2021-09-24

Analysis of Radiation Therapy in Cancer Treatment using Machine Learning

Yarschenko, Adam H

UNIVERSITY OF CALGARY

Analysis of Radiation Therapy in Cancer Treatment using Machine Learning

by

Adam Howard Yarschenko

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF SCIENCE

GRADUATE PROGRAM IN MECHANICAL ENGINEERING

CALGARY, ALBERTA

SEPTEMBER, 2021

ABSTRACT

Advancements in machine learning and data science have allowed researchers and clinicians to generate key insights from the vast amount of data generated in healthcare. This is currently a topic of research with great interest. With the advancement in algorithm design, and computing power, machine learning has proven to be a capable tool to augment or partially automate decision making. In this thesis, patient reported outcome surveys (PROs) for head and neck radiotherapy, and the relationship between the radiation dose distribution and breast size for whole-breast radiotherapy were investigated using statistical and machine learning methods.

Two PRO measures; the M.D. Anderson symptom inventory for head and neck cancer, and the M.D. Anderson dysphagia inventory, were examined for a cohort of patients post radiotherapy for head and neck cancer. A strategy for administering a single PRO instrument is proposed which would reduce the questionnaire burden on patients, and allow physicians to identify patients who require specialized treatment for dealing with radiotherapy side-effects, such as referral to a dietician or speech language pathologist.

Dosiomic features were extracted from the 3D radiation dose cloud in whole-breast radiotherapy plans. Feature reduction was achieved through hierarchical clustering and random forests were trained to stratify treated volume based on the distribution in the dose. Permutation feature importance was used to rank features' classification utility in this task. The top 3 features were used to achieve superior performance when compared to the entire feature set. Dosiomics gives new insight into 3D dose distribution, and these features can be used in future studies to relate to treatment outcomes associated with whole breast radiotherapy for large volumes.

PREFACE

This thesis is original work by the author Adam Yarschenko. The data used in the study in Chapter 2 was covered by ethics certificate HREBA.CC-19-0119, issued by the Health Research Ethics Board of Alberta. The study in chapter 3 was covered by ethics certificate HREBA.CC-19-0333 issued by the Health Research Ethics Board of Alberta.

ACKNOWLEDGEMENTS

Firstly, I would like to thank my supervisors Dr. Qiao Sun, and Dr. Wendy Smith for taking me on as a graduate student, for their sharing their knowledge, careful guidance, and support during my graduate studies. I was very lucky to have such capable and prominent supervisors to help me with this work. I would also like to acknowledge members of my thesis examination committee, Dr. Charles Kirkby, and Dr. Deyi Xue for attending the exam, and their feedback on the final thesis.

I would like to thank my co-authors for their hard work and contributions to the studies: Dr. Lisa Barbera, Dr. Harvey Quon, Demetra Yannitsos, Dr. Sarah Weppler. I would also like to thank my fellow student Lingyue Sun for all her help and guidance.

I would like to thank my loving wife Midori for her help and support during this period. Her support made it possible to focus on my research, and I couldn't have done it without her.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES AND ILLUSTRATIONS

LIST OF SYMBOLS, ABBREVIATIONS, AND NOMENCLATURE

Symbol	Definition
3DCRT	3-Dimensional Conformal Radiation Therapy
AI	Artificial Intelligence
ALARA	As Low As Reasonably Achievable
AUC	Area Under the Curve
BMI	Body Mass Index
CBCT	Cone Beam Computed Tomography
CNN	Convolutional Neural Network
CPQR	Canadian Partnership for Quality Radiotherapy
CT	Computed Tomography
CTV	Clinical Target Volume
D_2	Minimum radiation dose to the hottest 2% of the volume
D_{2CC}	Minimum radiation dose to the hottest 2 cubic centimeters of the volume
D_{98}	Minimum radiation dose delivered to the hottest 98% of the volume
DICOM	Digital Imaging and Communications in Medicine
DVH	Dose Volume Histogram
EBCTCG	Early Breast Cancer Trialists' Collaborative Group
ER	Emergency Room
FP	False Positive
FN	False Negative
GLCM	Grey Level Cooccurrence Matrix
GLDM	Grey Level Dependence Matrix
GLN	Grey Level Nonuniformity
GLRLM	Grey Level Run Length Matrix
GLSZM	Grey Level Size Zone Matrix
GTV	Gross Tumor Volume
Gy	Radiation Units of Gray
H&N	Head and Neck
HNC	Head and Neck Cancer
HPV	Human Papilloma Virus
HREBA	Health Research Ethics Board of Alberta
IBSI	Image Biomarker Standardisation Initiative
ICRU	International Commission on Radiation Units and measurements
IGRT	Image Guided Radiation Therapy
IMC	Internal Mammary Chain
IMRT	Intensity Modulated Radiation Therapy
MDADI	M.D. Anderson Dysphagia Inventory
MDASI-HN	M.D. Anderson Symptom Inventory Head and Neck Module
ML	Machine Learning

MLC	Multi-Leaf Collimator
MRI	Magnetic Resonance Imaging
MV	Mega Voltage
NN	Neural Network
OAR	Organs At Risk
PRO	Patient Reported Outcomes
PROM	Patient Reported Outcome Measure
PTV	Planning Target Volume
RF	Radiofrequency
ROC	Receiver Operating Characteristic Curve
ROI	Region of Interest
RT	Radiation Therapy
SD	Standard Deviation
SML	Small, Medium, Large
TN	True Negative
TP	True Positive
V105	Volume to receive 105% of the prescribed radiation dose
V107	Volume to receive 107% of the prescribed radiation dose
VMAT	Volumetric Modulated Arc Therapy

CHAPTER 1

INTRODUCTION

1.1 Cancer and Cancer Treatments

There is an abundance of data being generated in cancer clinics today. From treatment planning and delivery, to patient outcomes. These data contain a wealth of information previously undiscoverable by traditional analysis methods. With the advancement of data science and machine learning (ML), data has become an important tool in informing treatment decisions, and assessing the effectiveness of medical interventions. These tools utilize the power of computation to churn through this vast amount of data and discover patterns that were traditionally hidden, and that humans by themselves are unable to understand. By pairing these techniques with clinical expertise, advancements in patient care have taken a leap forward. This thesis examines the use of these technologies in patient reported outcome measures for head and neck cancer patients, and extracting information from the 3D radiation dose cloud in whole breast radiotherapy cancer treatments.

Cancer is the number 1 cause of death in Canada and 1 in 2 Canadians will develop it within their lifetime [1]. In 2020, an estimated 225,800 Canadians developed cancer and 83,300 died [2] and breast cancer was the second most diagnosed cancer overall, following lung [3][2]. Progress has been made with cancer fatalities steadily declining since peaking in 1988 due to improved treatment techniques and screening; however, more work needs to be done. This thesis examines some applications of ML in external beam radiotherapy treatment of head and neck (H&N), and breast cancer.

Cancer is characterized by unregulated cell growth, often resulting in tumors. Diseased cells reproduce, giving rise to new cancerous cells [4]. The proliferation of these cells is the fundamental characteristic of the disease-causing aspect of cancer. Cancerous cells invade healthy tissues and proliferate, pushing away the healthy cells in the region and destroying adjacent tissues, causing death if left untreated. The cells can also metastasize, and send colonizing cells to other regions of the body through the blood, lymphatic fluid, or cerebrospinal fluid [4][5].

Cancer treatments have developed greatly over the years. The main methods of treatment generally fall under three categories: Surgery, systemic therapy such as chemotherapy and immunotherapy, and radiation therapy [6]. Surgical methods involve physically removing the cancerous tissue, or tissue the disease is likely to propagate to in the case of preventative surgery. Systemic methods such as chemotherapy treat the entire body of the patient. Chemotherapy agents are injected into the blood stream and circulate throughout the patient in an attempt to attack the disease at the cellular level [6]. Radiation therapy is a technique in which high energy radiation is directed at a region of the body containing the cancerous cells, causing DNA damage to the tissue resulting in cell death [7]. In radiotherapy, the dose is planned to be maximal to the diseased area, while avoiding healthy tissues. Another form of radiotherapy known as brachytherapy, involves implanting radioactive seeds inside of the tumor or seroma region [8]. The seeds irradiate the tumor from the inside and spare the surrounding healthy tissue. Often these methods are used in combination such as surgery followed by chemotherapy and radiation. This thesis focuses primarily on breast, and H&N external beam radiotherapy, which will be discussed in detail;

however, many of the patients included in the studies in this thesis could have also undergone other treatments.

Overall and disease-specific survival rates are usually reported as 5-year survival, with 10, 15, and 20 year also being cited [2]. Depending on the stage of the disease and the age of the patient, cancer treatments can have different goals: cure, palliation, control, and prevention [6]. Cancer treatments with curative intent aim to increase the 5-year probability of survival. With some treatments being highly effective, many patients can survive 15 years or more. With some patients living this much longer, the long-term side-effects and impacts on quality of life are becoming more and more important [9]. Palliative treatments aim to reduce the discomfort of the patient until they pass away, or to extend life. For palliative treatments, the emphasis is on quality of life, while long-term side-effects are considered less important [10]. Control aims to cease the spread of the disease to other regions of the body, with the understanding that the main tumor or disease may be in an area too sensitive to treat, or is resistant to treatment. Preventative measures are taken when patients have genetic predisposition to certain cancers, such as breast, and the area is surgically removed, or treated in such a way as to prevent the occurrence, or recurrence of the disease [6].

1.2 Radiation Therapy

Shortly after the discovery of x-rays in 1895 by Wilhelm Röntgen [11], Leopold Freund and Eduard Schiff suggested they be used in treating diseases [12], and it was noticed early on that the clinical ability of x-rays to destroy cancerous cells could be used in therapy. In 1896 Victor Despeignes published on the treatment of a stomach tumor of a 32-year-old man, with

“considerable improvement in the condition of the patient” [13]. As early as 1903-1904 radium 226 had begun to be used to treat skin and uterine cancers [14].

While the early production of x-rays was done using cathode ray tubes, high energy clinical radiation was supplied using Cobalt 60. Prior to 1951, most radiation therapy treatment units were only capable of producing kilovoltage radiation (know as orthovoltage), which have limited ability to penetrate into tissue [15]. These types of units are still used in some centers to treat skin cancers, an example orthovoltage unit is shown in Figure 1-1. The first Cobalt 60 radiation therapy machine was commissioned in 1951 in the Saskatoon Cancer Clinic, and the first patient was treated on November 8th that year [16]. Cobalt 60 machines were able to produce megavolt energy photons, which are capable of penetrating deeper into the tissue layers than orthovoltage units. Cobalt-60 machines were beneficial because of their reliable beam output and a well-defined half-life. A disadvantage is the need to replace the source roughly every four to five years [15]. It also has lower beam penetration, delivers higher skin dose, and fewer degrees of freedom than modern high energy megavolt units [15].

Figure 1-1: Orthovoltage radiotherapy unit. They deliver lower energy radiation and are typically used to treat skin cancers. An appropriately sized cone is selected which constrains the shape of the x-ray beam, and is directed overtop the area to be treated.

Modern radiotherapy techniques now involve the use of linear accelerators that produce the high energy ionizing radiation required, with maximum energy ranges from 4-15 MeV [15]. A representation of a medical linear accelerator is shown in Figure 1-2.

Modern linear accelerators (linacs) have similar designs with some variations across manufacturers. Radiofrequency (RF) waves are directed into a klystron which amplifies them and sends them into a waveguide. Electrons are produced from an electron source, typically a tungsten filament that is heated and the electrons are ejected and fed into the waveguide. Once in the wave guide, packets of electrons are accelerated nearly to the speed of light and the beam is directed and shaped by magnets. In some designs the beam is bent 270° to provide more focus at the electron target. The target is typically made of tungsten, and when the electrons strike it, x-ray photons are produced in a variety of directions. Only forward moving photons are allowed to travel through

the primary collimator. The photons travel through the collimator and head into a flattening filter. The flattening filter is a cone shaped target that increases the uniformity of the beam intensity before it exits the head; however, some flattening filter free beams are in clinical use [17]. Further beam shaping is done by the multi-leaf collimator (MLC); which has an array of tungsten leaves arranged such that they can be positioned to block sections of the beam, conforming the beam geometry to the shape of the tumor [15].

Figure 1-2: Medical Linear Accelerator. Patients lay on the treatment couch which is in the foreground of the image. The gantry of the linear accelerator rotates around a central axis known as the isocenter and the treatment couch can have up to six degrees of freedom movement.

Radiation interacts with the molecules of the tissue through complex relationships that are a function of many parameters such as: the type of radiation, beam energy, depth of penetration, and type of tissue [15]. Higher energy photon beams penetrate more deeply into tissue, but they also travel further before delivering maximum dose. An example of a depth-dose curve is depicted in

Figure 1-3 for a typical high-energy photon beam like what would be used clinically. The photon beam enters the skin delivering a low dose to the surface D_s, as the beam penetrates further into the tissue the rate of delivered dose builds up until it reaches its maximum value at a depth inside of the tissue d_{max}. Between the surface and d_{max} is the build-up region. After d_{max}, the rate of energy absorption decreases almost exponentially with increasing depth due to attenuation of the beam in the tissue [15]. In order to deliver the prescription dose to the entire target volume when the depth of the tissue is much larger than the buildup region, the number of monitor units, or the time that the beam is left on, is increased, and more beams are added. This can result in hot spots in regions of the tissue [18] which are areas of high dose exceeding the prescribed dose.

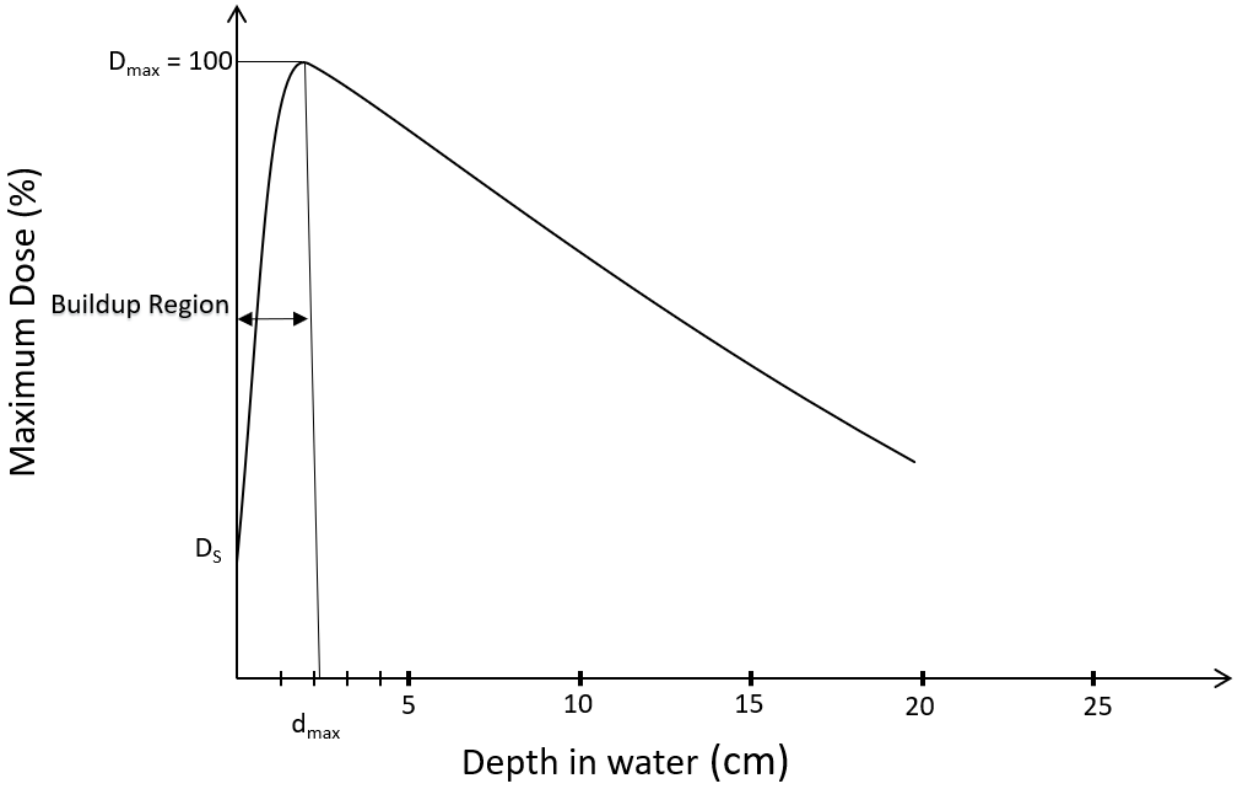

Figure 1-3: Depth-Dose Curve for a 6 MV photon beam in water. Water is used as a tissue surrogate since humans contain about 70% water. As the beam enters, a small amount of dose (Ds) is deposited at the surface of the skin. The buildup region occurs until the maximum percent dose is reached at a depth of d_{max}. This is where the beam deposits the highest amount of energy in the tissue, dose deposition falls off almost exponentially as the depth increases.

1.2.1 Treatment Planning and Delivery

Breast and H&N treatment planning share similar processes. Once a patient has been diagnosed with cancer, the treatment planning process begins. Through discussions with their doctor, a decision is made whether to pursue one, or a combination of treatment methods. If radiotherapy is chosen as one of the treatment methods, diagnostic imaging is done through a computed tomography (CT) scan, or in some cases magnetic resonance imaging (MRI) is conducted. The initial CT scan is known as the planning CT. From the planning CT the target volumes are delineated as well as avoidance volumes such as the heart and lungs, which are considered organs at risk (OARs). A treatment plan involves a complex set of instructions dictating the motion and energy output of the treatment machines. A treatment couch can have up to 6 degrees of freedom movement, while the gantry itself can rotate around the patient, and the collimator assembly can rotate along the axis of the beam. The MLCs shape the beam, and the energy level of the delivered photons of the linear accelerator can be changed. In breast plans, treatment planners carefully adjust these parameters to optimize the dose to the target volumes while minimizing dose to OARs and healthy tissues. This is an iterative process that often requires compromise be made between treatment objectives and healthy tissue avoidance. In Volumetric Modulated Arc Therapy (VMAT) these parameters are adjusted by an optimization algorithm based on desired goals and constraints.

