 0.18(0.05,0.68) | $10^{-3}$ |
| Median dose received by $ROI_{cont}(30\ mm)$ | 0.71 | 0.28(0.09,0.82) | $10^{-2}$ |
| **PTV volume** | 0.50 | / | / |
| **Homogeneity Index** | 0.56 | / | / |
| **Median dose received by PTV** | 0.51 | / | / |
| **Mean dose received by PTV** | 0.51 | / | / |

[a] with respect to distant metastasis
[b] with respect to distant metastasis
[c] Confidence Interval

Table A.6 **Summary of statistical analysis (sub-cohort D restricted to patients receiving three fractions of radiation)**

| | AUC[a] | OR[b] [95% CI[c]] | $p$-value |
|---|---|---|---|
| **Differential 1 mm thick region at distance x** [$ROI_{diff}(x)$] | | | |
| Mean dose received by $ROI_{diff}(16\ mm)$ | 0.82 | 0.07(0.02,0.27) | $< 10^{-5}$ |
| Median dose received by $ROI_{diff}(16\ mm)$ | 0.80 | 0.10(0.03,0.33) | $< 10^{-5}$ |
| **Continuous 30 mm spherical region** [$ROI_{cont}(30\ mm)$] | | | |
| Mean dose received by $ROI_{cont}(30\ mm)$ | 0.83 | 0.06(0.01,0.26) | $< 10^{-5}$ |
| Median dose received by $ROI_{cont}(30\ mm)$ | 0.81 | 0.09(0.03,0.31) | $< 10^{-5}$ |
| **PTV volume** | 0.72 | 0.23(0.08,0.63) | $10^{-4}$ |
| **Homogeneity Index** | 0.53 | / | / |
| **Median dose received by PTV** | 0.57 | / | / |
| **Mean dose received by PTV** | 0.57 | / | / |

[a] with respect to distant metastasis
[b] with respect to distant metastasis
[c] Confidence Interval

Table A.7 **Summary of statistical analysis (sub-cohort E restricted to patients treated at MUHC)**

| | AUC[a] | OR[b] [95% CI[c]] | $p$-value |
|---|---|---|---|
| **Differential 1 mm thick region at distance x** [$ROI_{diff}(x)$] | | | |
| Mean dose received by $ROI_{diff}(16\ mm)$ | 0.75 | 0.23(0.07,0.68) | $< 10^{-3}$ |
| Median dose received by $ROI_{diff}(16\ mm)$ | 0.76 | 0.19(0.06,0.58) | $< 10^{-4}$ |
| **Continuous 30 mm spherical region** [$ROI_{cont}(30\ mm)$] | | | |
| Mean dose received by $ROI_{cont}(30\ mm)$ | 0.79 | 0.14(0.04,0.49) | $< 10^{-4}$ |
| Median dose received by $ROI_{cont}(30\ mm)$ | 0.78 | 0.15(0.05,0.53) | $< 10^{-4}$ |
| **PTV volume** | 0.80 | 0.06(0.01,0.40) | $10^{-5}$ |
| **Homogeneity Index** | 0.53 | / | / |
| **Median dose received by PTV** | 0.54 | / | / |
| **Mean dose received by PTV** | 0.53 | / | / |

[a] with respect to distant metastasis
[b] with respect to distant metastasis
[c] Confidence Interval

Table A.8 **Summary of statistical analysis (sub-cohort F restricted to patients treated at CHUM)**

| | AUC[a] | OR[b] [95% CI[c]] | $p$-value |
|---|---|---|---|
| **Differential 1 mm thick region at distance x** [$ROI_{diff}(x)$] | | | |
| Mean dose received by $ROI_{diff}(16\ mm)$ | 0.81 | 0.10(0.02,0.26) | $< 10^{-5}$ |
| Median dose received by $ROI_{diff}(16\ mm)$ | 0.73 | 0.23(0.08,0.64) | $< 10^{-3}$ |
| **Continuous 30 mm spherical region** [$ROI_{cont}(30\ mm)$] | | | |
| Mean dose received by $ROI_{cont}(30\ mm)$ | 0.75 | 0.15(0.05,0.48) | $< 10^{-4}$ |
| Median dose received by $ROI_{cont}(30\ mm)$ | 0.72 | 0.26(0.09,0.72) | $< 10^{-3}$ |
| **PTV volume** | 0.59 | 0.38(0.15,0.96) | 0.03 |
| **Homogeneity Index** | 0.53 | / | / |
| **Median dose received by PTV** | 0.53 | / | / |
| **Mean dose received by PTV** | 0.52 | / | / |

[a] with respect to distant metastasis

[b] with respect to distant metastasis

[c] Confidence Interval

# Appendix B

# Supplementary Information - Chapter 5

## Region of Interest Creation Algorithm

The ROI creation algorithm is identical to that which was used in the previous study. The corresponding Figure from said study is also reproduced here for completeness (Figure 1). First, the PTV point cloud (a list of voxel positions contained directly on the PTV contour) was used to create a 3D volume and superimposed upon the dose grid (contained within the DICOM dose file). A convex hull encompassing the PTV, defined as the smallest convex set that contains the PTV, was generated using this point cloud. Second, the PTV point cloud was "grown" by a variable amount. The growing was facilitated by determining the centroid of the PTV and generating a direction vector between the centroid and every point. The direction vectors were scaled by a variable amount. Third, the grown point cloud was used to generate a convex hull of the isotropically grown PTV. This convex hull was superimposed with the lung contour mask, ensuring that the region of interest did not encroach on the contralateral lung and only included voxels within the ipsilateral lung. Next, XOR logic was used to create a mask which only considered points within the grown region but not within the PTV itself. This new ring-shaped region is hereafter referred to as the continuous cumulative region of interest of width x ($ROI_{cont}$(x mm)). x was varied in a range of 1 mm and 100 mm in 1 mm increments and represents the number of millimeters each direction vector was grown by. The ROI masks were then applied to the dose grid, from which all voxels contained within each ROI were analyzed.

## Supplementary Statistics

The DM event rate for the higher-than-threshold branch of the CyberKnife cohort (n=192) was 13%. The DM event rate for the lower-than-threshold branch of the CK cohort (n=13) was 0%. Analysis on this partition was not included in the main text of the manuscript due to the lack of a sufficient number of patients within the lower branch to infer any meaningful conclusions.

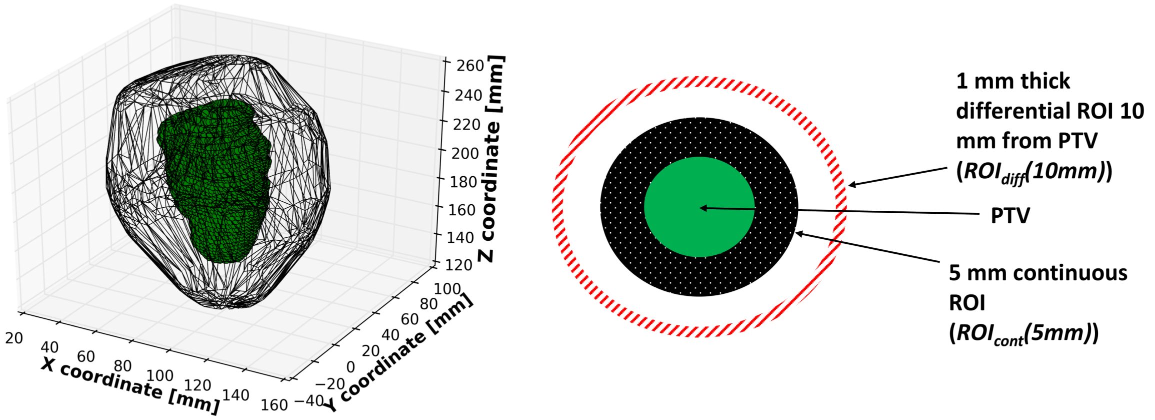


Figure B.1 **Left side: Depiction of the ROI algorithm. Inner green volume represents the PTV, black represents the boundary of the 30-mm thick shell-shaped region. Right side: Two-dimensional example of the region of interest creation algorithm.** Shown are the PTV, $ROI_{cont}$(5 mm) (the continuous region up to 5 mm outside the PTV) and $ROI_{disc}$(10 mm) (1 mm region, 10 mm away from the PTV). *Reproduced from previous study [3].*