An example of an oropharynx (a cancer of the H&N region) treatment plan with the calculated dose distribution is shown in Figure 1-4. The avoidance volume containing the OARs, planning target volume (PTV), and gross tumor volume (GTV) are delineated on the figure. The GTV demonstrates the extent of the malignant growth, whereas the clinical target volume (CTV)

includes the GTV plus an expanded region for malignant cells that are not visible or palpable around the GTV. The planning target volume (PTV) includes the CTV plus an expanded region to account for errors in patient setup, or target location, patient movement, and beam inaccuracies [15]. The top image in Figure 1-4 is an axial view looking in the direction from the feet towards the head of the patient. The bottom image is a view through the coronal plane. The avoidance volume is the summation of many critical structures such as the parotid glands, oral cavity, mandible, esophagus, ears, and several other structures which are not individually outlined in the image but are individually delineated in the treatment plan. In Figure 1-4 the irradiated volume is delineated by the green dose region and some of the healthy tissue was necessarily dosed in order to achieve the prescription dose inside of the PTV.

Once the treatment plan is produced it goes through a series of checks by medical physicists and oncologists to ensure plan quality. When the patient is ready to be treated, they are positioned on a treatment couch. The treatment machines are equipped with imaging devices capable of taking a series of x-rays of the patient, known as cone beam CT (CBCT). The CBCT is used by the radiation therapists during patient setup to check the positioning of the patient relative to the planned position. Comparison is made between tissues on the planning CT and the CBCT; when there is a discrepancy, the patient is shifted to align more closely with the planning CT. The method of aligning the patient with the treatment plan is known as image guided radiotherapy (IGRT). Efforts are made to position the patient with as few CBCTs as possible. Each time a patient receives a CBCT, they receive extra radiation dose to a large area of the body. This dose is added to the treatment dose, and should be minimized according to the principal of ‘as low as reasonably achievable’ (ALARA). Once satisfactory positioning is achieved the treatment is delivered to the

patient according to the fractionation schedule. A single fraction is delivered per day, for 5 days per week, totaling typically 33 fractions for H&N patients, and 5-25 fractions for breast cancer patients. In some cases, the geometry of the patient changes to the degree such that it requires replanning of the treatment. This can occur if the patient loses or gains a large amount of weight, tumor shrinkage, or swelling.

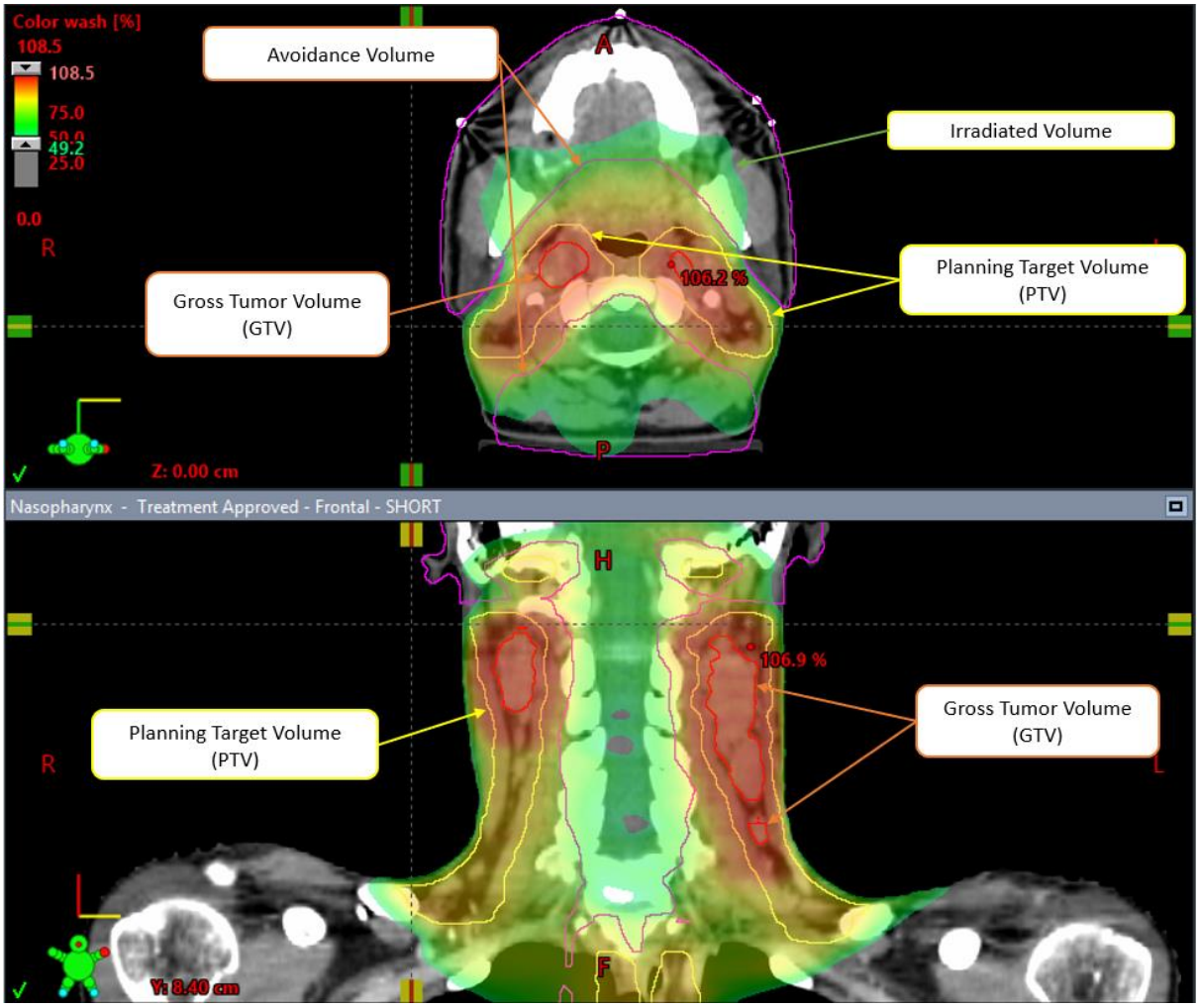

Figure 1-4: Nasopharynx treatment plan. Color wash represents the radiation dose distribution, with red being high dose and green representing low dose. The avoidance volume containing the OARs is delineated by a dark pink contour, and the planning target volume (PTV) is delineated by the yellow contour, the gross tumor volume (GTV) is the visible tumor and is delineated inside of the treatment volume.

The relationship between volume irradiated and dose is typically visualized using a dose-volume histogram (DVH) such as in Figure 1-5, which is easy for humans to interpret and is a simplification of higher dimensional data [19]. Parameters such as the minimum dose to the hottest 2% of the of volume (D_2), and the highest dose delivered to 98% of the target (D_{98}) are read off of the DVH; however, these parameters are only available for regions that have been contoured on the treatment plan. A contour is drawn on the treatment plan to outline ROIs as in Figure 1-4, then the DVH is calculated for each of these contours. Whole-breast plans do not have the breast contoured, so no DVH is available for the breast, although some plans have OAR contours such as heart and lungs as in Figure 1-5. Instead, min/max dose, and other parameters such as dose to the hottest 2cc's of tissue (D2cc), the volume to receive 105% of the prescription dose (V105), and the volume to receive 107% of the prescription dose (V107) can be calculated from the irradiated volume or 50% isodose. These measures are a 2D representation of a 3D dose distribution and some information is lost as the data is projected. The now standard practice for treatment planners is to keep these values below certain levels [20]. This involves making trade-offs between coverage of the target volume, and extra dose to healthy tissue [20]. DVH parameters have also been shown to have limited predictability for acute-phase weight loss when used in machine learning applications [21].

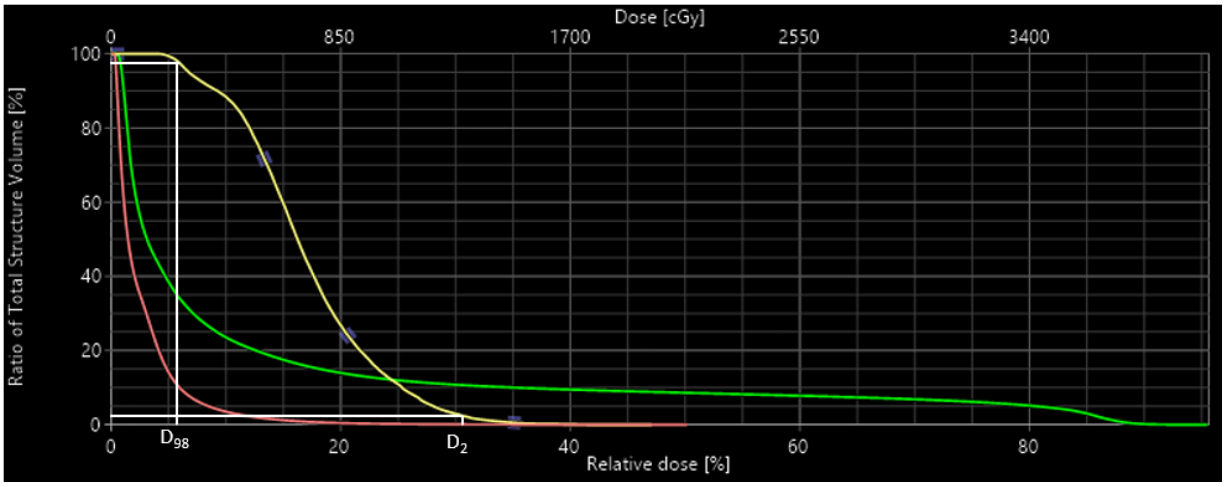

Figure 1-5: Example of a Dose-Volume Histogram. Each line in the figure represents a structure that has been contoured on the treatment plan. The structures in this example from left to right are: heart (pink), left lung (green), and the left anterior descending artery (yellow), which are all avoidance volumes or organs at risk (OAR). The D_2 and the D_{98} for the left anterior descending artery are shown on the plot which represent the minimum dose to the hottest 2% of volume, and the maximum dose covering 98% of volume respectively.

Breast treatment plans have much less complexity than H&N plans. In whole breast treatments the breast tissue is not contoured, there is no GTV, or PTV delineated on the plan. Often there are only a couple of beam angles, and the beams are blocked along the chest wall and remain open anterior to the patient. Recently treatment planning systems have been capable of automatically contouring the heart and the lungs, these are often the only regions that have contours in a breast plan. Breast plans will be discussed in more detail later in the chapter.

1.3 Head and Neck Cancer

In Canada, H&N cancers are the third leading cause of death, and they are the seventh most common cancer worldwide [1]. Historically, the usage of alcohol and tobacco products correlates with H&N cancer incidence; however, the majority of diagnoses in Canada are for HPV positive oropharyngeal cancers [22]. Worldwide, oral and pharyngeal carcinoma has increased by 225% from 1988 to 2004. The emerging population tends to be younger males 40-60 with no history of alcohol and tobacco use, who have had many sexual partners, and have participated in oral sex

[23]. Oral sex tends to be a primary mode of HPV transferal. HPV is recognized as a risk factor for H&N cancers and is associated with increased incidence. Since the majority of these patients are younger and the expectation is that they will live for many years after treatment, the long-term side-effects of treatment are particularly important when treatment planning decisions are made [24]. H&N cancer encompasses a variety of tumor types occurring in areas such as the pharynx (nasopharynx, oropharynx, hypopharynx), nasal cavity, oral cavity, salivary glands, larynx and lips. These areas are sensitive to treatment and can result in permanent undesirable side effects, particularly from radiotherapy [25]–[27].

1.3.1 Treatment

Treatment for head and neck cancers often involves a combination of surgery, radiation therapy, and chemotherapy. There are many combinations of these treatments, but radiotherapy will be the focus in this thesis, although many of the patients in the enclosed study in Chapter 2 will have also received chemotherapy. Radiotherapy is often used for sites that are not easily accessed by surgery, or where the goal is to preserve organ function [28]. Radiotherapy is able to target the region of diseased tissue without the requirement of surgical removal, but surgery typically requires radiotherapy to treat the peripheral area, which could still contain diseased cells [28]. It is vitally important that the radiation dose be localized to the disease tissue, but dose wash often spills into healthy tissue, resulting in radiation induced toxicities [29].

Dose regimens vary from 50-70 Gray (Gy) depending on the region and stage of disease and are typically divided into 30-33 fractions over 6-7 weeks of treatment [30]. In the case of standard fractionation patients are given 1 fraction per day, 5 days per week, until treatment is concluded.

Different fractionation schemes have been tested such as hyperfractionation, where the total number of fractions is increased with decreased dose per fraction, but maintain roughly the same amount of treatment days. In a meta-study of 5 clinical trials it was shown that this method results in an increase in overall 5-year survival of 3.1% compared to standard fractionation; however, altered fractionation alone with no chemotherapy showed a decrease in 5-year survival of 5.8% compared with conventional chemo radiotherapy where both chemotherapy and radiotherapy are conducted together [31]. The cohort of patients examined in this thesis received radical (chemo)radiotherapy with doses between 66-70 Gy in 30-33 fractions.

1.4 Breast Cancer

Breast cancer is one of the most common forms of cancer and accounts for 25% of all cancers in females. Since its peak in 1986 the mortality rate has decreased around 48%, likely due to increased screening and technological advancements in therapy [1]. A common form of therapy is whole breast irradiation in which the entirety of the breast volume is irradiated, usually after lumpectomy [32].

1.4.1 Whole Breast and Chest Wall Irradiation

The standard practice at most centers in Canada is 3D conformal radiation therapy (3DCRT) in which a pair of beams are used, blocked along the chest wall and left open above the patient. A standard beam setup is shown in Figure 1-6 and dose distribution in Figure 1-7. Other methods are being developed which can increase the conformity of the radiation dose which helps to reduce the dose to the lung while treating the breast. One such method being researched is the use of protons for treatment, called proton therapy. Proton therapy has the benefit of spreading out the Bragg peak

which enables the limitation of dose downstream of a given target, and to increase the conformity of the plan to the area being treated. Unfortunately, proton therapy is not in wide use around the world, as the machines are extremely expensive, and none exist currently in Canada. All the methods to try to reduce lung and heart dose have trade-offs, such as extra dose to other healthy tissues, complexity in planning, and increased treatment times or increased overall dose to the patient. This thesis focuses on conventional 3DCRT plans and other methods are not investigated in depth.

Whole breast irradiation is unique in that typically the area is not contoured on the treatment plan and the entirety of the breast tissue is irradiated. Breast tissue not only includes the visible volume but continues laterally down into the armpit and is delineated by the physician or therapist after palpation. The edges of the beams follow a line from the center of the chest to edge of the breast tissue into the armpit. This ensures the entirety of the breast tissue is irradiated, but it also captures a section of the lung, and often a section of the heart [20]. A representation of a patient setup with medial and lateral beams is show in Figure 1-6, and the dose distribution in an axial slice of the patient is shown in Figure 1-7. Partial breast irradiation involves treating only a sub volume of tissue, typically after surgery. The fluid-filled cavity that is left after a tumor is removed is known as the seroma; the area around the seroma could have remaining traces of microscopic disease, so it is often irradiated [33]. In a partial breast plan, the tumor or seroma, and lymphatic nodal regions are contoured. Since there are no contours in a whole-breast plan, the edges of the beam are aligned along the chest wall, and remain open outside of the body; however, with the advance of automatic contouring, more whole-breast treatment plans are having the heart and lung regions contoured, allowing for the calculation of dose to these organs at risk.

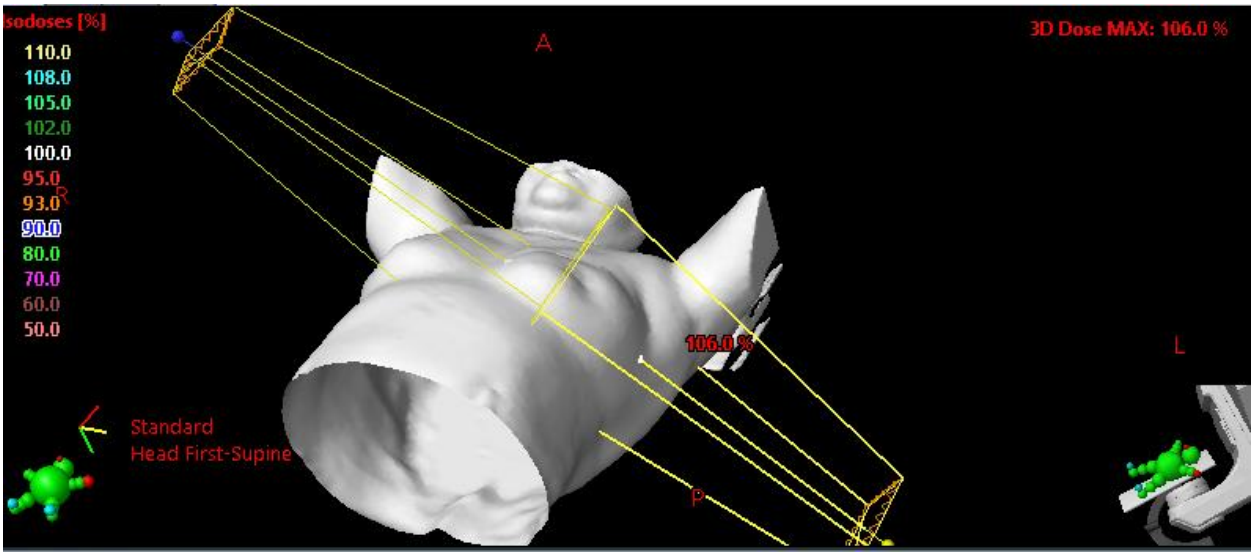

Figure 1-6: Typical patient setup from a whole-breast radiotherapy treatment plan. Two tangential beams are shown. One entering medially and the other entering laterally. The yellow region shows the extent of the fields which diverge gradually as a function of the distance from the source. The lateral beam edge extends into the armpit where the extent of the breast tissue has been identified by palpation. The medial beam enters the patient in the center of the chest, but its extents can change depending on whether lymph nodes are included in the field.

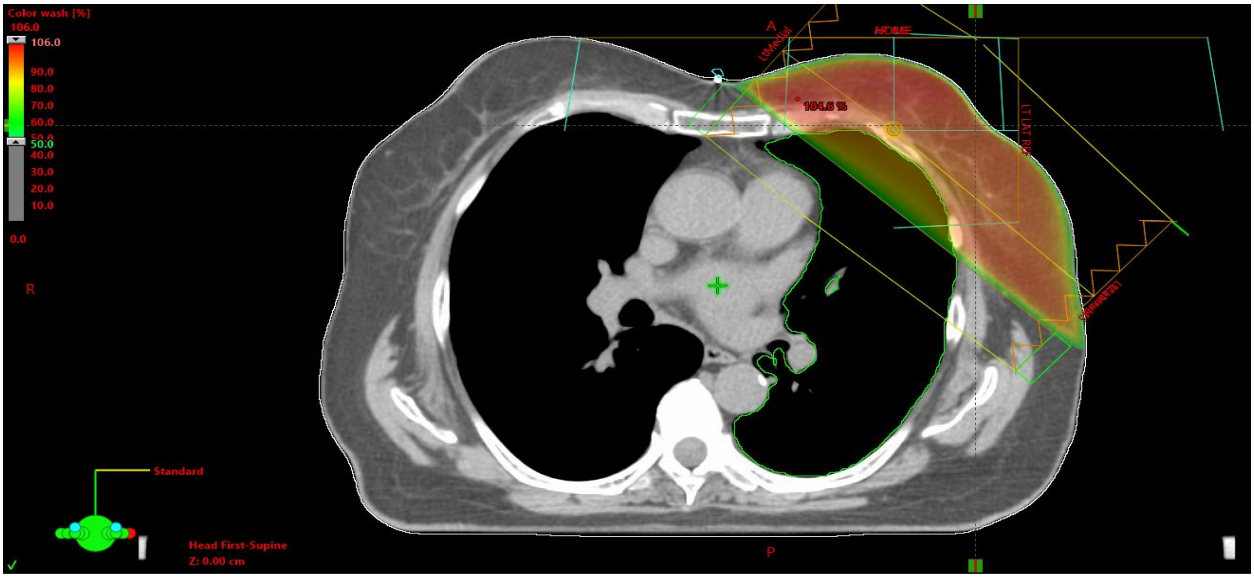

Figure 1-7: Axial slice of a whole-breast treatment plan. The radiation dose follows a line medial to lateral between the extents of the breast tissue. Part of the lung is irradiated inside of the chest cavity. The view is taken from the feet to the head. This image represents a left sided treatment. The dose cloud is shown by the color wash as a percentage of the prescribed dose. The left lung is outlined by a green contour. The point representing the hottest region with the highest dose is shown near the medial edge of the beam and shows a value of 104.6% of the prescribed dose.

1.4.2 Fractionation

The prescribed radiation dose is split into fractions. Randomized trials which showed the benefit of breast radiotherapy employed a dose fractionation scheme of 50Gy over the course of 5-6 weeks; however, shorter courses of fractionation of 13-16 fractions, with the corresponding reduction in radiation dose, have been shown to be equally effective in terms of overall survival, tumor control, cosmetic outcomes and normal tissue toxicity [34]–[36]. Whelan *et al*. showed that 42.56Gy in 16 fractions was effective compared to 50Gy in 25 fractions in terms of 10-year local recurrence and cosmetic outcomes for patients who had radiotherapy following breast conservation surgery [37]. This fractionation scheme has become a standard practice in many regions around the world and is known as 'Canadian Fractionation,' although some provinces such as British Columbia have practiced 40Gy in 16 fractions for mastectomy patients for decades [38].

1.4.3 Breast Size, Dose Homogeneity, and Outcomes

It is generally considered that large breasts are harder to plan, have increased dose inhomogeneity and experience greater toxicity [19], [39]–[41]. Some plans can result in high dose regions where the tissues can receive over 107% of the prescribed dose [42]. Poor treatment outcomes are typically attributed to hot spots in the dose distribution resulting in breast toxicity, where layers of skin can die and peel off [39], [43]–[45]. This is due to inadequate skin sparing and individual patient sensitivity. Skin toxicities are very common in breast cancer patients, and include cosmetic outcomes such as erythema, moist or dry skin desquamation, and occur one to two weeks post treatment. Long-term skin reactions can include fibrosis, dispigmentation, and telangiectasia [19]. Aside from these cosmetic outcomes, other toxicities would include damage to the lungs and heart.

Inhomogeneity in the dose distribution can occur particularly for large breasts. An example of a dose distribution in a large breast is shown in Figure 1-8. The prescribed dose is unevenly distributed throughout the tissue, with concentrations occurring near the beam entrances at the medial and lateral sections of the breast. The aim of the treatment plan is to deliver 100% of the prescription dose to the breast tissue and minimize it to healthy tissue. The red shaded region in Figure 1-8 represents the dose above 100% of the prescription dose. Ideally, this would be uniformly distributed throughout all of the breast tissue but as can be seen in the figure it is concentrated in particular areas. There are limits as to the maximum dose such as <10CC should receive 107% of the prescribed dose, and this treatment plan has a maximum dose of 105.7%, which is within guidelines [46].

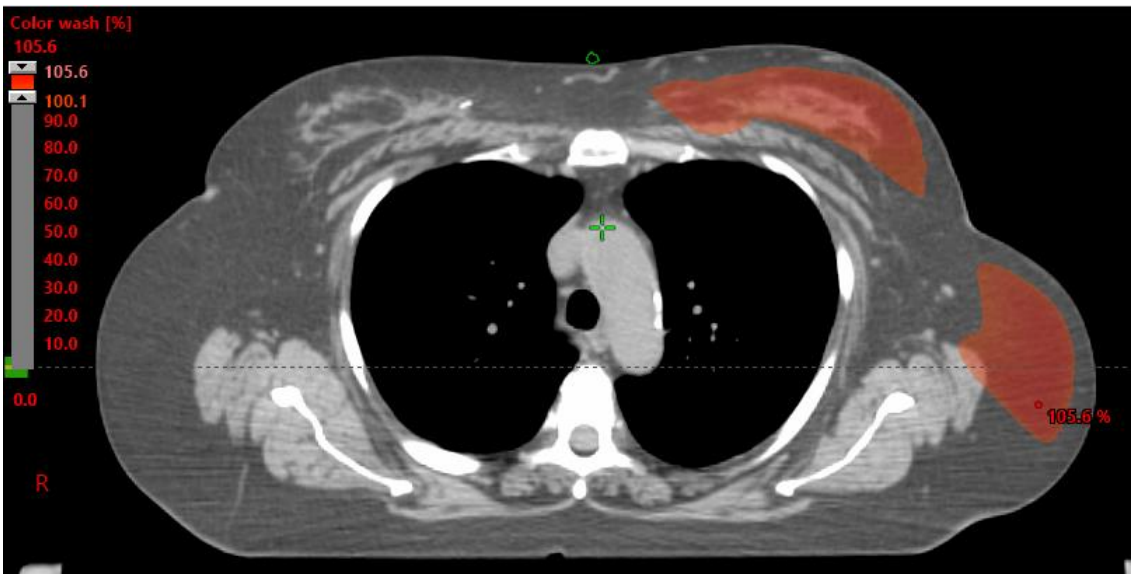

Figure 1-8: Dose Inhomogeneity in large breasts. Large breasts typically see hot spots develop and inhomogeneity in the dose. The red shaded area represents the volume to receive more than 100% of the prescribed dose. The maximum dose is kept below 106% and is within guidelines. The dose distribution is not uniform throughout the volume of tissue.

Figure 1-9 depicts a treatment plan for a small breast. The area to receive 100% or more of the prescribed dose is more evenly distributed around the cancerous tissue, and throughout the whole

breast. The maximum dose in this plan is 104.5% of the prescription. The field has been shaped to lower the dose to the chest wall.

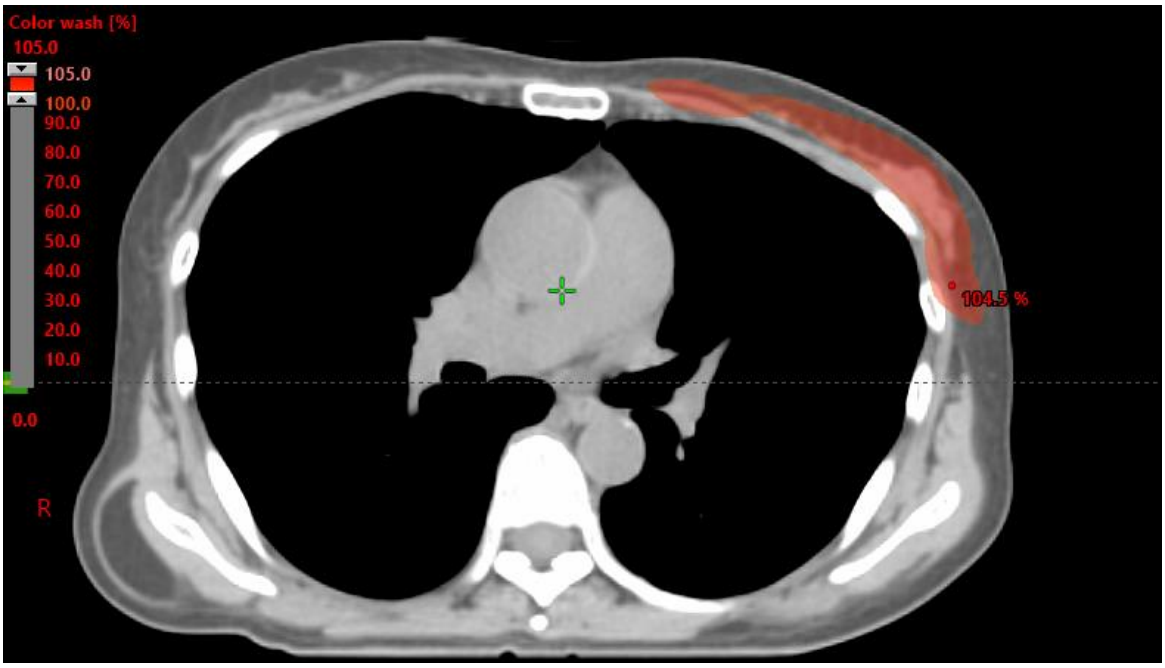

Figure 1-9: Dose Distribution in small breast. The dose distribution is more uniform in smaller breasts as the prescribed dose covers the tissue intended to be treated. Hot spots are less prevalent than in large breasts. The red shaded area represents the region of tissue receiving 100% or more of the prescription dose, with the maximum dose being 104.5%.

1.5 Problems with radiotherapy – Toxicity, secondary cancers, cosmetic outcomes

Following the initial enthusiasm after the discovery of x-rays, they were propounded as a method to treat a wide range of conditions, from tuberculosis to diabetes. It was not until later that the damaging effect of radiation was starting to be recognized [47], and many of the early researchers who worked with radioactive material died from radiation induced diseases. Although beneficial in the treatment of cancer, radiotherapy has many side effects: secondary cancers can be caused when healthy tissue is dosed, areas that receive high doses can have necrosis, and other toxicities specific to the region being treated.

1.5.1 Xerostomia and Dysphagia

Because of the sensitive nature of the organs located in the H&N region, patients often experience post-treatment side effects that have a negative impact on their quality of life (QoL). Xerostomia is the inability to produce saliva, and dysphagia is difficulty swallowing. These are a couple of the most common and most impactful toxicities experienced by H&N cancer patients and can be caused by receiving radiation therapy for cancer treatment [27][48].

Xerostomia is characterized by the inability to produce saliva and is often caused by radiation damage to the parotid glands [27]. This leads to dry mouth, and can consequently lead to dysphagia, or difficulty swallowing. Patients with xerostomia suffer from oral discomfort and pain, and find it difficult to chew, speak, or swallow, and can have heightened susceptibility to other oral conditions such as cavities and infection. Xerostomia can start early in treatment where in the first week 50-60% decrease in saliva flow can occur, with a final flow rate of approximately 20% for most patients after 7 weeks of treatment [49].

The severity of xerostomia can be measured to get a sense of the degree to which saliva production has been impacted, resulting in symptom grades, although the methods are not straightforward. Measurement of saliva production is important to be able to determine the impact that preventative or curative efforts are having, although there is weak correlation between measured salivary flow and xerostomia symptom grades [27]. This may be due to variation in normal saliva production between patients and hydration status of the mucosa [50]. This makes it difficult to correlate measurements of saliva production with QoL; however, xerostomia is often cited as a major impact factor in patient reported QoL [27].

Dysphagia is characterized by difficulty swallowing, typically resulting from movement abnormalities of the oral cavity, hypopharynx, oropharynx, larynx, and upper esophageal sphincter. This results in more time and effort required to move food or drinks from the mouth into the stomach and in the most severe cases, swallowing is impossible. Other symptoms of dysphagia can be: pain while swallowing, feeling like food is getting stuck in the esophagus, drooling, gagging, heartburn, weight loss, and coughing while swallowing [51]. Dysphagia can be caused by surgery for tumor removal, or from damage to muscles and other organs used in swallowing after radiotherapy [48]. Dysphagia can be accompanied by xerostomia causing difficulty chewing, and poor hydration of food prior to transfer into the stomach, causing food to get caught in the throat, or travel slowly. Dysphagia can also be a primary symptom of H&N cancer when a tumor impedes or blocks swallowing function.

Xerostomia and dysphagia contribute significantly to post treatment QoL. In conjunction with xerostomia and dysphagia, patients often report difficulty chewing, loss of taste, altered speech, difficulty with dentures, tooth decay, and pain in the mouth and throat. They can also experience effects on their mood, have difficulty socializing, and become embarrassed and avoid social interaction because of the perceived judgement of others while having difficulty swallowing [52][27]. They can become depressed because of the feeling of social isolation, since many interactions with family and friends are centered around meals. Patient reported outcome measures (PROM) help to track the impact on patient QoL so we can get a better understanding of the broader influence of these symptoms on patients' experiences post-cancer treatment.

1.5.2 Toxicities of the Heart and Lungs

The standard tangential beams in breast radiotherapy give rise to dose to the heart. There is well documented evidence that heart dose leads to increased cardiac toxicity and morbidity [53]. This result is more pronounced for left-sided breast irradiation when compared with women who have had right-sided breast irradiation. With patients being diagnosed at earlier stages and life expectancy increasing to greater than 15 years, these patients have an increased rate of mortality due to cardiac failure. The risk of cardiac complications increases with younger age and presence of other risk factors [53].

It is not clear how much of an increase in risk there is associated with cardiac dose. Phase 3 clinical trials from the Early Breast Cancer Trialists' Collaborative Group (EBCTCG) assessed that an increased risk of heart disease was 3% per Gy of radiation [22]; however, this study was retrospectively conducted for patients treated prior to the days when CT scans were taken. The study utilized a single representative CT and applied the treatment plan for each patient to the representative scan. Data from the Hiroshima bombing survivors showed an increase in cardiac death risk of 14% per Gy and there is a higher dose response relationship for survivors exposed at < 40 years of age [53]. This large discrepancy could be a result of homogeneous dosing in the case of the Hiroshima survivors, versus inhomogeneous dose in the case of breast cancer patients as well as receiving the dose in fractions versus all at once. Although it is unclear of the cardiac mortality risk for doses below 0.5 Gray, at some point above that there is a clear association between the level of exposure and cardiac death risk. Modern cardiac dose constraints for whole breast radiotherapy recommend limiting mean heart dose to <4 Gy, and this is generally considered to be an achievable standard.

Toxicities of the lungs can develop in patients who have received radiation therapy for a different part of the body. Incidental lung dose can happen while trying to dose other areas such as the breast [49]. Radiation pneumonitis is particularly prevalent in patients receiving whole breast radiotherapy including the chest wall. The ipsilateral lung, or the lung on the same side of the body, receives radiation dose while treating the chest wall. The incidence of radiation pneumonitis also increases with the use of systemic treatments such as chemotherapy combined with radiation [54]. Radiation pneumonitis incidence as much as 13% can be observed with patients who have had prior systemic therapy [49]. The combination with other treatments such as antihormonal and antibody protocols also increase the risk of pneumonitis [55].

1.6 Patient Reported Outcomes

PROMs are clinical surveys that are presented to patients during or after treatment. These tools are useful to help identify areas of concern, symptoms, and impacts on patient's QoL following treatment from the lens of the patient reported outcome (PRO). Often the goal in cancer treatment is disease control, or long-term survival. When only these metrics are used to determine the success of a treatment, considerations such as a patients' QoL can be overlooked or ignored in pursuit of gains in measured outcomes. Treatments that can give marginal benefit in terms of local control or long-term survival can sometimes have significant and damaging side effects which cause the patient great discomfort or difficulty living their life, and have impacts on their social interactions which play a key role in maintaining mental health [56]. Administering PRO instruments has shown to increase survival, improve quality of life, decrease hospitalizations, improve patient and caregiver communication, and increase patient satisfaction [9][52]–[55].