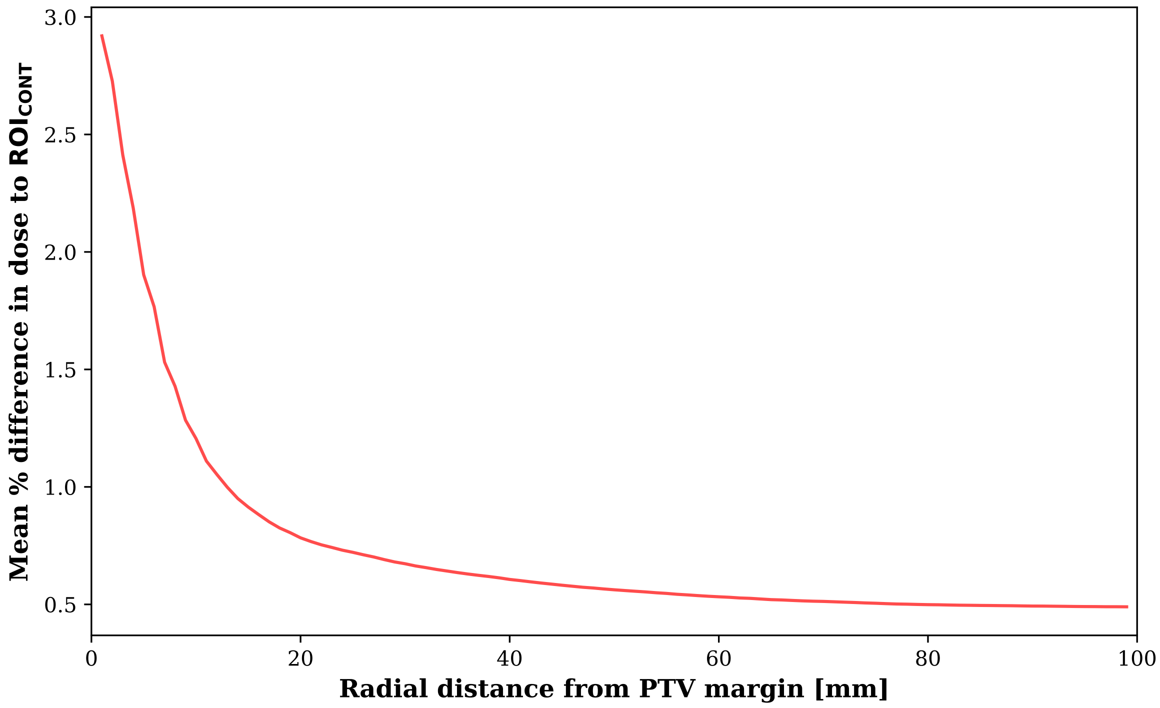
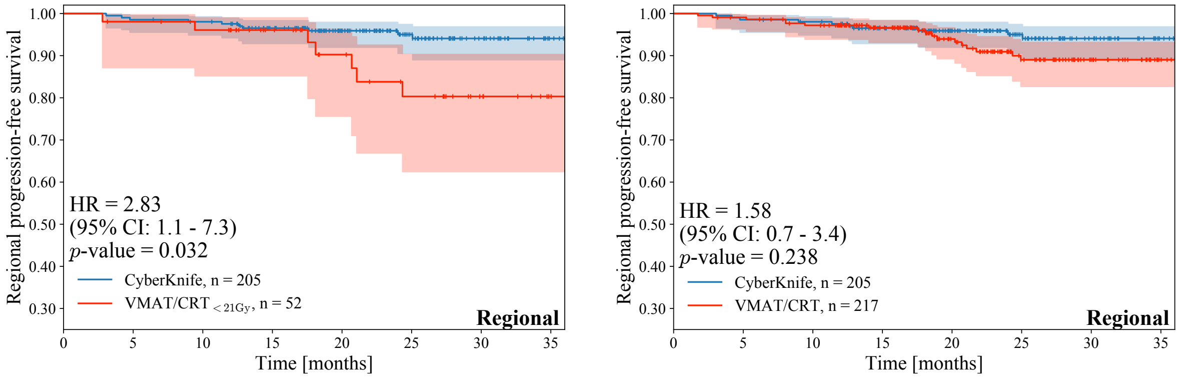


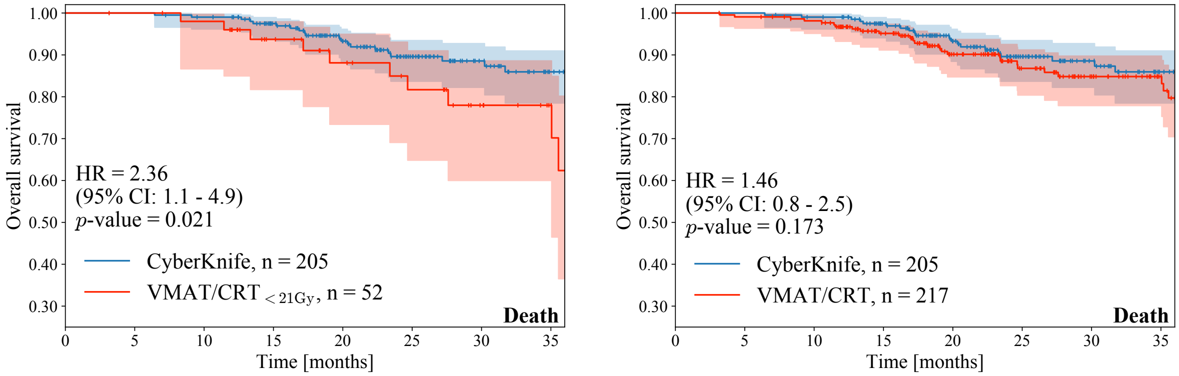
Figure B.2 **The impact of considering intrafraction motion on the dose to the** $ROI_{cont}$ **as a function of distance** $x$ **from the PTV.** This was done for a subset of 18 VMAT/CRT patients. Beyond 10 mm, the mean difference in the dose to the $ROI_{cont}$ as calculated with the non-shifted vs. the motion-averaged dose distribution is less than 2%



(a) The top curve (blue) represents the entire CyberKnife cohort while the bottom curve (red) represents the $\text{VMAT/CRT}_{<21\text{Gy}}$ cohort.

(b) The top curve (blue) represents the entire CyberKnife cohort while the bottom curve (red) represents the entire VMAT/CRT cohort.

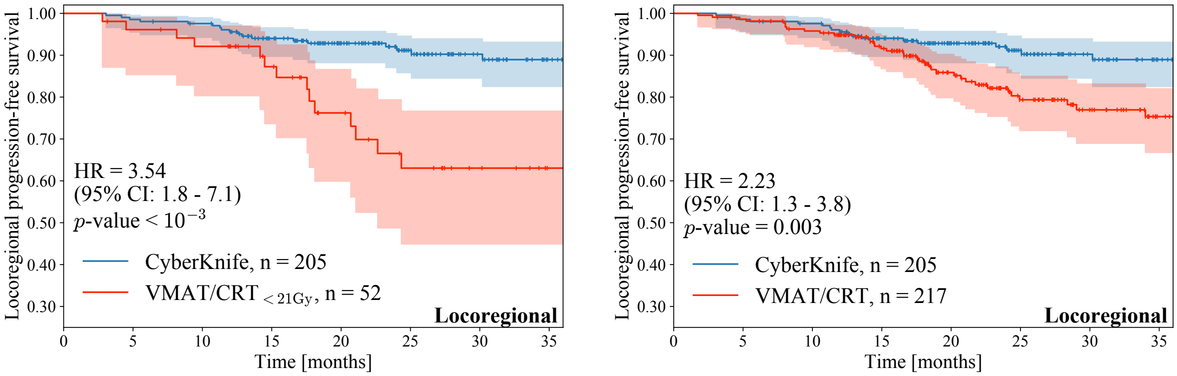
Figure B.3 **Kaplan-Meier regional progression-free survival curves.** The shaded regions correspond to the 95% confidence band of their respective survival curves. Crosses represent censored datapoints.



(a) The top curve (blue) represents the entire CyberKnife cohort while the bottom curve (red) represents the $\text{VMAT/CRT}_{<21\text{Gy}}$ cohort.

(b) The top curve (blue) represents the entire CyberKnife cohort while the bottom curve (red) represents the entire VMAT/CRT cohort.