1.6.1 MDASI-HN

The M.D. Anderson symptom inventory is a validated PRO instrument used to measure symptom burden and the interference of those symptoms in daily life. The H&N module (MDASI-HN) consists of a set of questions specific to patients being treated for H&N cancer [61]. Cancer treatment in these areas can result in a wide range of symptoms related to surgery, radiation therapy or lead to general feelings of overall sickness and social isolation. The MDASI-HN was developed not only to measure functional impairment but also the patient's perceived impact of treatment related symptoms on their QoL. Physicians often measure functional impairment, but this does not take into account symptom distress caused to the patient. Measuring symptom distress, or symptom interference helps the clinician understand the most impactful symptoms to a patient's daily life and permits them to apply appropriate interventions to minimize impact regardless of functional impairment.

The MDASI-HN utilizes a 0 to 10 scale to indicate symptom severity and presence. A score of 0 indicates not present, and 10 indicates, “as bad as you can imagine”. The patients are asked to rate their symptoms for the period of the last 24 hours.

1.6.2 MDADI

The M.D. Anderson Dysphagia Inventory (MDADI) is another PROM administered to head and neck cancer patients. Its purpose is to assess the effect of dysphagia, or difficulty swallowing, after receiving head and neck cancer treatment on QoL. The MDADI is a more general PRO instrument in terms that it measures overall impact on quality of life, rather than measuring specific symptoms

like the MDASI-HN. It includes four question groups: emotional, functional, physical, and global totalling 20 questions scored on a 5-point Likert scale (strongly agree, agree, no opinion, disagree, strongly disagree). The majority of questions are phrased such that the worst symptoms are associated with “strongly agree” being given a score of five (5), and none to mild symptoms associated with “strongly disagree” given a score of one (1); however, two questions: “I do not feel self-conscious when I eat,” and “I feel free to go out and eat with friends, neighbors, and relatives,” have reverse scoring. “Strongly agree” would correspond to a score of one, or none to mild symptoms, while “strongly disagree” would correspond to severe symptoms or a score of five.

These two PROMs are both administered to head and neck cancer patients but they have slightly different purposes. The MDASI-HN is primarily a symptom reporting device where patients are able to label the severity of their symptoms or side effects [61][26]. The MDADI is primarily a QoL measure, where patients label the impact of dysphagia on their ability to perform normal activities, and the impact on their social and mental well-being [52]. Where the MDASI-HN will have questions that ask to what degree is a symptom painful, the MDADI would ask to what degree does the symptom affect normal functioning. The MDASI-HN includes a wide range of symptoms common to head and neck cancer patients, which also includes dysphagia. The MDADI is specific to dysphagia, but many other symptoms can also impact quality of life such as xerostomia which is closely related to dysphagia [45][58]. It is useful to be able to compare the symptoms scoring from the MDASI-HN with the quality-of-life scoring on the MDADI to get a sense of which other symptoms can have an impact on a person's ability to function normally. This comparison is made in chapter 2 of this thesis.

1.7 Analysis Methods

1.7.1 Radiomics

The study of radiomics involves extracting statistical features from medical image pixel or voxel data. Imaging modalities such as computed tomography (CT), magnetic resonance imaging (MRI), or other medical imaging technologies generate large amounts of data which can include patterns that are beyond human capability to discover [64][65]. With the help of software, mathematical models can extract thousands of features from the 2D or 3D pixel (‘voxel’) data [66]. Recently, there has been growing interest in using radiomics features in the prediction of cancer specific outcomes [67] and diagnosis [68]. An interesting discovery, is that patients with Human Papilloma Virus (HPV), a known cause of H&N cancers, have smaller tumors [69]. Radiomics has also been used to predict treatment outcomes in head-and-neck [70], and lung cancer [71]; however, radiomics does not consider the radiation dose to the patient.

The process of extracting radiomic features begins with the CT scan. A region of interest (ROI) is delineated on the scan, typically as contours from a radiotherapy plan. The ROI is turned into a mask which has the same shape as the ROI, but has binary values for regions inside and outside of the ROI. This constrains the feature extraction to the area inside of the mask. The mask is then applied over top of the CT slices, typically over top of the contour that the mask was generated from, but the mask can be applied to any region of an image. The slices are stacked together to generate 3D volumes. The intensity of the individual voxels then serves as the basis from which the features are extracted. Unlike a color image, a CT image only has a single channel of grey level

intensities that are sensitive to the image acquisition and reconstruction settings of a particular scanner.

Descriptions of classes of features that can be extracted are described in the Imaging Biomarker Standardization Initiative (IBSI) documentation [52]. Some of the feature classes that are available include:

- Shape based: These features are 2D and 3D descriptions of the geometry of the mask such as the volume of the ROI, surface area of the mesh, compactness, surface to volume ratio, sphericity, max/min diameter, and others.
- Local intensity features: Common statistics describing the intensity distribution in the ROI such as: mean, variance, min/max, skewness, energy, where first order energy is the summation of all the intensities inside the ROI.
- Grey Level Co-occurrence Matrix (GLCM): Expresses how combinations of intensities are distributed throughout the image directions, for 3D images there are 13 directions. It contains the frequency in which combinations of grey levels occur in neighboring voxels in each direction. Several statistics are calculated from the GLCM such as: joint average, joint variance, difference average, and the angular second moment.
- Grey Level Run Length Matrix (GLRLM): The number of consecutive voxels with the same grey level intensity along a particular direction are counted and classified as a run. Statistics from the distribution of these runs can be calculated such as: long runs emphasis, high grey level run emphasis, and. grey level non-uniformity (GLN).
- Grey Level Size Zone Matrix (GLSZM): Looks at the number of groups or zones of voxels that are linked to neighboring voxels by having an equivalent grey value. It represents the number

of groups of voxels with a given intensity. Features calculated from this matrix include: small zone emphasis, large zone emphasis, and gray level non-uniformity.

1.7.2 Dosiomics

The same methods to extract radiomic features can be applied to a radiotherapy dose cloud, which is known as dosiomics [72]–[74]. The dose cloud is calculated on the CT scan, taking into account the complex interaction between radiation and tissue. More recently, studies have begun to utilize dosiomics to predict cancer outcomes and toxicities. It's been effective in predicting radiation pneumonitis [74]–[76], fibrosis [44], weight loss [21] and shown to be better at modelling toxicity probabilities for H&N cancer than analytical methods utilizing DVHs [77]. Dosiomics have even shown promising results in predicting cancer recurrence in H&N patients [77]. This is a recent field with most studies published in 2020, and to the best of our knowledge, no work had been done for breast cancer patients.

The dosiomic study in Chapter 3 of this thesis utilizes the 50% Isodose to generate the mask for the ROI. The 50% Isodose is the volume to receive 50% of the prescribed radiation dose, and is considered the irradiated volume. Voxels with intensity values 50% or higher of the prescribed dose are included in the ROI. Unlike the voxel intensities of imaging modalities such as CT which are given in grey-level intensities and are arbitrary, the voxel intensities for the dose cloud are converted to the radiation units of Gray (Gy), which are a measure of absorbed energy in joules per kilogram.

1.7.3 Machine Learning

Machine learning techniques fall into two broad categories: supervised and unsupervised, although reinforcement learning could be considered a third category, it is not discussed here.

1.7.4 Unsupervised Learning

Unsupervised machine learning involves attempting to extract patterns from unlabeled data. It aims to discover structure hidden in data points that would otherwise be intractable for humans. Unsupervised machine learning methods are useful for data sets that lack a known outcome or reward function and for dimensionality reduction tasks [78][79][80]. Individual data points are scored based on their similarity to other data and groups or clusters are formed based on the distance between features [78]. These methods were used in the studies in chapter 2 and 3 of this thesis.

Hierarchical clustering is a method of unsupervised machine learning that groups data together based on distance metrics and then organizes the clusters into a hierarchy [78]. The aim is to group items while maximizing the difference between groups. As the data are grouped, the distance between clusters is calculated then the most similar clusters are combined. The process continues in this manner and clusters are combined at higher levels of the hierarchy until a single cluster remains. This is especially useful for dimensionality reduction where the number of clusters to examine can then be selected by choosing a level of the hierarchy with a threshold value [79]. A lower threshold provides a greater number of clusters, while a high threshold reduces the number of clusters [81]. After choosing the number of clusters to use, a single feature can be selected from each cluster. This is most valuable for data sets where there is high correlation between the

features, the correlation matrix can be clustered, and features can be selected which have low correlation with each other. High dimensional datasets with many correlated features can cause overfitting, bias, and poor generalizability if the features are used in supervised machine learning models [59][62][83]. This can be due to the high complexity of the data set where the extraneous features act as noise, or more specifically, the sparsity of the data increases, known as the curse of dimensionality [84]. As the dimensionality of the data increases, the volume that the data must occupy increases exponentially. For example, if data are arranged on a number line of x units in length and you add another feature or axis along which the data can be arranged, then the data occupies a space consisting of x^2 $units^2$. Adding yet another axis increases the feature space to x^3 $units^3$ of which the same number of data samples must occupy. Previous groups or cluster of data become harder to distinguish in this large dimensional space [84].

1.7.5 Supervised Learning

In supervised techniques, data is provided which has been labeled with a particular category or outcome of interest; usually labelling is done by humans. The supervised algorithm is then trained on this labeled data, and learns patterns in a similar manner to how humans learn [83]. Some methods involve directly inputting the raw pixel or voxel data into ML algorithms such as neural networks (NN) to learn patterns which correlate with the labeled outcome [42][64]. The NN is then used to make predictions from previously unseen data. Algorithms such as NNs lack interpretability since their decision pathway is hidden in a black-box type structure, making them less useful in medical applications even when their predictive ability is high [85]. Other methods involve feature engineering where features of the data can be manually extracted and combined in meaningful ways in an attempt to decrease the dimensionality of the data [79]. The remaining

discussion on supervised machine learning classifiers in this thesis will focus on decision trees and random forests.

1.7.6 Decision Trees

A decision tree is a type of supervised machine learning classifier which splits the data set based on its input features [86]. An example of a decision tree is shown in Figure 1-10. Decision trees for machine learning utilize algorithms such as information gain, or Gini importance to determine the optimal split [87][86]. The data is split into smaller groups at branch nodes which have arrows pointing into and out of the node. The termination, where a label is provided in terms of the model's prediction, is known as a leaf node. Leaf nodes can contain one or more data points although having leaf nodes with single data points could be an indication of over fitting [87]. Decision trees on their own have low predictive capacity, and are highly susceptible to the subset of features chosen to compare at each branch node [87]. A method of addressing these limitations is to group the trees into a forest.

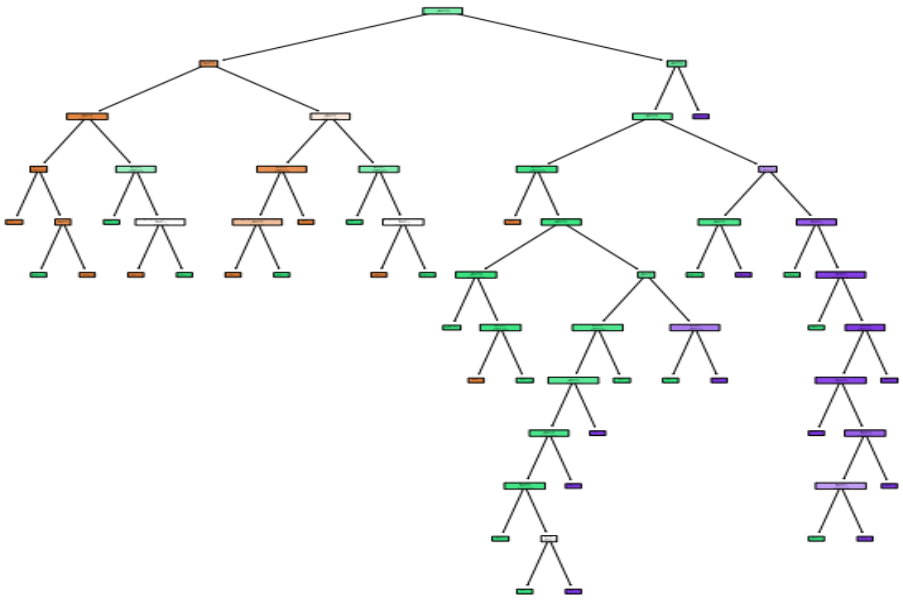

Figure 1-10: Decision tree. Each branch node as arrows pointing to, and away from it. Data is split at branch nodes until termination at a leaf where a prediction is given.

1.7.7 Random Forests

Random forests have better generalization ability and are less susceptible to overfitting than single decision trees [61][62]. Demonstration of their use is detailed in Chapter 3 of this thesis. They are built by drawing a random bootstrapped sample from the data set with replacement that is the same size as the original data set. A decision tree is grown from each bootstrapped sample and at each node a random selection of features is chosen without replacement to test for a split. The node is split using the feature providing the best results according to the objective function such as information gain, or Gini importance [88]. The process is repeated for the number of trees in the forest. Then the class label, or prediction, is assigned by aggregating the majority vote from all the trees in the forest.

1.7.8 Feature Importance

Random forests suffer from a decrease of interpretability when compared to decision trees [87], [88]. Since the class label is determined through the aggregation of hundreds of decision trees, it is difficult to determine the best series of splits that was used by the random forest to make a particular prediction. A method of lending interpretability to random forests is through permutation feature importance [88]. The values of a feature in the data set are permuted, where the values are shuffled in a random order. This breaks the link between the feature and the classification label. The decrease in accuracy of the prediction is measured for each permutated feature. The feature that results in the largest decrease in accuracy, receives the highest importance value. This process is repeated several times and the results are recorded to get a distribution of each feature's importance.

In high dimensional data sets containing a large number of correlated features, permutation feature importance can suffer bias because the model is able to access a feature through its correlate [82]. This results in both features receiving lower importance values even if they are highly important [89]. A method of dealing with this limitation is to hierarchically cluster the features from their correlation matrix [79]. A correlation matrix can be transformed for use in hierarchical clustering by taking the complement of the absolute vale of the correlations. This results in highly correlated features being given a low value for use as a distance measure in hierarchical clustering. The desired number of clusters are then selected, and a single feature from each cluster is used in the random forest classifier, reducing the correlation between features.

1.8 Thesis Overview

This thesis addresses some applications of machine learning to problems in radiation therapy cancer treatments. Chapter 2 looks at the relationship between two patient reported outcome instruments the MDASI-HN, and the MDADI. The two surveys are compared for a cohort of cancer patients who underwent treatment for head and neck cancer at the Tom Baker Cancer Centre. It examines how symptoms reported on the MDASI-HN could be reflected in quality of life impacts as reported on the MDADI. Correlation analysis found little overlap between items in the instruments. A group of patients was identified who reported no symptoms on the MDASI-HN but reported quality of life issues on the MDADI; hierarchical clustering was able to show common patterns in their responses. A recommendation is made to help clinicians use the MDASI-HN to flag patients who may require referral to specialists to deal with problematic side effects.

Chapter 3 is analysis of dosiomic features extracted from whole breast radiotherapy plans. Feature reduction was achieved through grouping in hierarchical clustering, then selection of a single feature from each cluster. The patients were then stratified based on treatment volume, and the dosiomic features were examined to determine if relationships exist between breast size and patterns in the radiation dose distribution. A small group of features was identified that provide a higher prediction accuracy inside of a random forest classifier than using the entire feature set.

CHAPTER 2

A COMPARATIVE STUDY OF THE M.D. ANDERSON SYMPTOM INVENTORY WITH THE M.D. ANDERSON DYSPHAGIA INVENTORY FOR HEAD AND NECK CANCER PATIENTS

Adam Yarschenko, Demetra Yannitsos, Sarah Weppler, Lisa Barbera, Harvey Quon, Qiao Sun, Wendy Smith

2.1 General Introduction

Patients treated with radiotherapy for head and neck cancer often experience a wide range of side effects. It is important to understand how these side effects impact patients' quality of life (QoL) post treatment. For this purpose, QoL instruments have been developed which allows patients to give feedback through questionnaires based on their symptom burden and the effect of these symptoms on their life in general. This study examines two QoL instruments, the M.D. Anderson Symptom Inventory - Head and Neck Module (MDASI-HN), and the M.D. Anderson Dysphagia Inventory (MDADI). It compares the two items through statistical and machine learning analysis to determine if administering the MDASI-HN alone alerts clinicians to all patients who require referral to specialists to address radiotherapy related side effects. Hierarchical clustering was used to identify what severe QoL impacts patients who were missed on the MDASI-

HN, were experiencing as captured by the MDADI. This paper was written for a clinical journal with less focus on technical details.

The work conducted in this chapter contains a study comparing patient reported outcome instruments. I was the primary author, and the contributing authors were Demetra Yannitsos, Dr. Sarah Weppler, Dr. Lisa Barbera, Dr. Harvey Quon, Dr. Qiao Sun, and Dr. Wendy Smith. I performed the analysis, contributed to the first draft of the manuscript, and wrote the subsequent revisions and edits. Demetra Yannitsos contributed to writing and editing the manuscript. Dr. Sarah Weppler designed the study, collected the data, and provided feedback for the analysis. Dr. Barbera and Dr. Quon read the manuscript, provided feedback, clinical perspective, and contributed to the design of the study. Dr. Sun read the manuscript and provided feedback on the machine learning methods and analysis. Dr. Smith contributed to the study design, supervision of the study, and provided guidance and feedback for the manuscript. Correlation analysis, unsupervised machine learning, and sensitivity analysis was used to compare the two surveys and to help make recommendations based on a clinical perspective.

2.2 Abstract

Background: Head and neck cancer patients often report a wide range of symptoms after treatment impacting quality of life (QoL). In a busy clinical setting where patient reported outcome measures (PROMs) may be administered at multiple patient visits, it is advantageous to capture these symptoms with as few questions as possible. In this study the M.D. Anderson Head and Neck Symptom Inventory (MDASI-HN), and the M.D. Anderson Dysphagia Inventory (MDADI) are

compared to determine if using the MDASI-HN alone would overlook symptoms identified with MDADI.

Methods: The MDASI-HN and the MDADI were completed by 156 patients, post radiotherapy for head and neck cancer. Associations between the two instruments were analyzed using correlation analysis, unsupervised machine learning, and sensitivity analysis.

Results: Overall, little correlation was found between the two surveys; however, there was overlap between MDASI-HN dry mouth and many MDADI items, confirming that dry mouth is an important factor in difficulty swallowing, and patient QoL. Taking longer to eat (MDADI), was the most commonly reported item overall, with 85 (54%) patients rating it as moderate/severe. Dry mouth was the most endorsed MDASI-HN item (68, 44%). There were 51 patients missed by the MDASI-HN, reporting no moderate/severe symptoms on MDASI-HN, but reported one or more moderate/severe QoL impacts on MDADI. If patients who reported a score of 2 or higher on the MDASI-HN Dry Mouth item are flagged as requiring follow-up, the number of patients missed by MDASI-HN drops to 15.

Conclusions: In a head and neck cancer clinic where MDASI-HN is routinely administered, assessment of symptoms and QoL might be enhanced by including the following MDADI items: Taking longer to eat, why can't you eat this, and limiting food intake, or by reducing the value at which MDASI Dry Mouth is considered moderate/severe to 2.

2.3 Plain English Summary

Individuals with head and neck cancer can experience many distressing symptoms after treatment. Clinicians must identify and asses these symptoms, and this is often done through questionnaires. For repeated use in a cancer clinic setting, the optimal questionnaire minimizes the number of

questions while capturing the broadest range of the most important symptoms impacting patients' quality of life. This study compares the MD Anderson Symptom Inventory – Head and Neck (MDASI-HN), to a quality-of-life questionnaire, the MD Anderson Dysphagia Inventory (MDADI). The study aimed to answer two questions: Does the MDASI-HN fully capture the range of symptoms impacting head and neck cancer patients' quality of life, and can this be improved by including items from the MDADI for use in a clinical setting? We found that there was little correlation between the two surveys, and that one major item from the MDASI-HN, 'dry mouth,' was responsible for most of the impact on quality of life; however, certain issues were not captured by the MDASI-HN. To alert clinicians to these symptoms, augmenting the MDASI-HN with the following MDASI-HN items should be considered: "It takes me longer to eat because of my swallowing problem", "People ask me, 'Why can't you eat that?'", and "I limit my food intake because of my swallowing difficulty". Another option would be to consider patients who gave a score of ≤2 on the MDASI-HN dry mouth item as requiring further inquiry. Overall, if physicians are only collecting MDASI-HN from patients, then it is important to be aware of the gaps in symptom assessment, some of which can be captured by using select MDADI items.

2.4 Background

Individuals diagnosed with head and neck cancer (HNC) comprise a complex patient population, with many undergoing multimodal cancer treatments and requiring extensive symptom management [29][90][91][92]. Treatment for HNC is determined by site and stage and includes a combination of surgery, radiation, or chemotherapy [93]. Despite technological advancements in surgical procedures [94] and Intensity Modulated Radiotherapy (IMRT)[95], HNC patients still experience a heavy symptom burden impacting their physical, functional, and psychosocial

wellbeing [96][97][98][99][56]. Common symptoms include oral pain, decrease in appetite, dysphagia, dry mouth, and fatigue [97][98][99][56]. These patients may frequently use the emergency department or be admitted to hospital for support while undergoing radiation therapy, making symptom management essential in minimizing these hospital visits [100]. Therefore, early identification and support is necessary to effectively manage and reduce symptom severity.

Patient reported outcome measures (PROMs) can be used to identify a patient's priority symptom concerns in clinical practice [96][60][58]. Symptom scoring above a threshold can flag a need for intervention. In addition, having patients report their symptom experiences through PROMs has been shown to improve and facilitate patient-clinician communication [96][60][58], reduce ER visits [9], and increase overall survival [101]. Clinicians can then intervene or prompt patient specific supports including lifestyle modifications, prescribing medications, referring patients to other healthcare professionals, or other personalized symptom supports.

There are a wide number of PROMs validated for the HNC population [97][98][58]. The MD Anderson Symptom Inventory – Head and Neck (MDASI-HN) is a validated HNC-specific measure that has been increasingly implemented into routine care [61]. This tool has been adopted into various Canadian health systems and endorsed by the Canadian Partnership for Quality Radiotherapy (CPQR) [57]. The MDASI-HN includes items regarding the most common general cancer symptoms, HNC-specific symptoms, as well as items describing the extent to which symptoms interfere with daily life [61].

Another PROM relevant to the HNC population is the MD Anderson Dysphagia Inventory (MDADI) [52]. The MDADI has a specific symptom focus, with multiple items measuring how patients experience dysphagia. Dysphagia is amongst the most common symptoms HNC patients experience [98] with approximately 30% of locally advanced HNC patients experiencing grade 2 or worse dysphagia [102] based on objective toxicity assessments. Xerostomia is also amongst the most common symptoms experienced by head and neck patients [9][10]; however, dysphagia has been shown to have a greater negative effect on quality of life compared to xerostomia, even though xerostomia is typically reported more often [63].

Each HNC PROM has its own unique elements and captures a set of items differently from other measures [103]. While it is desirable to assess the broadest set of symptoms, in a busy clinical setting it is not feasible to administer multiple lengthy PROMs in order to capture this data. Therefore, it is important to understand the limitations of a questionnaire and to address these with direct interviewing or by supplementing with additional items. The MDASI-HN is currently being used routinely in many head and neck cancer clinics, as it captures a wide range of the most common symptoms; however, it is important to understand if using the MDASI-HN solely is sufficient at capturing highly endorsed HNC symptoms from patients. Therefore, the aim of this study was to identify if there are gaps in the MDASI-HN that are captured by the MDADI.

2.5 Methods

2.5.1 Study Patients and Data Collection

A cohort of 156 patients were recruited at Tom Baker Cancer Centre in Calgary, AB between June, and October 2019, while undergoing routine follow up appointments after completing treatment.

All participants received radiation therapy for head and neck cancer. Time since treatment completion ranged from 3 months to 6 years. Eligibility included adult patients who received treatment with radical volumetric modulated arc therapy (VMAT) (chemo)radiation with doses between 66-70 Gy, and weekly cone beam CTs (CBCT) for image guided radiotherapy. Patients with confirmed recurrence of head and neck cancer prior to survey completion were excluded.

This was a cross-sectional study in which patients were asked to complete a one-time PRO questionnaire on paper. The PRO questionnaire included the MDASI-HN and MDADI instruments and were completed concurrently. Written consent was provided by patients according to ethics approval (HREBA.CC-19-0119).

2.5.2 Patient Reported Outcome Instruments

The MDASI-HN includes a total of 28 items, with subgroups including core cancer symptoms (13 questions), head and neck specific symptoms (9 questions) and the extent that symptoms interfere with daily living (6 questions) [61][62]. Each item is ranked on a scale from 0 to 10, with 0 indicating the symptom is not present and 10 indicating the symptom is severe. Symptom burden scores are categorized as: none (rating of 0); mild (1 - 4); moderate (5 - 6); or severe (7 - 10). Rosenthal *et al*., and Cleeland *et al*. reported findings for patients with moderate to severe (5-10) symptoms when choosing questions during the development of the MDASI-HN [61][62]. To focus on clinically relevant symptom scores, and to align with the methods used during the development of the MDASI-HN, moderate and severe symptom scores were amalgamated. Therefore, scoring for individual items were dichotomized into None/Mild (<5) and Moderate/Severe (≥5) responses

[61]. The full questions along with their abbreviations used in the analysis are shown in supplementary Table S 1.

The MDADI includes 20 items with three subgroups: emotional, functional, physical, plus a single general question on how swallowing impacted daily routine. Responses for each item are indicated on a 5-point Likert scale (strongly agree, agree, no opinion, disagree, strongly disagree) [52]. Questions are typically phrased such that a response of "strongly agree" indicates severe symptoms. Each Likert response was assigned a number from 1 to 5, with 5 being the most severe. Two questions: "I do not feel self-conscious when I eat," and "I feel free to go out and eat with friends, neighbors, and relatives" have reverse scoring. A response of "strongly agree" corresponds to a score of 1, indicating mild symptoms and "strongly disagree" corresponds to 5, or severe symptoms. In order to compare symptom severity groups with MDASI-HN, responses of strongly disagree (1), disagree (2), and no opinion (3) were grouped into the none/mild category. Agree (4) and strongly agree (5) were grouped into the moderate/severe category. This cut off was chosen to reflect clinical practice of treating moderate to severe symptoms, and facilitate comparison with MDASI-HN. Full MDADI questions with their abbreviated version used in the analysis are shown in supplementary

Table S 2.

2.6 Analysis

2.6.1 Correlations

Descriptive statistics were used to summarize patient demographics and disease characteristics. A correlation analysis was conducted between every question pair using Spearman correlation coefficient. Inter-instrument and intra-instrument correlations were examined. The results of the comprehensive correlation analysis are not shown for brevity. Selected individual questions with the highest correlation were examined in detail. MDASI dry mouth is included because of its frequency of moderate/severe endorsement by patients. For this family of comparisons, a significance level of 0.05 was used and Bonferroni multiple testing correction applied.

2.6.2 Clustering

Hierarchical clustering is an unsupervised machine learning method which groups items based on similarity, and was used to identify groupings of patients and questions. These groups were used to assess the relative presence of symptoms compared to the overall questionnaire, and to identify highly endorsed symptoms that may not be captured on the MDASI-HN. The relationships of the groupings are illustrated by dendrograms on the top and left side of the heatmap. Groups are sequentially merged based on similarity until one group remains. Similarity is calculated using a distance metric based on how patients score on survey items. Distances were calculated using Euclidean distance. Patients who reported low scores on all MDASI-HN items but reported moderate/severe on any MDADI item were clustered. An examination of the MDADI scores for

these patients was undertaken to identify groups of questions that may be highly endorsed on MDADI even if patients have no symptoms captured by MDASI-HN.

2.6.3 Sensitivity Analysis

Question scores were sorted into none/mild and moderate/severe, then a sensitivity analysis was performed for each question pair between the two surveys. Patients who reported moderate/severe on both questions were considered true positives (TP), if moderate/severe was indicated on only the MDADI question, and none/mind on the MDASI-HN question, they were considered false negative (FN). Then sensitivity would be TP/(TP+FN). This was done to determine the relationship between highly endorsed MDADI questions and highly endorsed MDASI-HN questions. Specificity is not reported since the research interest did not include symptoms with little or no impact.

All statistical analysis was performed using Python 3.7.7 and machine learning models used scikit learn 0.23.1 [104].

2.7 Results

MDASI-HN and MDADI questions and their abbreviations used in the analysis are presented in the supplementary data in Table S 1 and

Table S 2 respectively. Patient demographics and disease characteristics are summarized in Table 2-1.

Table 2-1: Patient Demographics

Demographic	**No. (n=156)**	**(%)**
Age, mean (SD)	57.5 (10.9)	
Gender		
Male	132	(85)
Female	24	(15)
Initial BMI (SD)	28.2 (5.6)	
ECOG, Median (Range)	1 (1-3)	
Disease Characteristics		
Primary Tumor Location		
Larynx	7	(4.5)
Hypopharynx	3	(1.9)
Oral Cavity	3	(1.9)
Oropharynx	99	(63.5)
Nasal Cavity	7	(4.5)
Nasopharynx	26	(16.7)

Unknown	11	(7.0)
T-Stage		
T0-T2	72	(46.2)
T3-T4	73	(46.8)
Tx	11	(7.0)
N-Stage		
N0	24	(15.4)
N1	34	(21.8)
N2	83	(53.2)
N3	14	(9.0)
NX	1	(0.6)
Chemotherapy Agent		
Carboplatin	3	(1.9)
Cetuximab	13	(8.3)
Cisplatin (Cisplatinum)	128	(82.1)
None	12	(7.7)

The 20 most endorsed moderate/severe items are shown in Table 2-2. Taking longer to eat because of swallowing problems, as identified by the MDADI, was the most endorsed item between the two surveys, with 85 (55%) patients rating it as moderate/severe. Having dry mouth was the most endorsed MDASI-HN item, with 68 (44%) patients. Out of the top 5 items most endorsed by patients, four of them are from MDADI.

Table 2-2: Most Endorsed Questions

Question	**Number of Patients**	**(%)**
MDADI_LongerToEat	85	(54.5)
MDADI_NotSelfConscious	71	(45.5)
MDASI_DryMouth	68	(43.6)
MDADI_LimitFood	51	(32.7)
MDADI_WhyNotEat	49	(31.4)

MDASI_Taste	42	(26.9)
MDASI_Mucus	40	(25.6)
MDADI_Effort	39	(25)
MDADI_FreeToGoOut	38	(24.4)
MDASI_SwallowChew	38	(24.4)
MDADI_Activities	37	(23.7)
MDASI_Fatigue	34	(21.8)
MDADI_Upset	33	(21.2)
MDASI_Sleep	31	(19.9)
MDADI_Cooking	29	(18.6)
MDADI_LiquidsCough	28	(17.9)
MDASI_Activity	28	(17.9)
MDASI_Drowsy	25	(16)
MDASI_ChokeCough	25	(16)
MDADI_EndOfDay	25	(16)

The number of patients to indicate a moderate/severe response for each question is given as 'Number of Patients'. MDASI-HN items are presented in blue for ease of reading.

2.7.1 Correlations

Overall, correlation between individual MDASI-HN and MDADI questions were weak to moderate (mean 0.27, SD 0.12, min -0.12, max 0.64). Figure 2-1 shows a correlation matrix between selected MDASI-HN and MDADI questions. The chart is read by following the row and column corresponding to the question pair to be investigated. The portion below the diagonal of the matrix has scatter plots representing the raw scores of each question combination. For instance, a point is plotted for a patient who scores 10 on MDASI dry mouth and 5 on MDADI effort. To see if there is a non-linear relationship between questions, a second order regression is fitted to the scatter plot data. Along the diagonal are histograms showing the distribution of scores for each item, with a kernel density estimate line to visualize a predicted distribution. The correlation

strength is reported above the diagonal, where the size of the circles is proportional to the correlation strength, and the correlation strength is reported inside the circles. The p-value for the correlation is indicated in the bottom right corner of the scatter plots. The strongest correlations are seen within intra-survey items.

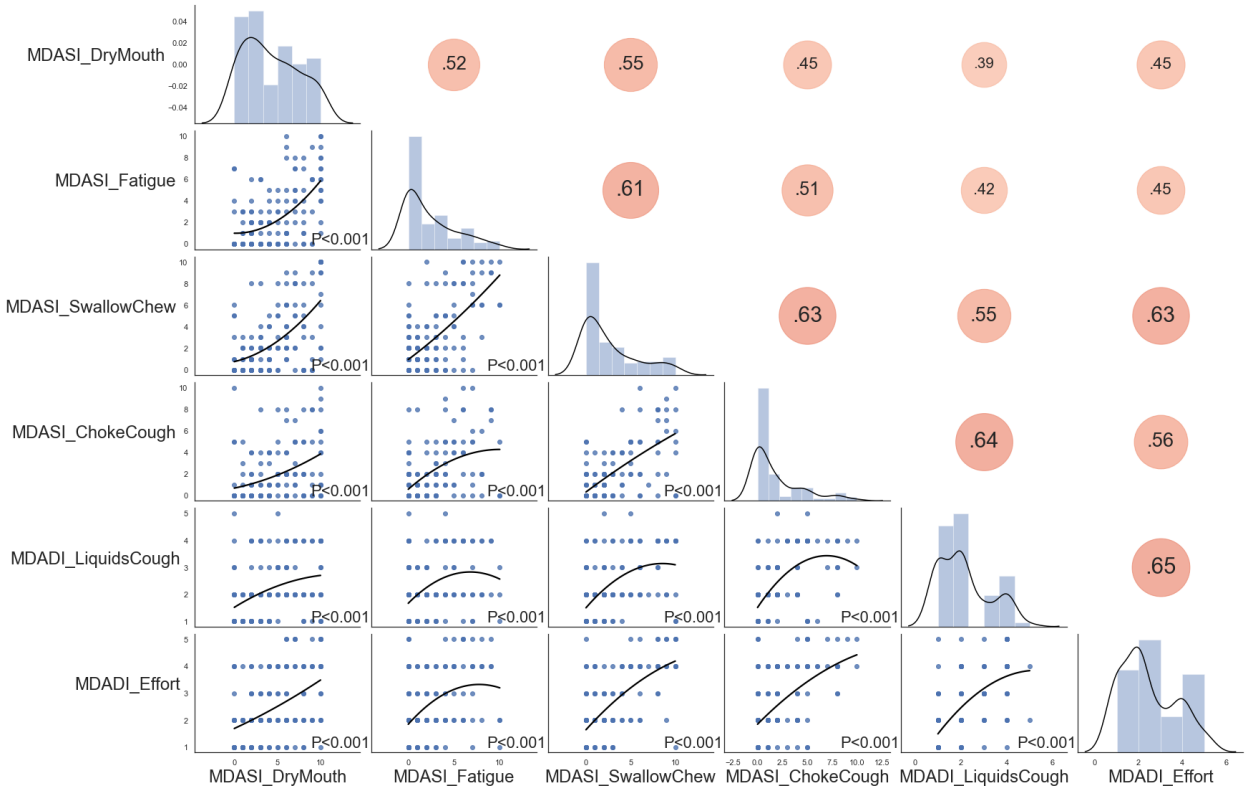

Figure 2-1: Correlation matrix of select questions. Question pairs between instruments with the highest correlation are shown. MDASI dry mouth is included because it is the most frequent question associated with high sensitivity values. All results were significant with Bonferroni multiple testing correction applied (P = 0.05/15 = 0.003). Scatter plots which show raw scores on each item are shown below the diagonal of the matrix with second order regression line. The score distribution along with a kernel density estimate are plotted along the diagonal. Above the diagonal contains correlation circles with the correlation strength printed inside. The size and intensity of the circles reflect the correlation strength

2.7.2 Clustering

Figure 2-2 shows hierarchical clustering for the subgroup of 51 patients (33% of the total) who did not indicate a moderate/severe response on any MDASI-HN question, but did indicate a moderate/severe response on at least one MDADI question. The most common MDADI question

receiving a moderate/severe response for this group of patients is, “I do not feel self-conscious when I eat,” with 33 (65% of the subgroup) moderate/severe responses. This is clustered near “I feel free to go out and eat with friends, neighbors, and relatives,” (box V in Figure 2-2) with 12 (24%) moderate/severe responses, in section V-2. The patients in box 2 generally reported none/mild on nearly all of the other questions across both surveys.