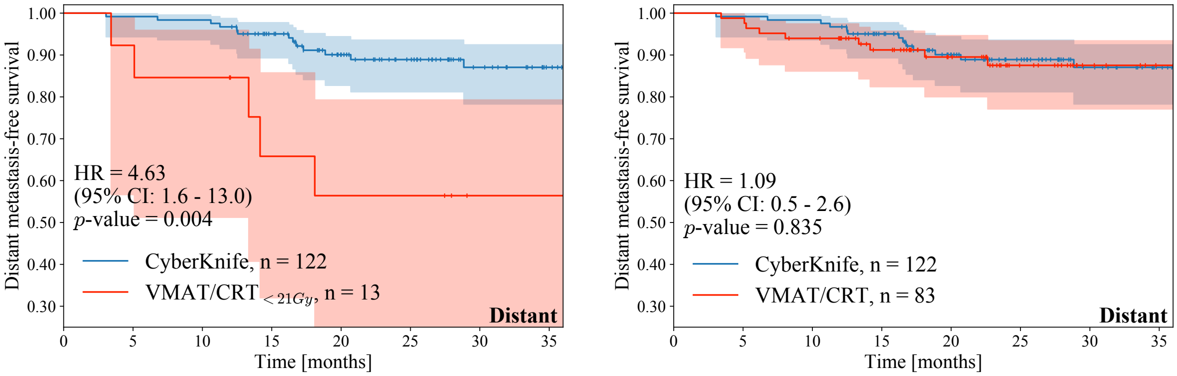
Figure B.4 **Kaplan-Meier overall survival curves.** The shaded regions correspond to the 95% confidence band of their respective survival curves. Crosses represent censored datapoints.



(a) The top curve (blue) represents the entire CyberKnife cohort while the bottom curve (red) represents the VMAT/CRT$_{<21\mathrm{Gy}}$ cohort.

(b) The top curve (blue) represents the entire CyberKnife cohort while the bottom curve (red) represents the entire VMAT/CRT cohort.

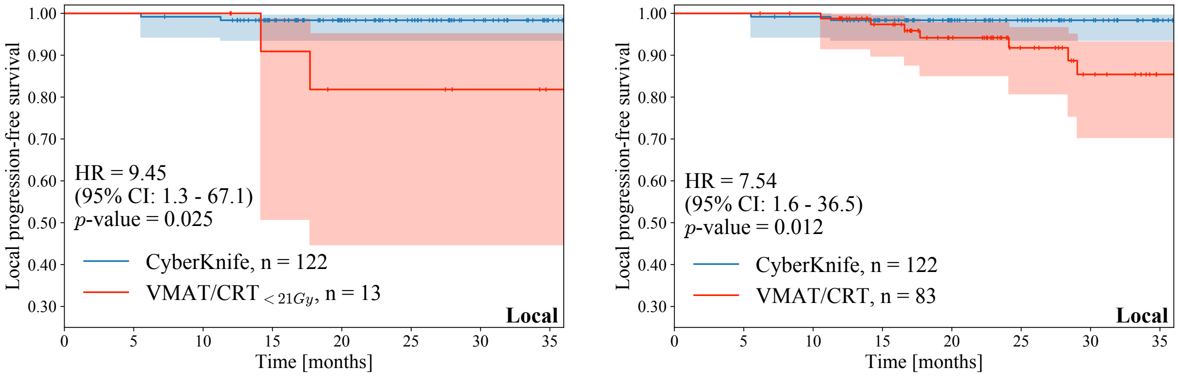
Figure B.5 **Kaplan-Meier locoregional progression-free survival curves.** The shaded regions correspond to the 95% confidence band of their respective survival curves. Crosses represent censored datapoints.



(a) The top curve (blue) represents the entire CyberKnife cohort while the bottom curve (red) represents the VMAT/CRT$_{<21\mathrm{Gy}}$ cohort.

(b) The top curve (blue) represents the entire CyberKnife cohort while the bottom curve (red) represents the entire VMAT/CRT cohort.

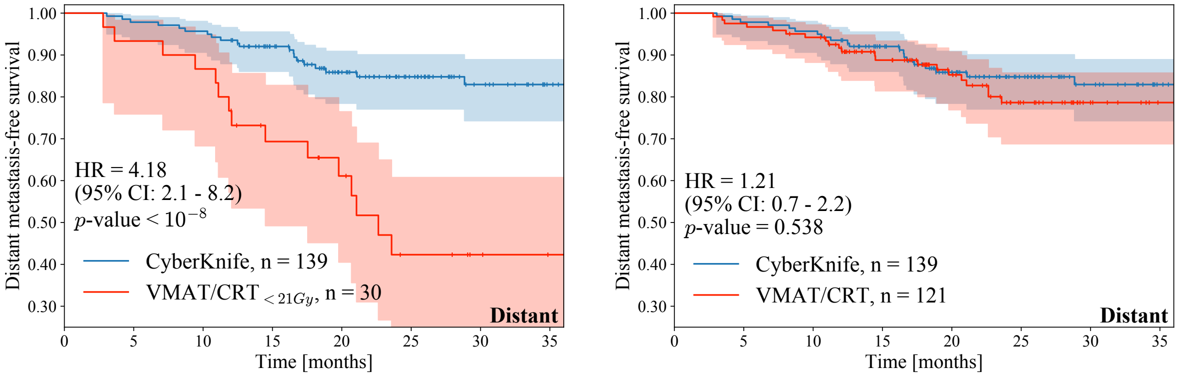
Figure B.6 **Kaplan-Meier distant metastasis free survival curves restricted to only patients with a prescription dose of 60 Gy in 3 fractions.** The shaded regions correspond to the 95% confidence band of their respective survival curves. Crosses represent censored datapoints.



(a) The top curve (blue) represents the entire CyberKnife cohort while the bottom curve (red) represents the VMAT/CRT$_{<21\text{Gy}}$ cohort.

(b) The top curve (blue) represents the entire CyberKnife cohort while the bottom curve (red) represents the entire VMAT/CRT cohort.

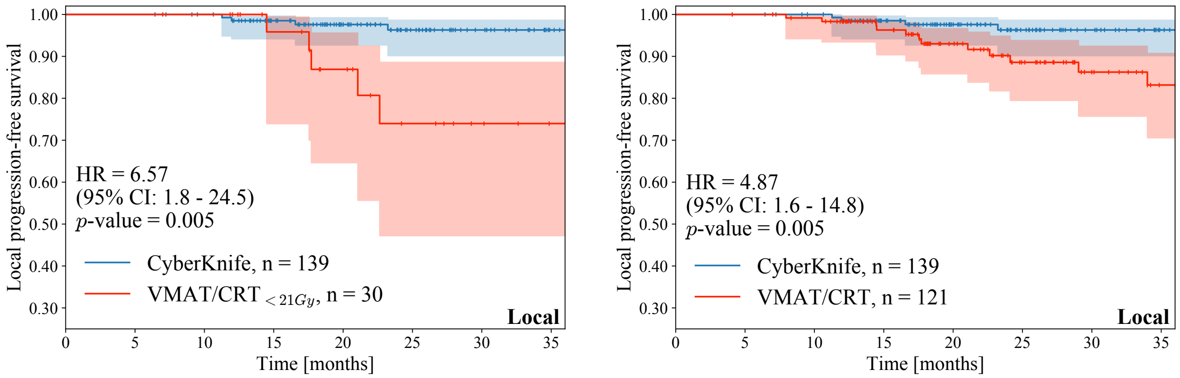
Figure B.7 **Kaplan-Meier local progression-free survival curves restricted to only patients with a prescription dose of 60 Gy in 3 fractions.** The shaded regions correspond to the 95% confidence band of their respective survival curves. Crosses represent censored datapoints.



(a) The top curve (blue) represents the entire CyberKnife cohort while the bottom curve (red) represents the VMAT/CRT$_{<21\text{Gy}}$ cohort.

(b) The top curve (blue) represents the entire CyberKnife cohort while the bottom curve (red) represents the entire VMAT/CRT cohort.

Figure B.8 **Kaplan-Meier distant metastasis free survival curves restricted to patients without adenocarcinoma.** The shaded regions correspond to the 95% confidence band of their respective survival curves. Crosses represent censored datapoints.



(a) The top curve (blue) represents the entire CyberKnife cohort while the bottom curve (red) represents the VMAT/CRT$_{<21\text{Gy}}$ cohort.

(b) The top curve (blue) represents the entire CyberKnife cohort while the bottom curve (red) represents the entire VMAT/CRT cohort.

Figure B.9 **Kaplan-Meier local progression-free survival curves restricted to patients without adenocarcinoma.** The shaded regions correspond to the 95% confidence band of their respective survival curves. Crosses represent censored datapoints.

Table B.1 **Hazard ratios obtained from multivariable Cox proportional hazards models with respect to distant metastasis and local failure.** Ranges quoted in brackets are the 95% confidence intervals. Bolded values represent statistical significance. For categorical variables, the reference category, such as the “Unknown" histology, is marked with an asterisk. Categorical covariates for which one of its categories featured no events were omitted.

| | Distant metastasis | | Local failure | |
|---|---|---|---|---|
| | Hazard Ratio | *p*-value | Hazard Ratio | *p*-value |
| Sex [female vs. male*] | 1.03 (0.61 - 1.74) | 0.92 | 0.57 (0.29 - 1.13) | 0.11 |
| Age [years] | 0.98 (0.96 - 1.01) | 0.29 | **0.96 (0.93 - 1.00)** | **0.04** |
| Stage [T2 vs. T1*] | 1.58 (0.65 - 3.83) | 0.32 | 0.97 (0.28 - 3.42) | 0.97 |
| VMAT/CRT vs. CK* | 1.23 (0.70 - 2.16) | 0.46 | **3.12 (1.42 - 6.85)** | **0.004** |
| PTV volume [cm$^3$] | 0.99 (0.97 - 1.01) | 0.32 | 1.00 (0.98 - 1.01) | 0.63 |
| Prescription EQD2 [Gy] | 0.99 (0.98 - 1.01) | 0.60 | **0.97 (0.94 - 1.00)** | **0.04** |
| Mean PTV EQD2 [Gy] | 1.00 (0.99 - 1.02) | 0.62 | 1.01 (0.99 - 1.03) | 0.29 |
| ROI(30 mm) threshold | **0.28 (0.15 - 0.55)** | $\mathbf{< 10^{-3}}$ | 0.68 (0.29 - 1.55) | 0.36 |
| **Histology** | | | | |
| Unknown* | ref | ref | ref | ref |
| NSCLC NOS | 2.19 (0.99 - 4.83) | 0.05 | 1.10 (0.33 - 3.65) | 0.88 |
| Adenocarcinoma | 0.97 (0.51 - 1.86) | 0.93 | 1.34 (0.60 - 2.98) | 0.47 |
| Squamous Cell Carcinoma | 1.04 (0.48 - 2.25) | 0.92 | 1.07 (0.34 - 3.32) | 0.91 |
| Large Cell Carcinoma | 1.50 (0.34 - 6.64) | 0.60 | – | – |

Table B.2 **Hazard ratios obtained from a multivariable Cox proportional hazards model with respect to various endpoints.** Ranges quoted in brackets are the 95% confidence intervals. Bolded values represent statistical significance. For categorical variables, the reference category, such as the “Unknown" histology, is marked with an asterisk. Categorical covariates for which one of its categories featured no events were omitted.

| | Regional failure | | Locoregional failure | | Death | |
|---|---|---|---|---|---|---|
| | Hazard Ratio | *p*-value | Hazard Ratio | *p*-value | Hazard Ratio | *p*-value |
| Sex | 1.25 (0.58 - 2.73) | 0.57 | 0.72 (0.43 - 1.23) | 0.23 | 0.95 (0.54 - 1.69) | 0.87 |
| Age [years] | 1.03 (0.98 - 1.08) | 0.23 | 0.98 (0.95 - 1.01) | 0.12 | 0.99 (0.96 - 1.02) | 0.51 |
| Stage | 1.98 (0.70 - 5.58) | 0.20 | 1.64 (0.74 - 3.60) | 0.22 | **2.94 (1.46 - 5.92)** | **0.002** |
| VMAT/CRT vs. CK* | 1.64 (0.69 - 3.85) | 0.26 | **2.20 (1.24 - 3.91)** | **0.007** | 1.68 (0.90 - 3.11) | 0.10 |
| PTV volume [cm$^3$] | 1.00 (0.99 - 1.02) | 0.69 | 1.00 (0.99 - 1.01) | 0.71 | 1.00 (0.99 - 1.01) | 0.52 |
| Prescription EQD2 [Gy] | 1.00 (0.97 - 1.03) | 0.77 | 0.98 (0.96 - 1.00) | 0.10 | 0.99 (0.97 - 1.01) | 0.23 |
| Mean PTV EQD2 [Gy] | 0.99 (0.97 - 1.01) | 0.43 | 1.00 (0.99 - 1.02) | 0.81 | 1.00 (0.98 - 1.01) | 0.85 |
| ROI(30 mm) threshold | 0.52 (0.19 - 1.45) | 0.21 | 0.68 (0.35 - 1.34) | 0.27 | 0.74 (0.34 - 1.60) | 0.44 |
| **Histology** | | | | | | |
| Unknown* | ref | ref | ref | ref | ref | ref |
| NSCLC NOS | 2.56 (0.66 - 9.84) | 0.17 | 1.62 (0.64 - 4.13) | 0.31 | 1.44 (0.56 - 3.73) | 0.45 |
| Adenocarcinoma | 1.24 (0.41 - 3.78) | 0.70 | 1.33 (0.68 - 2.60) | 0.40 | 0.99 (0.48 - 2.06) | 0.99 |
| SCC | 2.63 (0.86 - 7.97) | 0.09 | 1.95 (0.91 - 4.19) | 0.09 | 1.60 (0.72 - 3.59) | 0.25 |
| LCC | – | – | – | – | 1.42 (0.18 - 11.37) | 0.74 |

Table B.3 **Hazard ratios obtained from a multivariable Cox proportional hazards model with respect to various endpoints restricted to patients without adenocarcinoma.** Ranges quoted in brackets are the 95% confidence intervals. Bolded values represent statistical significance. For categorical variables, the reference category, such as the “Unknown" histology, is marked with an asterisk. Categorical covariates for which one of its categories featured no events were omitted.

| | Distant metastasis | | Local failure | |
|---|---|---|---|---|
| | Hazard Ratio | *p*-value | Hazard Ratio | *p*-value |
| Sex [female vs. male*] | 1.05 (0.55 - 2.03) | 0.88 | 0.76 (0.28 - 2.01) | 0.57 |
| Age [years] | 0.99 (0.96 - 1.03) | 0.78 | 0.96 (0.91 - 1.01) | 0.09 |
| Stage [T2 vs. T1*] | 1.30 (0.43 - 3.89) | 0.64 | 1.07 (0.18 - 6.24) | 0.94 |
| VMAT/CRT vs. CK* | 0.96 (0.49 - 1.89) | 0.90 | **4.50 (1.27 - 15.95)** | **0.02** |
| PTV volume [cm$^3$] | 1.00 (0.98 - 1.02) | 0.63 | 1.01 (0.99 - 1.03) | 0.27 |
| Prescription EQD2 [Gy] | 0.99 (0.97 - 1.02) | 0.60 | 0.97 (0.92 - 1.01) | 0.15 |
| Mean PTV EQD2 [Gy] | 1.00 (0.99 - 1.02) | 0.91 | 1.02 (0.99 - 1.05) | 0.15 |
| ROI(30 mm) threshold | **0.33 (0.15 - 0.77)** | **0.01** | 0.47 (0.14 - 1.62) | 0.23 |
| **Histology** | | | | |
| Unknown* | ref | ref | ref | ref |
| NSCLC NOS | 2.09 (0.94 - 4.66) | 0.07 | 1.24 (0.35 - 4.32) | 0.74 |
| Squamous Cell Carcinoma | 0.95 (0.43 - 2.07) | 0.89 | 1.03 (0.31 - 3.42) | 0.96 |
| Large Cell Carcinoma | 1.44 (0.32 - 6.43) | 0.63 | – | – |

Table B.4 **Hazard ratios obtained from a multivariable Cox proportional hazards model with respect to various endpoints restricted to only patients with a prescription dose of 60 Gy in 3 fractions.** Ranges quoted in brackets are the 95% confidence intervals. Bolded values represent statistical significance. For categorical variables, the reference category, such as the “Unknown" histology, is marked with an asterisk. Categorical covariates for which one of its categories featured no events were omitted.

| | Distant metastasis | | Local failure | |
|---|---|---|---|---|
| | Hazard Ratio | *p*-value | Hazard Ratio | *p*-value |
| Sex [female vs. male*] | 0.80 (0.34 - 1.91) | 0.61 | 0.50 (0.13 - 1.89) | 0.30 |
| Age [years] | 0.99 (0.94 - 1.04) | 0.68 | 0.95 (0.89 - 1.02) | 0.13 |
| Stage [T2 vs. T1*] | 1.98 (0.42 - 9.40) | 0.39 | – | – |
| VMAT/CRT vs. CK* | 0.93 (0.31 - 2.79) | 0.89 | **5.88 (1.18 - 29.39)** | **0.03** |
| PTV volume [cm$^3$] | 0.99 (0.95 - 1.03) | 0.58 | 0.98 (0.90 - 1.06) | 0.57 |
| Mean PTV EQD2 [Gy] | 1.00 (0.99 - 1.02) | 0.88 | 1.01 (0.98 - 1.05) | 0.50 |
| ROI(30 mm) threshold | **0.20 (0.05 - 0.71)** | **0.01** | 0.61 (0.09 - 4.16) | 0.61 |
| **Histology** | | | | |
| Unknown* | ref | ref | ref | ref |
| NSCLC NOS | 3.37 (0.75 - 15.11) | 0.11 | – | – |
| Adenocarcinoma | 1.25 (0.43 - 3.64) | 0.68 | 0.60 (0.14 - 2.46) | 0.47 |
| Squamous Cell Carcinoma | 1.48 (0.43 - 5.06) | 0.53 | – | – |
| Large Cell Carcinoma | 1.54 (0.16 - 14.64) | 0.71 | – | – |