Other questions of consideration for this subgroup are (Figure 2-2, box W), MDADI longer to eat, with 23 (45%) moderate/severe responses, and MDADI why not eat, with 15 (29%). Most of the 14 group 3 patients reported none/mild symptoms apart from these two questions. While group 1 patients also frequently endorsed these two questions, they also reported a range of moderate/severe responses on MDADI items.

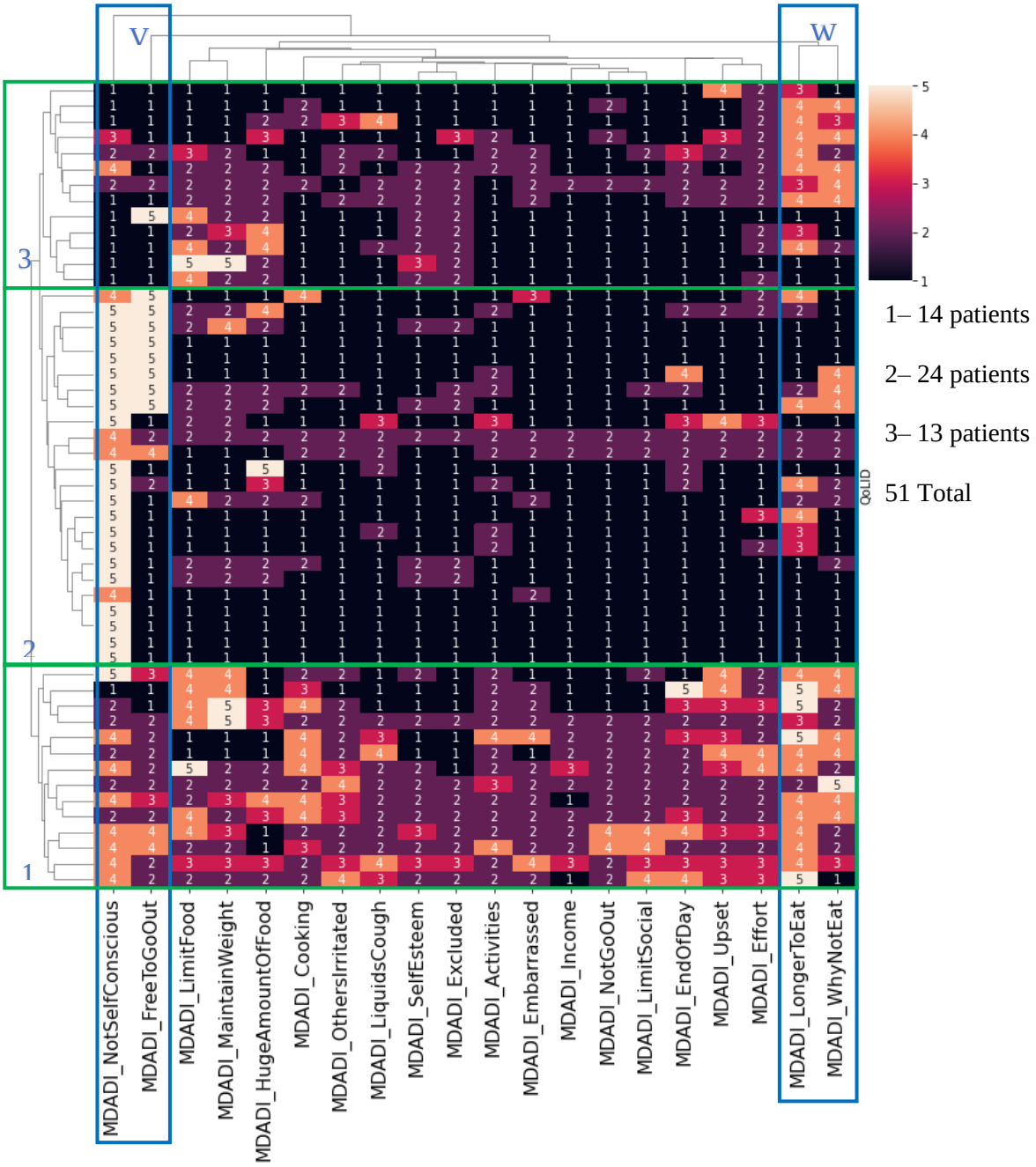

Figure 2-2: Questions missed by MDASI-HN. There were 51 (33%) patients who gave only none/mild responses on the entire MDASI-HN but gave moderate/severe responses on at least one question on MDADI; we show MDADI results for these patients. Raw scores are plotted with the clustered heatmap. Note: numbers superimposed on the figure indicate MDADI item scores. Euclidean distance was used as the distance metric, and Ward's method was used to calculate the linkages between clusters.

2.7.3 Sensitivity Analysis

Many patients who report moderate/severe symptoms for MDASI dry mouth also report moderate/severe on many of the MDADI items, the top 3 of the latter are; MDADI longer to eat

(30% of all patients), MDADI activities (20%), and MDADI limit food (19%). Table 2-3 shows MDADI items ranked by the number of moderate/severe responses, with corresponding MDASI-HN match. MDASI match is the MDASI-HN question with the highest number of moderate/severe responses out of the patients who indicated moderate/severe on the corresponding MDADI question. An example would be; out of the 85 patients who indicated moderate/severe on MDADI longer to eat, 47 of them also indicated moderate/severe on MDASI dry mouth. MDASI dry mouth would be the most endorsed MDASI-HN question for those patients.

Table 2-3: MDADI Items Ranked in Descending Order with MDASI-HN match

Number of Moderate/Severe Responses to MDADI Item	MDADI Item	MDASI Item Match	Frequency	Sensitivity
85	MDADI_LongerToEat	MDASI_DryMouth	47	0.55
71	MDADI_NotSelfConscious	MDASI_DryMouth	24	0.34
51	MDADI_LimitFood	MDASI_DryMouth	30	0.59
49	MDADI_WhyNotEat	MDASI_DryMouth	28	0.57
39	MDADI_Effort	MDASI_DryMouth	29	0.74
38	MDADI_FreeToGoOut	MDASI_DryMouth	21	0.55
37	MDADI_Activities	MDASI_DryMouth	31	0.84
33	MDADI_Upset	MDASI_DryMouth	23	0.7
29	MDADI_Cooking	MDASI_DryMouth	20	0.69
28	MDADI_LiquidsCough	MDASI_DryMouth	19	0.68
25	MDADI_EndOfDay	MDASI_DryMouth	19	0.76
24	MDADI_Embarrassed	MDASI_DryMouth	15	0.62
24	MDADI_MaintainWeight	MDASI_DryMouth	15	0.62
20	MDADI_HugeAmountOfFood	MDASI_DryMouth	11	0.55
19	MDADI_LimitSocial	MDASI_DryMouth	14	0.74
17	MDADI_NotGoOut	MDASI_DryMouth	12	0.71
12	MDADI_SelfEsteem	MDASI_DryMouth	9	0.75
9	MDADI_OthersIrritated	MDASI_DryMouth	6	0.67
7	MDADI_Excluded	MDASI_DryMouth	3	0.43
5	MDADI_Income	MDASI_Work	5	1

MDADI items are matched with the MDASI-HN item most endorsed by the same patients. Frequency is the number of patients to indicate moderate/severe on both questions. Sensitivity is equivalent to the frequency divided by the number of moderate/severe responses to MDADI item. Specificity is not reported since items with little or no impact are not considered for this research question.

Question combinations are ordered by sensitivity in Figure 2-3, showing the top 20 and their frequency of moderate/severe endorsement. Several question pairs had high sensitivity, but were not endorsed by many patients. The total number of patients to rate moderate/severe on each MDADI item can be calculated by dividing the frequency by the sensitivity. MDASI work is the most sensitive to MDADI income with a sensitivity of 1.0, although only 5 patients gave a moderate/severe response on MDADI income. This shows that each patient who experienced a drop of income due to their swallowing dysfunction, also experienced an impact on their work as reported on the MDASI-HN, although there were very few of these patients. Seven of the top 20 MDASI-HN questions are sensitive to MDADI income, but it is associated with a low frequency (≤5). MDASI dry mouth has the next highest sensitivity to MDADI actives with a sensitivity of 0.84 and is endorsed by 31 patients. These pairings do not indicate causation but help to show symptoms and QoL impacts occurring in parallel. It is helpful to understand the most important question groups by examining the sensitivity with the frequency.

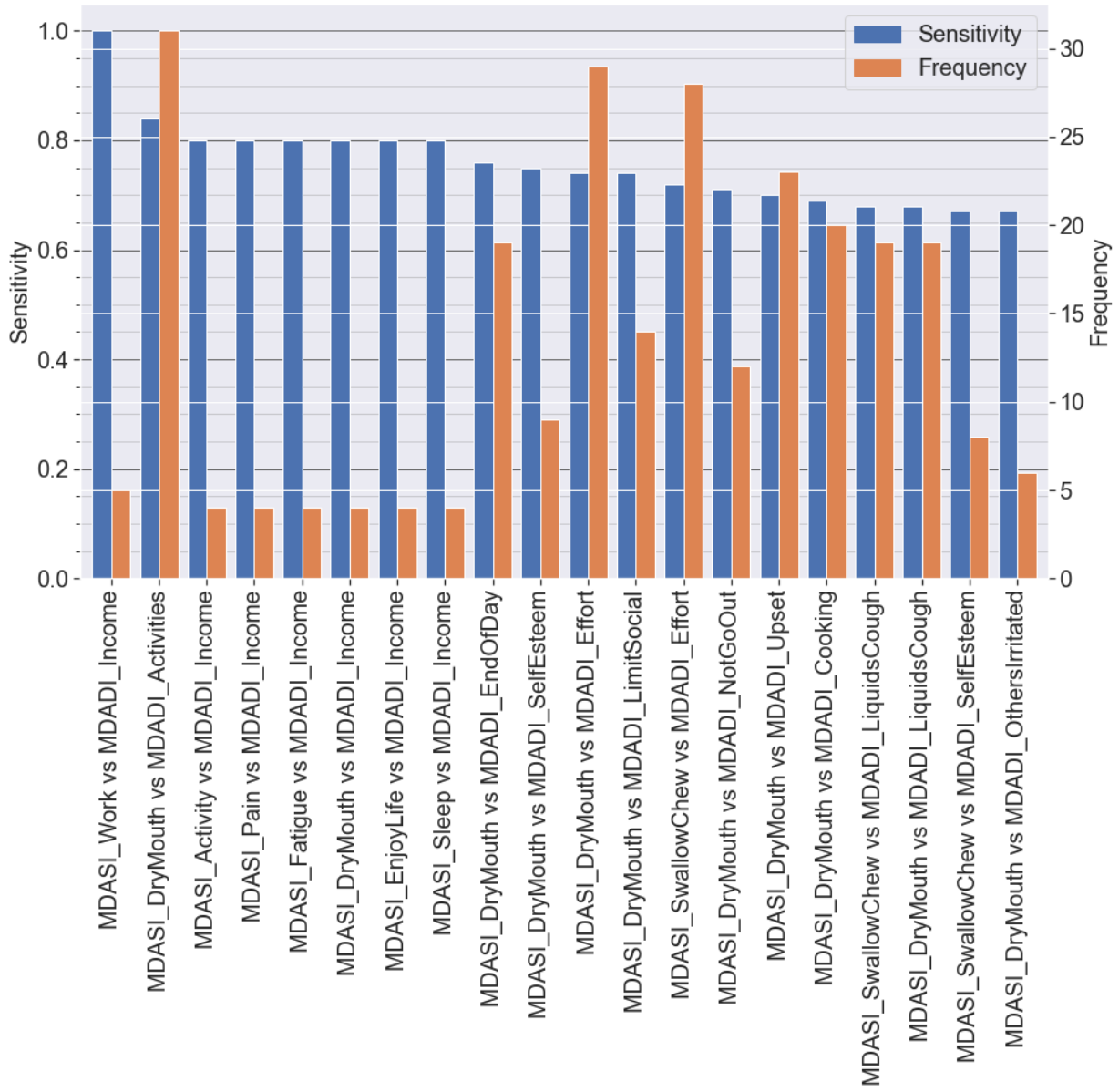

Figure 2-3: Sensitivity to MDASI-HN questions, and frequency chart for MDASI-HN and MDADI question pairs. Frequency is the number of patients to indicate moderate/severe on both questions. Sensitivity is equivalent to the frequency divided by the moderate/severe responses to MDADI item. Many of the items with the highest sensitivity

2.8 Discussion

This study compared all items between the MDADI and the MDASI-HN surveys to identify potential gaps in the MDASI-HN, when considering its use for routine clinical care. Correlations were weak between items across surveys, generally indicating that these two instruments had little

overlap in symptom assessment. We calculated sensitivity scores to detect the extent to which MDADI items are captured by the MDASI-HN. After ranking MDADI items by moderate/severe scores, we found MDASI dry mouth to be to be most sensitive to identifying MDADI items with sensitivity scores ranging from 0.34 to 0.84, the highest being MDASI dry mouth with MDADI activities (0.84). The highest sensitivity scores indicate that the MDASI-HN is sufficient at capturing how patients respond on the MDADI to the extent their daily activities have been impacted, how swallowing is more difficult by the end of the day, how their swallowing difficulties impact their self-esteem, how swallowing takes greater effort, and how their swallowing difficulties limit their social life.

Within our study population, we found approximately 33% of patients reported no moderate/severe MDASI-HN symptoms but at least one moderate/severe MDADI symptom, which would be a significant number of patients in a large population. Although this number would be impacted by our chosen cut-point values in our analysis, and the differences in reporting formats between the MDADI and MDASI-HN measures (5 and 10-point scales, respectively), we have taken a pragmatic and clinically relevant approach to the comparison between them.

Our results do indicate some gaps in the MDASI-HN symptom survey when used on its own, as seen with some of the highest endorsed MDADI items. The MDADI item ‘takes longer to eat’ had the highest percentage of patients (55%, 85) report a moderate-severe symptom score compared to all other items, with 55% (47) of those also endorsing MDASI dry mouth. Taking longer to eat was also the second most missed by MDASI-HN (15%, 23), suggesting it is taking patients longer to eat due to symptoms not captured on the MDASI-HN. Gaps could also be identified in other

highly endorsed MDADI items including: others asking ‘why can’t you eat this?’ and limiting food intake. As the MDASI-HN is not sensitive to some of these highly endorsed MDADI items, it may benefit physicians to refer to the above MDADI items during clinical assessments to aid physicians in referring or providing more tailored symptom support. For example, patients who regularly feel they need to limit their food intake may benefit from more direct support from a dietician or speech language pathologist.

Interestingly, the most common MDADI items that received moderate/severe scores were ‘I do not feel self-conscious when I eat,’ and ‘I feel free to go out.’ These two particular items differ from all other items as higher scores reflect higher functioning/more positive responses, and this was corrected for in our analysis by taking into account the reversal. These MDADI items may reflect some patient misinterpretation in reporting as the response scale was the reverse of previous questions, and reflected inconsistent reports from patients in our results. There were 13 patients who gave moderate/severe responses on at least one of these two questions, and none/mild on every other question across both surveys. If we assumed these patients misinterpreted these questions, and removed them from the subgroup, there would be 38 (24%) patients left who were missed by MDASI-HN. We cannot know for sure if questions were misinterpreted, without asking each individual. Otherwise, for patients who did not misinterpret these items, their responses may indicate they feel self-conscious when eating, and do not feel comfortable going out despite experiencing no symptoms as recorded on the MDASI-HN, indicating the potential for other symptoms not captured by either survey. Although these particular findings may reflect important distinctions from a research perspective, overall, these questions are less actionable in a clinical sense and may not influence how a physician would manage the patient.

No studies to our knowledge have directly compared all MDASI-HN and MDADI items. The MD Anderson Working Group has recently published a study looking at a single-item comparison of the MDASI-HN swallowing chewing item to the MDADI questionnaire in regards to correlations to quality of life [26]. The authors found that the MDASI-HN swallowing chewing item is an efficient way for physicians to obtain dysphagia experiences from patients. The authors suggest only patients who report a score of ≥ 6 could additionally complete the MDADI to provide a more detailed account of their dysphagia. While our analysis was based on a value to separate none/mild from moderate/severe of 5, reflecting the analysis during development of the MDASI-HN [61], our sensitivity analysis indicated that MDASI swallowing chewing had a mild-moderate sensitivity to three MDADI items, (MDADI effort, MDADI liquids cough, and MDADI self-esteem) (0.67-0.72). Generally, our study found the MDASI dry mouth captured most highly endorsed MDADI items with mild-moderate sensitivity, and was more sensitive to MDADI items compared to MDASI swallowing chewing.