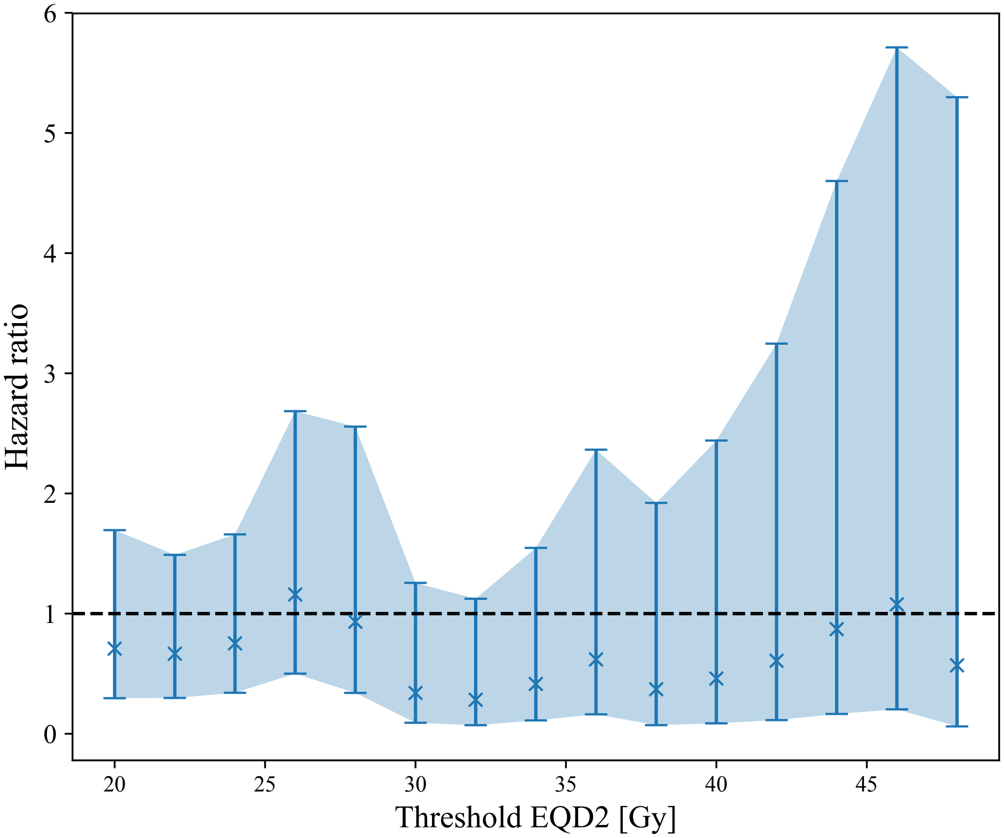


Figure B.10 **Hazard ratio with respect to local failure obtained from a multivariable Cox regression of the mean dose to the** $ROI_{cont}$**(30 mm) for differing threshold levels.** Hazard ratios in the range of 0.28 to 0.41 can be found for a threshold between 30 to 34 Gy. Shaded bands represent 95% confidence intervals.

# Appendix C

# Supplementary Information - Chapter 6

## Supplementary Note 1: Minimal activation map

The procedurally generated minimal activation map is shown in Figure C.1. Evidently, this figure represents noise. This is expected as the algorithm is trained to evaluate whether an image is likely to be an aggressive tumor. Thus, procedurally generating an image to which the network would assign a minimal score ends up resembling noise. This further reinforces that the maximal class activation map (and its properties) represent a difference between more and less aggressive tumors.

## Supplementary Note 2: Radiomic analysis

As mentioned in the manuscript, the radiomic analysis (for 94 radiomic features as described in IBSI [1]) was performed for the top 64 filters that make up the final convolutional block. The maximal activation maps for the 64 filters are shown in Figure C.2 (low-resolution version, high-resolution version requires separate download due to size). The radiomic analysis results are shown in Figure C.3.

## Supplementary Methods 1: Transfer learning approach

VGG19 was used as a transfer learning base, chosen for being the most recently successful network to win ImageNet while still maintaining similar architecture to ours (in comparison to

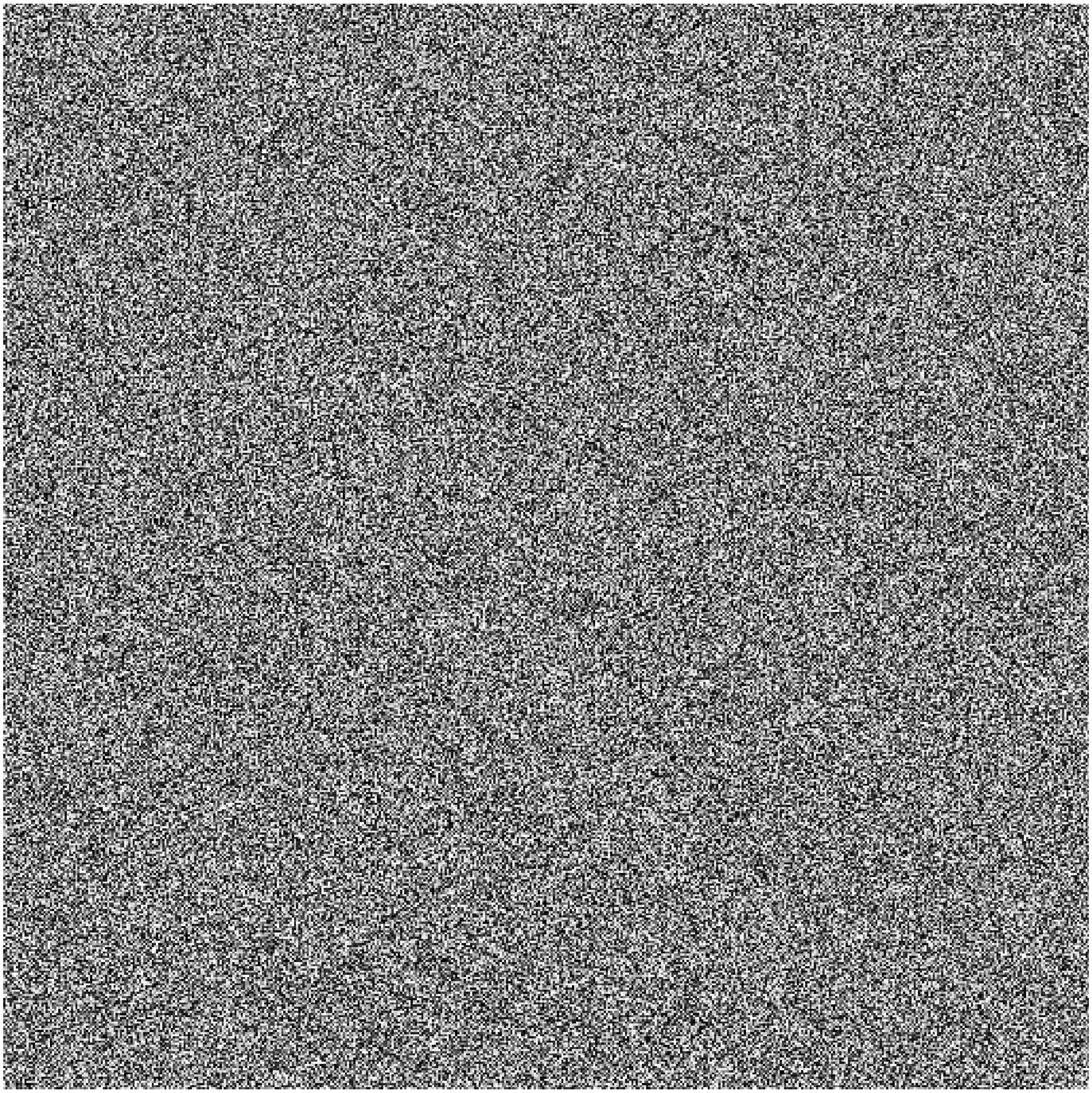


Figure C.1 **Minimal activation map**. Depicts a procedurally generated image input that would result in a minimal classification score of 0 (i.e. not distant metastasis).

the more advanced ResNet/GoogleNet). More explicitly, VGG19's final layers involve multiple fully connected layers, concluding in a fully connected layer with x nodes, where x is the number of image classes. This is in comparison to more advanced networks, which may not employ the same fully connected layer stack. Three approaches were taken. Firstly, the convolutional base (and weights) of VGG19 were used with a randomly initialized fully connected stack identical

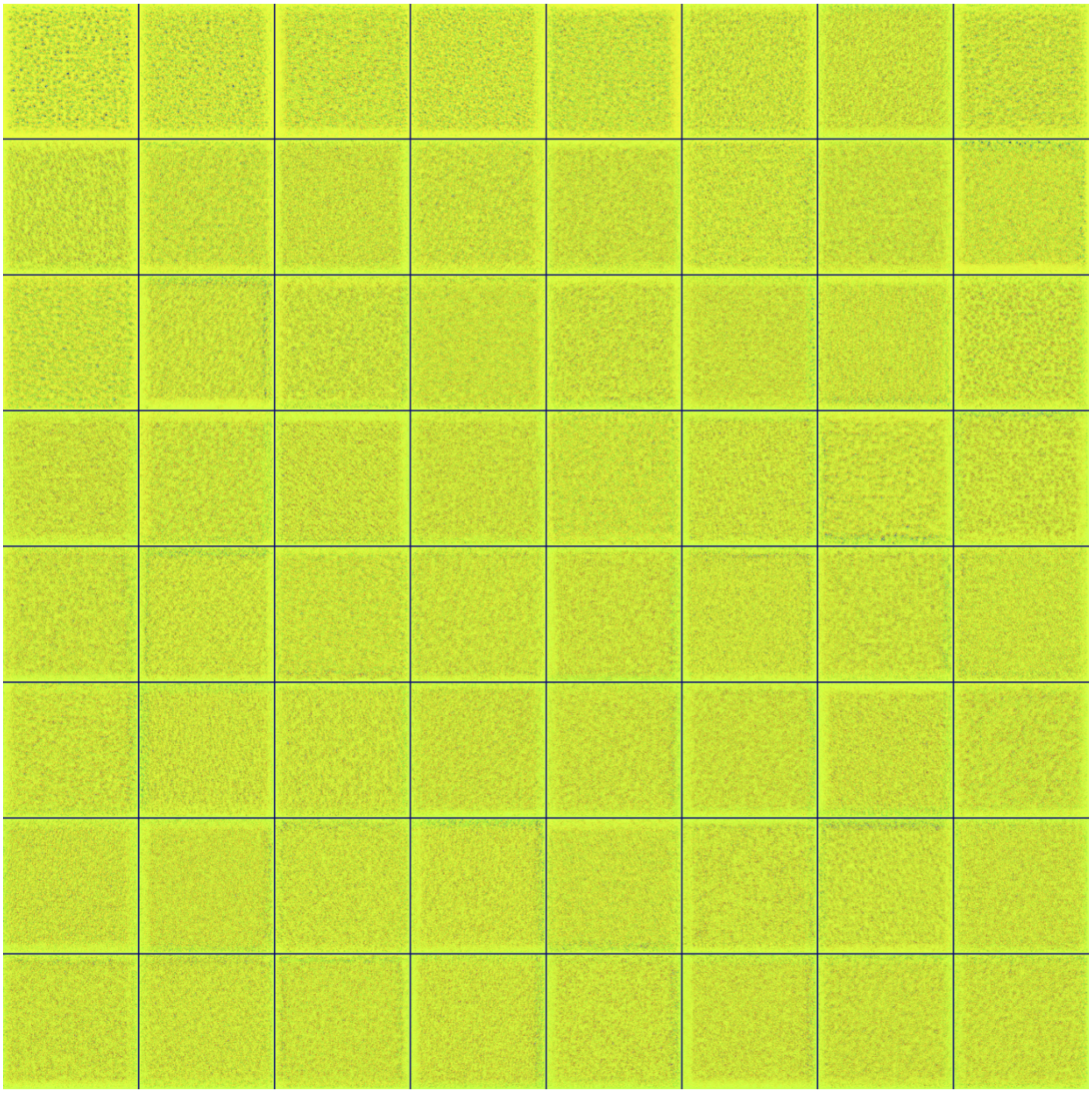


Figure C.2 **Maximal activation maps for top 64 filters in the final convolutional block.** Represents procedurally generated images that would each result in a particular filter being maximally activated. Color map chosen solely for visualization purposes.

to the bottom portion of Figure 5. This fully connected stack was subsequently fine-tuned, with a learning rate of 10-4, keeping the convolutional base weights frozen. As VGG19 requires images of shape 224 X 224, our images had to be cropped to this size. This was done while keeping the GTV itself as centered as possible in the cropped image. After 100 epochs of

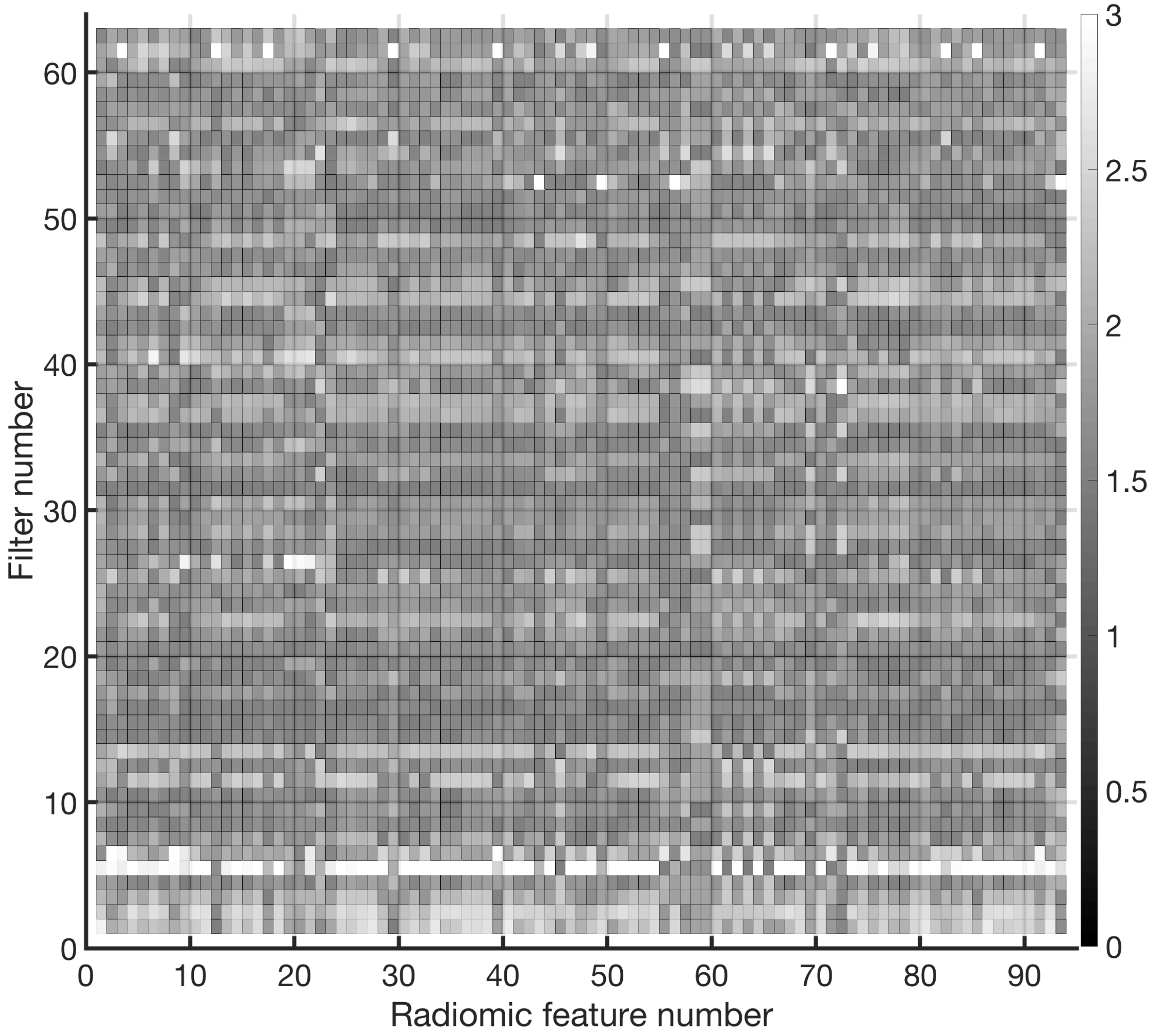


Figure C.3 **Radiomic analysis across all features and filters. x-axis represents the radiomic feature number as defined in the Image Biomarker Standardization Initiative (IBSI).** y-axis represents the filter number (arbitrary). Color indicates the normalized value of the radiomic feature (standard deviations away from the mean calculated for a specific feature across all filters). Color scale chosen solely for visualization purposes. This analysis represents that the range of filters represents a diverse selection of radiomic features, combined in a number of different ways.