Our results are interesting in that the MDASI dry mouth item demonstrated the highest sensitivity scores on highly endorsed MDADI questions compared to other MDASI-HN items, notably MDASI swallowing chewing and MDASI choking/coughing which would be expected to have higher sensitivity. Dysphagia generally includes issues with swallowing, such as choking and coughing and xerostomia is associated with dysphagia, but is distinct. Xerostomia and dysphagia are closely related so it can be difficult to resolve their differences without specific investigations such as measurement of salivary flow rates and sialochemistry. Our results may indicate patients experience worse quality of life due to dry mouth, than to dysphagia. If using only the MDASI-

HN, it may be useful to establish a symptom score on the MDASI-HN dry mouth item, such as ≥ 2, that prompts physicians to ask more detailed questions or refer to additional selective MDADI items. If the threshold for flagging patients as having a moderate/severe response on the MDASI dry mouth item is reduced to 2, the previous 51 patients missed by the MDASI-HN drops down to 15. This could be an efficient way to flag patients who could have been potentially missed while maintaining the question burden to just the MDASI-HN.

There are some limitations of the study that warrant discussion. The MDASI-HN was only comparted to one other survey that was symptom-specific. Further, the nature of the items is different between the two surveys, such that the MDASI-HN survey is cross-cutting and includes common cancer symptoms as well as HNC specific symptoms, while the MDADI survey is specific to how patients experience dysphagia. The cross-sectional study design also reflects a limitation by only capturing patients who are finished treatment. All patients included in the study were in follow-up which would limit the generalizability of these results to patients on active treatment. Although scoring of the MDASI-HN is done by grouping patients into none, mild, moderate, severe, the categories were dichotomized into none/mild, and moderate/severe to focus on clinical relevance and to align with the method of reporting during the development of the MDASI-HN [61]. Additionally, although our sample size was sufficient to complete a sensitivity analysis, a larger sample size could help to further clarify the relationship between the MDASI-HN and MDADI items.

2.9 Conclusions

We conducted a study to compare the MDASI-HN and MDADI quality of life instruments. There is interest from physicians to know to what extent symptom burden is captured solely by the MDASI-HN. Inter instrument correlation was generally weak suggesting low overlap between the two; however, some questions such as MDASI dry mouth, were highly endorsed and sensitive to most MDADI questions, suggesting this single item is associated with many dysphagia related symptoms. A significant proportion of patients reported mild symptoms on the MDASI-HN but reported at least one severe symptom on the MDADI. The MDADI questions that captured the highest number of patients missed by the MDASI-HN are: longer to eat, why not eat, and limit food. These questions could serve to augment the MDASI-HN in a clinical setting, and to get a more complete sense of symptom burden.

CHAPTER 3

DOSIOMIC ANALYSIS ON BREAST RADIOTHERAPY PLANS

3.1 General Introduction

Dosiomics is a new field of research in which features representing patterns in the 3D dose distribution for radiotherapy cancer treatments are extracted. This area of study borrows tools from the more developed area of radiomics where features are extracted from imaging data such as CT or MRI. Patterns in the dose distribution were investigated to understand if they can be used to stratify patients based on irradiated volume for whole breast radiotherapy. It is generally understood that large breasted patients are more difficult to plan and have worse treatment outcomes. This study investigates the use of dosiomics and its relationship to the treated volume of whole breast radiotherapy plans and is a proof of concept for future studies which may be used to predict treatment outcomes.

The work presented in this chapter contains a study on whole-breast radiotherapy plans. I was the primary author, and contributing authors were Dr. Qiao Sun, and Dr. Wendy Smith. I collected all the data, performed the feature extraction and analysis, and wrote the manuscript. Dr Smith conceived of the project and provided guidance, valuable suggestions, and edits to the manuscript. Dr Sun provided guidance on the statistical and machine learning methods, proofread and made suggestions to the manuscript, and provided feedback on the analysis.

3.2 Abstract

Background: Large breast size is associated with poor radiotherapy (RT) outcomes as a result of hot spots, and standard dose volume histogram (DVH) measures are poor indicators of future toxicity. There is little consensus in the literature as to how breast size should be measured and its relationship to hot spots. This study shows how dosiomics can be performed in breast RT to stratify patients based on breast size and is a proof of concept on how to prepare datasets for future investigations.

Methods: A set of 631 whole-breast 3D conformal radiotherapy (3DCRT) plans were collected. The volume to receive 50% prescribed dose or greater was used to categorize breasts into small, medium, and large. Dosiomic features were extracted, and feature reduction was achieved by hierarchically clustering then selecting the highest variance feature from each cluster. A random forest classifier was trained to predict size from the reduced feature set, and permutation feature importance identified the most predictive features. DVH features such as the volume to receive 107% and 105% of prescribed dose (V107, V105) and dose to the hottest 2cc (D2cc) were compared with the dosiomic features in terms of predictive capacity.

Results: 1462 dosiomic features were extracted and reduced to 50. Prediction accuracy for the top 3 dosiomic features, and DVH features was 93% and 36% respectively.

Conclusions: Dosiomic features greatly outperform DVH features to stratify breast volume. Dosiomics gives greater information about the 3D distribution of hot spots and could be used in future studies to predict treatment outcomes.

3.3 Introduction

Studies have linked radiation therapy treatment outcomes with breast size [19][41][105][43]. They show that larger breast size increases planning difficulty, increases dose inhomogeneity, and produces hot spots, causing negative patient outcomes or toxicity after treatment[41][106]. This study investigates elements of the dose distribution that can stratify breast volume using dosiomics and compares them to standard dose volume histogram (DVH) measures. The relationship of these metrics to patient outcomes such as cosmesis, could then be tested in future work. Many studies have attempted to classify breast size, and several metrics have been used such as cup size, separation distance, normal distance from the chest wall to the nipple and fat thickness, clinical target volume, planning target volume, and visual inspection of photographs [43][45][19][40][106][41]. Issues arise because there is ambiguity in the extent of the breast tissue, which typically extends into the armpit and is not taken into consideration in many of the proposed breast size or volume measurement methods.

Before treatment planning, oncologists or radiation therapists typically palpate in the armpit to determine the extent of the breast tissue and mark it with a wire which is visible in planning imaging. In this study we propose to use the irradiated volume, or the 50% isodose from the treatment plan to investigate the volume of breast tissue [31][86][108]. Because tangential beams extend from the center of the chest wall or the edge of the medial portion of the breast to the lateral portion of the breast palpated by the oncologists, this should give a good indication of actual breast tissue volume. The treated volume does not include only breast tissues, often a small portion of the chest wall, lung, and even heart is present in 3D conformal radiation therapy (3DCRT) plans, but since we are ultimately interested in the dose distribution throughout the tissues, these areas

are included [105]. There are treatment planning methods available to increase the conformity of the dose to the breast tissue such as intensity modulated radiotherapy (IMRT), but such methods are beyond the scope of the study which focuses on tangential beam 3DCRT including field-in-field, with beam energies ranging from 6-15MV.

In many studies 3D dose data is mapped into a 2D DVH to make it easier for humans to interpret. Three features were used from the DVH in this study: volume to receive 107% of the prescribed dose (V107), volume to receive 105% of the prescribed dose (V105), and dose to the hottest 2cc of tissue (D2cc). With the advancement of machine learning techniques (ML), researchers are increasingly using ML to identify patterns in high dimensional data that was previously intractable for humans. Dosiomic and DVH features were investigated as a means to stratify breast radiotherapy treatments on treated volume, with the intent of gaining an understanding as to how patterns in the dose cloud relate to treatment volume. A deeper understanding of this relationship can direct further study of patient outcomes, leading to the ability to inform treatment decisions in the era of personalized medicine.

3.4 Materials and Methods

3.4.1 Patients in Study, and Data Collection

A set of 631 whole-breast 3DCRT plans treated at the Tom Baker Cancer Centre between June 2015 to November 2019 was collected. The plans each contained two tangential fields: medial anterior, and opposing lateral posterior, and included plans with wide tangents. Plans with supraclavicular nodes irradiated were not included. Table 3-1 summarizes the patient group and

shows the number of patients whose chest wall was treated, the number of patients with breast implants, and whose internal mammary lymph node chain (IMC) were treated.

Table 3-1: Patient Characteristics

Characteristic	**No. (n=631)**	**(%)**
Sex		
Female	627	(99)
Male	4	(0.6)
Treated Side		
Left	522	(83)
Right	109	(17)
Chest Wall Treated	241	(38)
Implant Treated	49	(8)
IMC Nodes Treated	379	(60)
Prescribed Dose (Gy)		
40 - 46	566	(90)
46 - 52.5	29	(5)
Other	36	(5)

The irradiated volume was represented using the 50% isodose [83][87]. The data were split into small (S), medium (M), and large (L) volumes using the median value, plus and minus one standard deviation as the bin edges. The S bin was considered having a volume less than 1,273 cc (n=78), M being between 1,273 and 2,266 cc (n=466), and L volumes being over 2,266 cc (n=85). The bin edge separating the M and L volume was varied by ± 10% of the volume to assess feature stability.

Classification performance was measured using balanced accuracy, which is the mean of sensitivity and specificity. The area under the receiver operating characteristic curve (ROC AUC) was not used since it is misleading for imbalanced data sets [109]. Balanced accuracy is one of the best methods for an imbalanced data set [110] and is the main metric used in this study.

In order to extract dosiomic features, a mask is required which designates the region of interest (ROI) of each slice of the dose array. This process is illustrated in Figure 3-1. The 3D dose cloud is calculated from the radiotherapy plan on the patient's CT scan. The region to receive 50% of the prescribed dose is then used to generate a mask. The mask contains binary values, with ones designating the ROI and zeros elsewhere. The mask is then applied over top of the dose cloud and features are extracted. The DICOM files were read into a script using pydicom v2.1.2 [111].

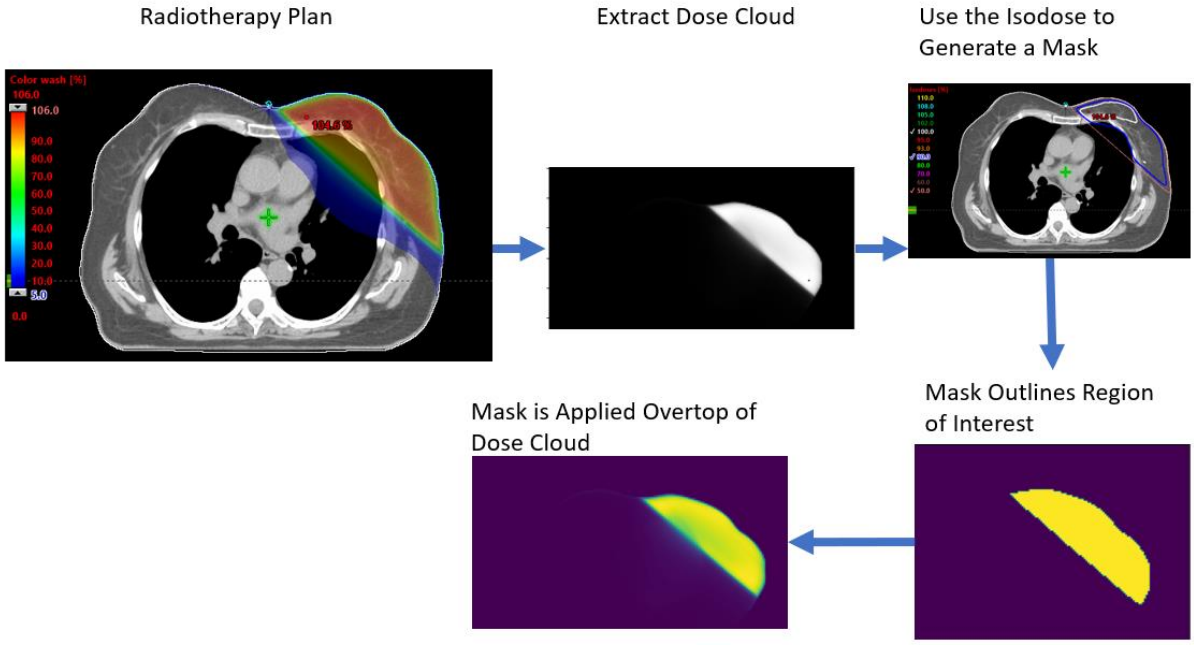

Figure 3-1: Mask Extraction from Isodose. The first image shows an axial CT slice through the patient and a colored dose wash representing the radiation dose. The radiotherapy plan is constructed on the CT scan and the dose cloud is calculated from the plan. The 50% isodose profile which represents the region of tissue to receive 50% of the prescribed radiation dose, otherwise known as the irradiated volume, is used to create a binary mask containing 1 in the ROI and 0 elsewhere. The mask is overlaid onto the dose cloud so dosiomic features can be extracted from the ROI.

3.4.2 Dosiomic Feature Extraction

Dosiomic feature extraction was performed using pyradiomics v3.0.1 [66], 2D features are extracted from the slices and 3D features are extracted from the combined stack. Before feature extraction, the voxels are resampled to an isotropic 3 mm through interpolation using pyradiomics [66]. Isotropic resampling is required for the calculation of texture features so the voxel spacing is rotationally invariant, and to allow for comparison between batches [112]. The range of Gy level intensities are discretized into bins for image preprocessing to reduce computation to a tractable level [65][112]. For this study, a fixed bin width of 1 Gy was used. The advantage of using a fixed bin width is that the direct relationship with the Gy values is maintained [112].

Several classes of features can be extracted from each dose set as described in the Imaging Biomarker Standardization Initiative (IBSI) documentation [112]. Classes of features extracted in this study as defined by IBSI include:

- First Order Statistics: Common metrics which describe the distribution of intensities throughout the ROI. Examples would include: mean, median, minimum, and energy. Where first order total energy adds all voxel values together in the volume.
- Grey Level Cooccurrence Matrix (GLCM): Calculates patterns of voxel values which occur together. The number of times two intensity values occur within a distance from a central voxel in any given direction are used to build the matrix. Several statistics can be calculated such as difference variance, and difference entropy.
- Grey Level Run Length Matrix (GLRLM): Constructed by counting the length in number of consecutive voxels with a given intensity. Features such as short run emphasis, grey level non-uniformity (GLN), and grey level variance are calculated from this matrix.

- Grey Level Size Zone Matrix (GLSZM): Looks at the number of connected voxels that share the same intensity. Some of the features calculated from this matrix include: size-zone non-uniformity, small area emphasis, and zone percentage.
- Grey Level Dependence Matrix (GLDM): Constructed by counting the number of voxels of a given intensity which are connected to a center voxel of the same intensity. Some of these features include: small dependence emphasis, large dependence emphasis, and dependence variance.

Features can also be extracted from filtered images. Filters are applied to test if the different texture features are better able to quantify the characteristics of the ROI [112]. Texture optimization could have significant influence on the prognostic capability of features [113]. Several filters were used in this study: wavelet, square, square root, logarithm, exponential, and gradient. A Laplacian transformation of Gaussian (LoG) filtered images with sigma values of [3.0, 4.0, 5.0] was also applied. Sigma < 3.0 is not included since the images were resampled to a 3mm voxel size. A total of 1462 features were extracted from each patient. A detailed description of the features can be found in the pyradiomics documentation [66] and the IBSI documentation [112].

Features that are direct surrogates of volume such as total energy, surface area, or mesh volume, do not enhance the understanding of the dose distribution in larger breasts. For most of the analysis below, these features are excluded. Focus is paid to patterns in the voxel distribution which may serve as other indicators of treatment volume.

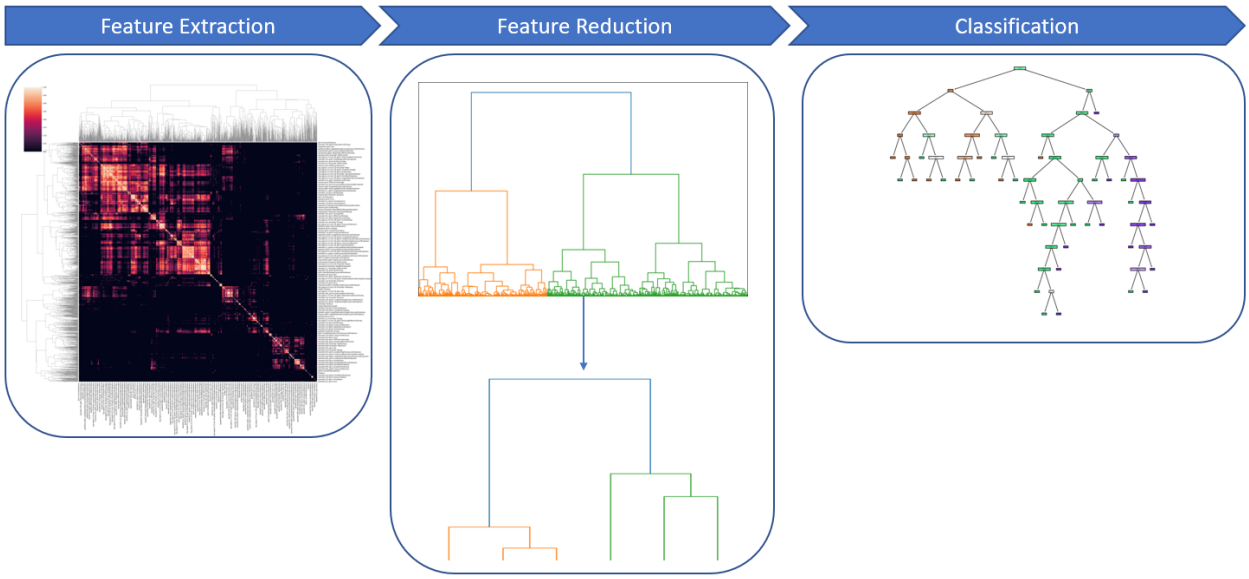

Figure 3-2: Analysis process schematic. Dosiomic features are extracted from the data, resulting in a large number of correlated features as shown in the clustered heat map. In order to reduce the feature space, the absolute value of the compliment of the correlation matrix was hierarchically clustered and the feature with the highest variance from each cluster was selected. Classification is performed using the selected features in a random forest which is an ensemble of decision trees.