training using the same partition scheme as the benchmark study, this resulted in an AUC of 0.69 predicting DM. The second method used an identical architecture, except the fully connected stack was not randomly initialized, but rather used the learnt weights from our top performing model. It was subsequently fine-tuned for 100 epochs, using a learning rate 10-4. This resulted in an AUC of 0.72. The final approach was identical to the second approach; however, we only trained the final fully connected layer. That is, only trained 4096 parameters. This resulted in an AUC of 0.76, slightly better when compared to the previous transfer learning methods utilized, however still substantially worse than our de novo methodology. More work is needed to investigate the full implications of using transfer learning in this context; for our study we further concentrated on the performance of our own CNN constructed for this purpose.

## Supplementary Methods 2: Parameters tested when building the CNNs

A variety of architectures and parameters were tested during the course of this research. A summary of what values were tested is shown in Table C.1. A more detailed explanation of each parameter and the range tested is described below. We note that the exact choice of architecture and parameters is a result of an educated trial-and-error process, and the intuition gleaned from it. There is a decent degree of robustness in the variation of many parameters (e.g. size of fully connected layers, number of filters, type of non-linearity), as slight modifications will not qualitatively alter the results.

Table C.1 **Summary of parameters tested during the course of this research.** Bolded value represents the value(s) settled on in the final model.

| Parameter tested | Values tested |
|---|---|
| **Batch size** | 16, **32**, 48 |
| **Number of filters** | **32, 64, 128**, 256, 512 |
| **Number of convolutional blocks** | 2, **3**, 4 |
| **Type of non-linearity** | **PReLU**, ReLU |
| **Filter size** | **3x3, 5x5**, 7x7, 9x9, 12x12 |
| **Type of pooling** | **Max pooling**, Average pooling |
| **Size of pooling** | 2x2, **4x4**, 6x6 |
| **Number of fully connected layers** | 1, **2**, 3 |
| **Size of fully connected layers** | 64, **128, 256**, 512 |
| **Learning rate** | $10^{-2}$, $\mathbf{10^{-3}}$, $10^{-4}$, $10^{-5}$ |
| **Momentum** | 0.25, **0.5**, 0.75, 0.9 |
| **Rotation range (data augmentation)** | **20**°, 30°, 40° |
| **Shift range (data augmentation)** | 0.2, 0.3, **0.4** |

*Batch size*: The batch size represents the number of images that the network sees each iteration of the algorithm. One epoch consists of several batches such that the network sees every image once (i.e. # of batches/epoch = total # of images / batch size). The performance seemingly began to plateau at the chosen batch size (32), higher batch sizes did not improve the performance. Furthermore, at higher batch sizes, computational memory becomes a serious concern, limiting our possible range of batch sizes.

*Number of filters*: The number of filters represents the number of learned weight matrices within each convolutional layer. A variety of permutations was tested, always with deeper layers including either an equal number or more (often double) filters than the previous layer. The range (from 25 to 29) is a typical range tested when building a CNN framework. Our network leans towards using a smaller number of filters when possible, to avoid over-fitting.

*Number of convolutional blocks*: The number of convolutional blocks drastically impacts the complexity of the network. CNNs are prone to over-fitting, especially with a larger number of convolutional blocks, limiting our range to that shown above. In particular, the CNN began to overfit when using 4 convolutional blocks as opposed to 3.

*Type of non-linearity*: Parametric rectified linear units (PReLU) introduce a minor amount of complexity into the network, which in our case resulted in a performance increase of approximately 2%.

*Filter size*: Filter size represents the effective impact that a single pixel has on deeper layers. A particularly large range was chosen as a priori the relevant neighbourhood around any given pixel is unknown. As most tumors within our dataset were relatively small, the larger filter sizes (i.e. 9 x 9 or higher) were quickly dismissed.

*Type of pooling*: Average pooling was attempted, but quickly dismissed for generalization purposes. In our case, max pooling results in a network with significantly less over-fitting. This is as expected, as each max-pooling layer effectively reduces the amount of information (which possibly could be used to over-fit) by nearly 95%. In comparison, average-pooling reduces the dimensionality, but still allows much of the information to flow through to the next layer.

*Size of pooling*: Larger max pooling results in less over-fitting, but 6x6 resulted in the loss of a substantial amount of information (particularly with the filter size used).

*Number of fully connected layers*: The number of fully connected layers significantly impacts the complexity of the network. It also exponentially increases the computational power required, restricting the maximum range. Like the number of convolutional blocks, we found the network began to overfit when using more than 2 fully connected layers. On the other hand, using 2 fully connected layers instead of 1 resulted in a significant performance increase (on the order of 10%).

*Size of fully connected layers*: Similarly, the size drastically impacts the complexity and the computational power required. It is typical to test up to a size of $2^{10}$, however we did not due to computational reasons.

*Learning rate*: Learning rate affects the step size that the algorithm takes during the gradient descent algorithm. A learning rate that is too high can skip over a local/global minimum, while a learning rate that is too low can get caught in a local minimum. We also experimented with lowering the learning rate once the algorithm was close to a global minimum, however it did not result in a performance gain.

*Momentum*: Similar to varying the learning rate, momentum can help the algorithm not get caught in a local minimum. It does this by not taking a step solely based on the gradient, but also incorporates the current velocity (akin to classical momentum). A wide range was tested as a priori it is difficult to know what the proper value for our particular network is.

*Rotation range (data augmentation)*: The rotation range was chosen to enable most possible arrangements of a particular tumor (as each image was also flipped horizontally and/or vertically).

*Shift range (data augmentation)*: The shift range was chosen as to not create a situation where a tumor could be shifted entirely out of frame (<1% chance even at the maximum shift value.

*Number of epochs*: To ensure that each network had enough training time, they were trained for 400 epochs and the training/validation loss was visually inspected. An example of the training (blue)/validation (orange) loss progress for is shown in Supplementary Figure C.4. At approximately 100 epochs, the network began to over-fit the training data (as shown by the steeper slope on the training curve and drastic increase in the validation loss. Ultimately, the model reached its validation loss minimum at the $34^{th}$ epoch.

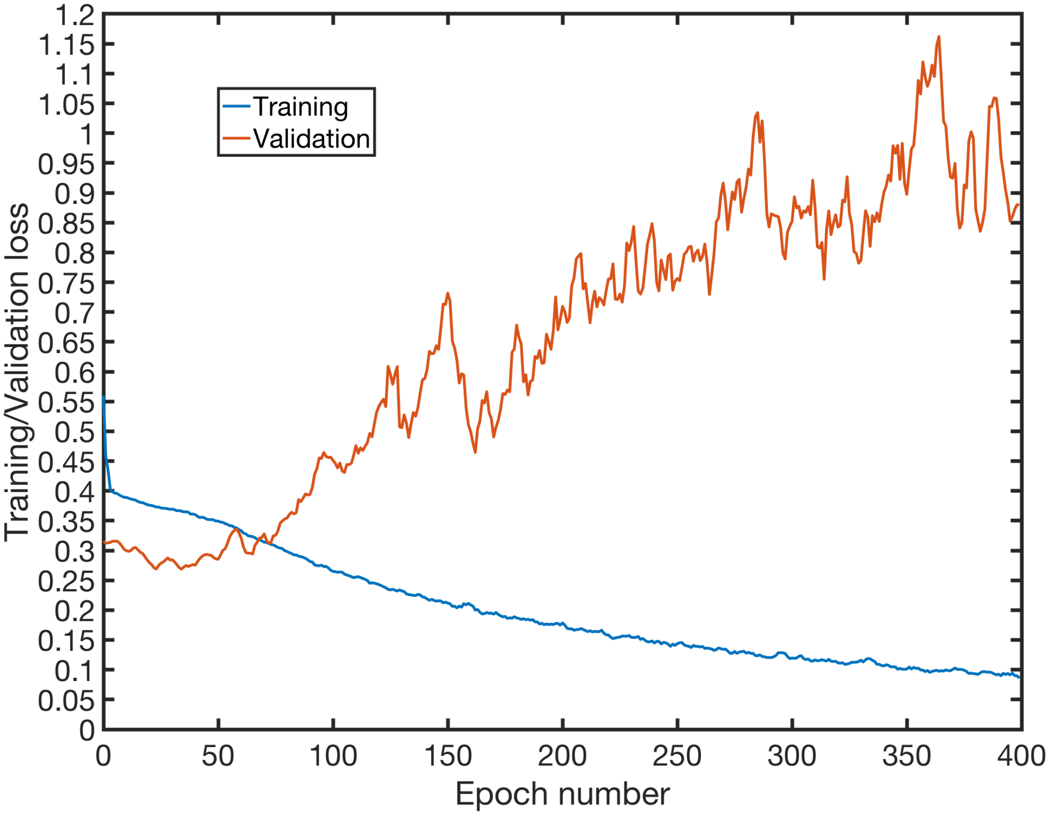


Figure C.4 **Training/Validation loss for a sample training of a network**. Orange represents the loss measured on the validation data-set, blue represents the loss measured on the training data-set.

# Appendix D

# Supplementary Information - Chapter 7

## Supplementary Note 1: Numerical results on artificial cohorts (for clinical interpretation)

When applying the model to the actual cohort, 33/106 high-risk patients (w.r.t. overall survival) are predicted (31%). In the first artificial set (T4 cohort), the model predicts 50/106 high-risk patients (47%), to be expected as it is known that T4 tumors have a worse prognosis. Similarly, in the second artificial set (T0 cohort), the model predicts 22/106 high-risk patients (21%). Further results are shown in Table D1.

Table D.1 **Numerical results when applying overall survival prediction model to artificial cohorts.**

| | Predicted high-risk patients | True Negative Rate (TNR) | True Positive Rate (TPR) |
|---|---|---|---|
| **Actual Cohort** | 33/106 (31%) | 68/82 (83%) | 19/24 (79%) |
| **All T4** | 50/106 (47%) | 52/82 (63%) | 19/24 (83%) |
| **All T0** | 22/106 (21%) | 73/82 (89%) | 19/24 (54%) |
| **All N3** | 36/106 (34%) | 65/82 (79%) | 19/24 (79%) |
| **All N0** | 27/106 (25%) | 71/82 (87%) | 19/24 (67%) |
| **All hypopharyngeal** | 101/106 (96%) | 5/82 (6%) | 24/24 (100%) |

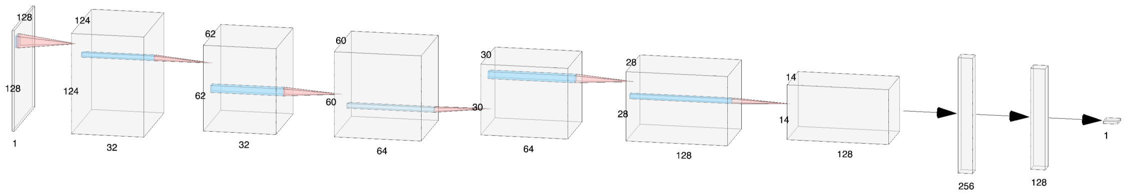


Figure D.1 **PET portion trained in Step 1 (red).** Numbers are indicative of the dimensions of each layer ("feature map"). Each layer alternates between a convolutional kernel (of size 5x5, 3x3, 3x3) and a max-pooling layer (2x2, 2x2, 2x2). Not shown is a dropout layer (0.5) after the final fully connected layer.

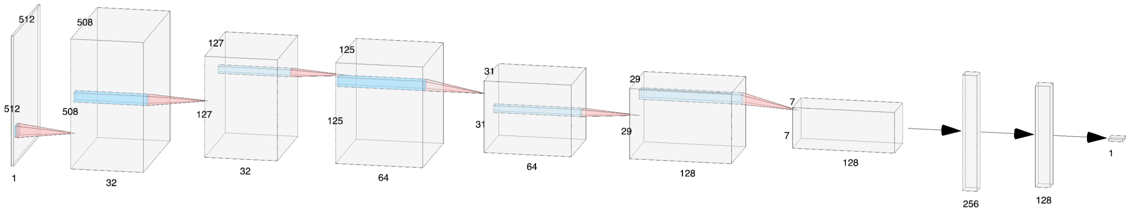


Figure D.2 **CT portion trained in Step 1 (blue).** Numbers are indicative of the dimensions of each layer ("feature map"). Each layer alternates between a convolutional kernel (of size 5x5, 3x3, 3x3) and a max-pooling layer (4x4, 4x4, 4x4). Not shown is a dropout layer (0.5) after the final fully connected layer.

# Supplementary Methods: Parameters tested when building the CNNs

A variety of architectures and parameters were tested during the course of this research. A summary of what values were tested is shown in Table D2. A more detailed explanation of each parameter and the range tested is described below. We note that the exact choice of architecture and parameters is a result of an educated trial-and-error process, and the intuition gleaned from it. Furthermore, informed by our previous work, we were able to perform a more efficient parameter search, ultimately eliminating the need to test the extreme ranges. There is a decent degree of robustness in the variation of many parameters (e.g. size of fully connected layers, number of filters, type of non-linearity), as slight modifications will not qualitatively alter the results.

Table D.2 **Summary of parameters tested during the course of this research.** Bolded value represents the value(s) settled on in the final model.

| Parameter tested | Values tested |
|---|---|
| **Batch size** | 16, **32**, 48 |
| **Number of filters** | **32, 64, 128** |
| **Number of convolutional blocks** | **3** |
| **Type of non-linearity** | **PReLU**, ReLU |
| **Filter size** | **3x3, 5x5**, 7x7 |
| **Type of pooling** | **Max pooling** |
| **Size of pooling** | 2x2, **4x4**, 6x6 |
| **Number of fully connected layers** | 1, **2**, 3 |
| **Size of fully connected layers** | 64, **128, 256**, 512 |
| **Initial Learning rate** | $\mathbf{10^{-3}}$, $10^{-4}$ |
| **Rotation range (data augmentation)** | **20°** |
| **Shift range (data augmentation)** | **0.2** |

*Batch size:* The batch size represents the number of images that the network sees each iteration of the algorithm. One epoch consists of several batches such that the network sees every image once (i.e. # of batches/epoch = total # of images / batch size). The performance seemingly began to plateau at the chosen batch size (32), higher batch sizes did not improve the performance. Furthermore, at higher batch sizes, computational memory becomes a serious concern, limiting our possible range of batch sizes.

*Number of filters:* The number of filters represents the number of learned weight matrices within each convolutional layer. A variety of permutations was tested, always with deeper layers including either an equal number or more (often double) filters than the previous layer. The range (from 25 to 27) was informed by our previous work.

*Number of convolutional blocks:* The number of convolutional blocks drastically impacts the complexity of the network. CNNs are prone to over-fitting, especially with a larger number of convolutional blocks; 3 were used as informed by our previous study.

*Type of non-linearity:* Parametric rectified linear units (PReLU) introduce a minor amount of complexity into the network, which in our case resulted in a performance increase of approximately 2%.

*Filter size*: Filter size represents the effective impact that a single pixel has on deeper layers. Informed by our previous study, we were able to reduce the range of filter sizes tested. This is further reinforced due to the size of each tumor.

*Type of pooling:* Max-pooling was used as informed by our previous study.

*Size of pooling:* Larger max pooling results in less over-fitting, but 6x6 resulted in the loss of a substantial amount of information (particularly with the filter size used).

*Number of fully connected layers:* The number of fully connected layers significantly impacts the complexity of the network. It also exponentially increases the computational power required, restricting the maximum range. Like the number of convolutional blocks, we found the network began to overfit when using more than 2 fully connected layers. On the other hand, using 2 fully connected layers instead of 1 resulted in a significant performance increase (on the order of 10%).

*Size of fully connected layers:* Similarly, the size drastically impacts the complexity and the computational power required. It is typical to test up to a size of 210, however we did not due to computational reasons.

*Initial learning rate:* Learning rate affects the step size that the algorithm takes during the gradient descent algorithm. A learning rate that is too high can skip over a local/global minimum, while a learning rate that is too low can get caught in a local minimum. Similar to the previous study, we also experimented with lowering the learning rate once the algorithm was close to a global minimum, however it did not result in a performance gain.

*Rotation range (data augmentation):* The rotation range was chosen to enable most possible arrangements of a particular tumor (as each image was also flipped horizontally and/or vertically).

*Shift range (data augmentation):* The shift range was chosen as to not create a situation where a tumor could be shifted entirely out of frame (<1% chance even at the maximum shift value).