3.4.3 Feature Reduction and Analysis

Feature reduction was conducted by first hierarchically clustering the feature set then selecting a single feature from each cluster, which were then used in a random forest classifier with permutation feature importance to indicate the most predictive features for treatment volume category. The general workflow for the analysis process is shown in Figure 3-2. A more detailed description of the model parameters and data preparation is shown in Figure 3-3. Shape features (n=17), and total energy features (n=17) which could be used as a direct surrogate for treated volume did not contribute to the understanding of the relationship between dose distribution and treated volume categories, and so were excluded in the rest of the analysis, leaving 1428 remaining features.

3.4.4 Hierarchical Clustering

For this study, the dosiomic features were correlated using the Spearman rank correlation. To convert the correlation matrix into a distance matrix for hierarchical clustering, the complement of the absolute value of the correlation was used. This allows highly correlated features to be grouped by providing a low distance value, then a single feature is selected from each cluster. This maintains variability and reduces redundancy [114]. Chormunge and Sudarson used a similar method to select a non-redundant feature from each cluster on microarray and text datasets [79]. To select an item from a cluster, the features were scaled to 0-1 using min-max scaling, then the feature with the highest variance in the cluster was selected.

The number of features selected from hierarchical clustering was chosen by plotting the number of features as a function of the balanced accuracy. First, one cluster is used and the accuracy is tested, then two clusters are used, three, and so on, until the entire feature set is used for classification. Balanced accuracy increases quickly as features are added. The point at which the accuracy curve begins to level off is where the algorithm is potentially using the extra features to overfit [114]. The number of features selected at this point are then used in the remainder of the study. Scikit-learn 0.23.1 was used for the machine learning models [104].

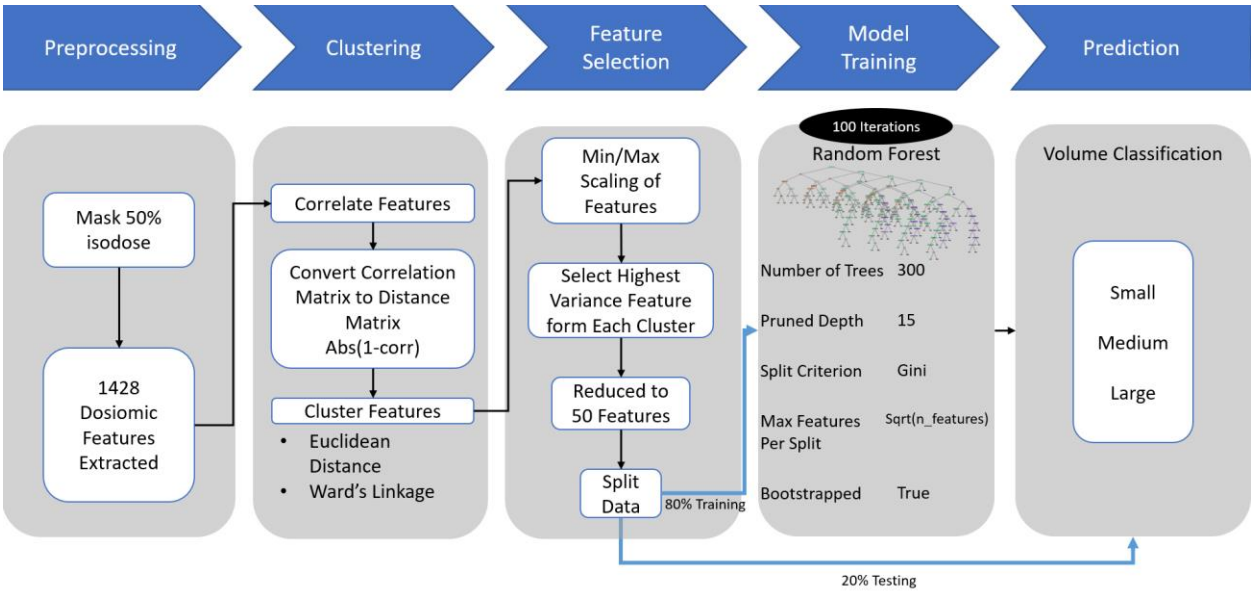

Figure 3-3:Flow chart depicting the data preparation process up to classification.

3.4.5 Permutation Feature Importance

Permutation feature importance was used to assesses the predictive capacity of features in the model [87][88]. It is defined as the decrease in the model score (balanced accuracy for this study) when a single feature is randomly shuffled [88]. The process is repeated 50 times for each feature, resulting in a range of values for the decrease in balanced accuracy. This study uses random forests as a classifier which has its own measure of feature importance based on model impurity [88]. Impurity-based importance is strongly biased and favours high cardinality features [115]. It should be noted that feature importance from fully developed forests (trees with a single observation in each leaf) is inaccurate and should not be used [87]. The trees in this study are pruned to reduce this error as outlined in the section below. Permutation importance can be misleading in datasets with strongly correlated features. If two features are highly correlated and one is permuted, the model can still access the feature through its correlate, resulting in lower importance for both even

if they are actually important [115]. A method to compensate for this is to use hierarchical clustering and only keep one feature from each cluster [114], and is the method used in this study.

3.4.6 Predicting Treatment Volume Class

The trees in this study were pruned to a maximum depth of 15, with 300 trees used in each forest. The minimum samples used per split was set at 2. Gini impurity was used as the splitting criteria and bootstrapped samples were used when building trees. The dataset was randomly split, with 20% used for testing and the 80% for training. Initial model parameters were selected by producing a matrix containing ranges of possible parameter values which contained over 6480 combinations. A subset of 200 models was selected through random search, and trained using 3-fold cross validation. The best performing model was then used as a template from which another matrix was constructed containing 270 models with ranges of parameters centered around the best performing parameters from the random search. Every model was tested, and the best parameters were selected. Volume classification predictions (S, M, L) were performed using different groups of features: group1 contained only the dosiomic features, group 2 contained dosiomic plus dose-volume (V105, V107, D2CC), group 3 contained only the dose-volume features. Since the main goal of the study was feature assessment, hyperparameter tuning was not conducted for every single analysis. The pursuit of slight gains in accuracy after a base model was determined were outside of the goal of the study.

3.5 Results

3.5.1 Feature Reduction

During hierarchical clustering the number of clusters was varied from 1 to 1428 representing the full feature set, and a single feature selected from each cluster. Balanced accuracy score was

plotted as a function of the number of features as shown in Figure 3-4. Due to the random nature of the random forest, the accuracy can vary depending on the random seed used to select features for the first split as well as the selection of the test data. For that reason, the measurement was repeated 15 times with varying random seeds and test data, and the full range of the variation in recorded accuracy is shown as a shaded region in Figure 3-4. The accuracy of the model continues to climb until it begins to level off at about 50 features with a mean balanced accuracy of 83%. After this point the model accuracy slowly continues to climb as the model uses the extra features to overfit. There is another jump at about 150 features, but we are interested in the smallest number of features that give the steepest gain in accuracy. The set of 50 features is used for the remainder of the analysis.

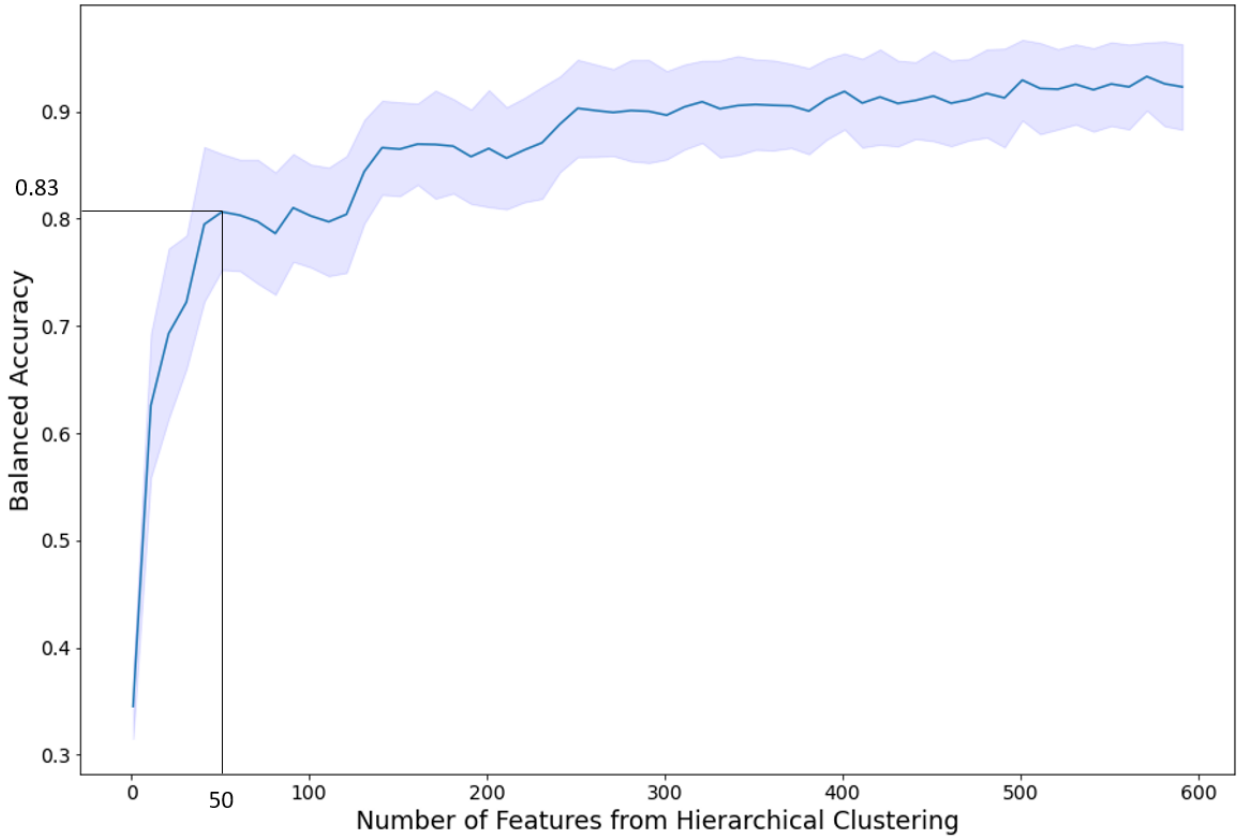

Figure 3-4: Balance accuracy score as a function of the number of features selected from hierarchical clustering. The measurement was repeated 15 times with varying random seeds for each random forest, as well as varying the selection of the test data. The variation in the measurements is shown as a shaded region, with the average as the solid line. Accuracy increases steeply until about 50 features where it begins to level off.

3.5.2 Feature Importance

The top features according to permutation feature importance are ranked from highest to lowest in a box plot in Figure 3-5. Feature definitions can be found in the IBSI documentation [112]. The top feature was wavelet-LLL GLRLM run length non-uniformity.[112]. If this single feature is used to train the model, a mean balanced accuracy score of 67% is achieved on a test set. The next feature is log-sigma-4.0mm 3D first order Median, which when used to train the model achieves a mean accuracy of 63%, the next feature is square GLDM Large Dependence High Gray Level emphasis and is able to achieve a mean accuracy of 48% when used alone. Reducing the feature set allows the model to focus on the most informative features and it no longer suffers from the curse of dimensionality. When only the top 3 features are used, mean balanced accuracy of 93%

is achieved, which is higher compared to using all 50 features. A summary of these results is seen in Table 3-2 where the mean balanced accuracy is given plus/minus the standard deviation of running the test through 100 iterations of varying random states. Features with zero importance have no effect on the model accuracy when they are removed and features with negative importance increase the accuracy of the model when removed. As can be seen in Figure 3-5 most features have low importance.

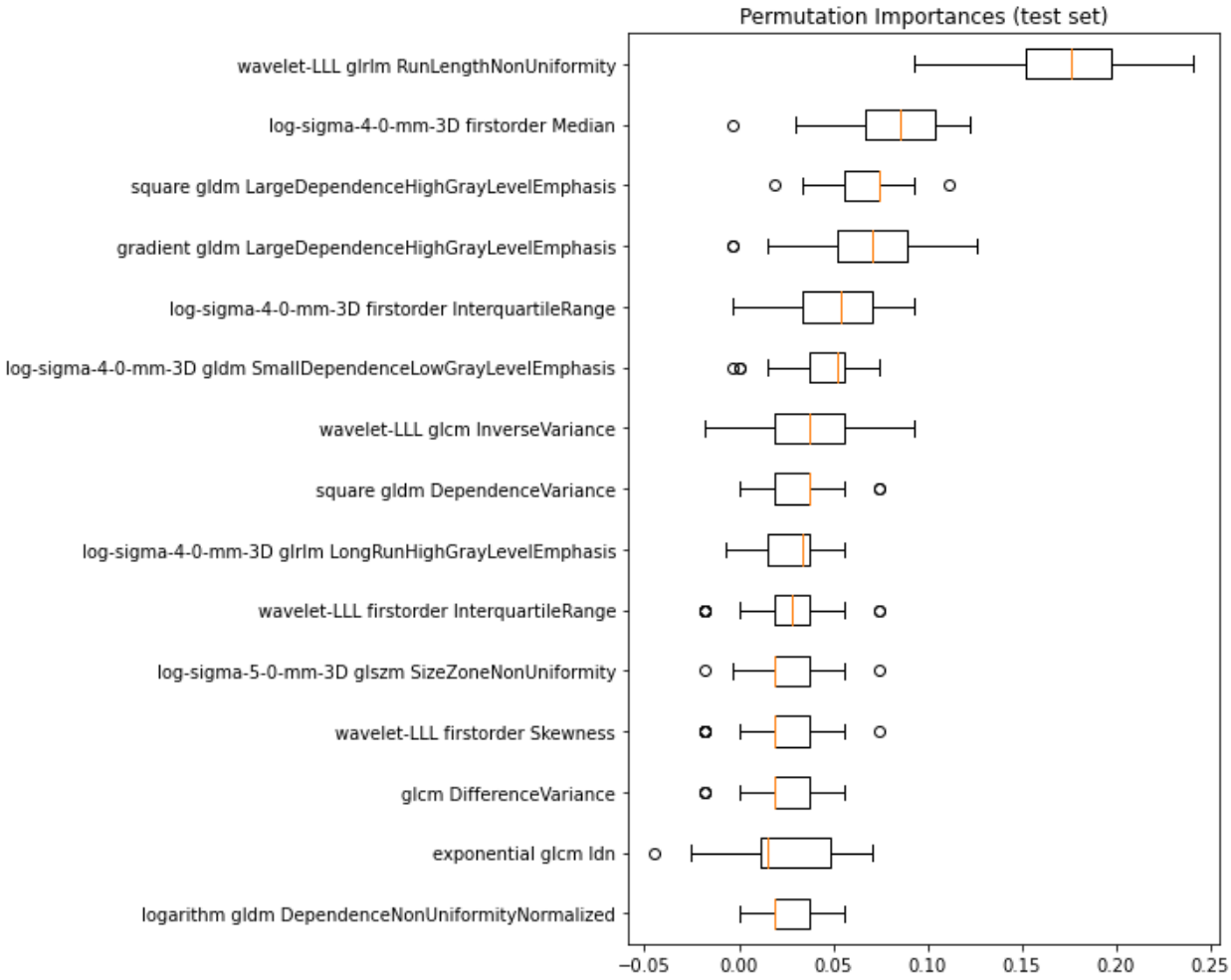

Figure 3-5: Permutation Feature Importance on the test set. The permutations were repeated for 100 iterations with varying random states, giving a range in importance values for each feature. The model consistently favoured wavelet-LLL GLRLM run length non-uniformity with much lower importance given to the other features.

3.5.3 Dose Volume Histogram Features

Including the DVH features (V015, V107, D2cc) in the feature set had a negative impact on prediction accuracy. When using both the 50 features from clustering, as well as the three DVH features, the mean balanced accuracy decreased from 84% to 80%. If using the DVH features alone, the mean balanced accuracy decreased to 36%. These values with standard deviations are recorded in Table 3-2.

Table 3-2: Feature Combinations to Classify Treatment Volume Category

Training Feature	**Balanced Accuracy (%)**
50 Features from Clustering	84 ± 3
50 Features + DVH (V107, V105, D2cc)	80 ± 3
DVH (V107, V105, D2cc)	36 ± 3
Wavelet-LLL GLRLM Run Length Non-Uniformity	67 ± 3
Log-Sigma 4.0mm 3D First Order Median	63 ± 5
Square GLDM Large Dependence High Gray Level Emphasis	48 ± 3
Top 3 Features	93 ± 1

Different sets of feature combinations and their accuracy when used to predict treatment volume category are shown in the table above. Each test was generated 100 times, giving a mean balanced accuracy and range. The mean and standard deviation of the balanced accuracy scores are reported for each combination of features. Including DVH parameters in the feature set decreased accuracy compared to using all 50 features (80% and 84% respectively) and using them alone resulted in low accuracy of 36%. Using the top 3 features from permutation feature performance resulted in the highest accuracy of 93%

3.5.4 Top Features

A distribution of the top three features' values relative to the volume category is shown in Figure 3-6. The S category generally shows lower values and a narrower range then the other two categories. The M category for each feature shows a large range across most of the values

overlapping with both L and S, with the L category having a slightly smaller range, but higher maximum values.

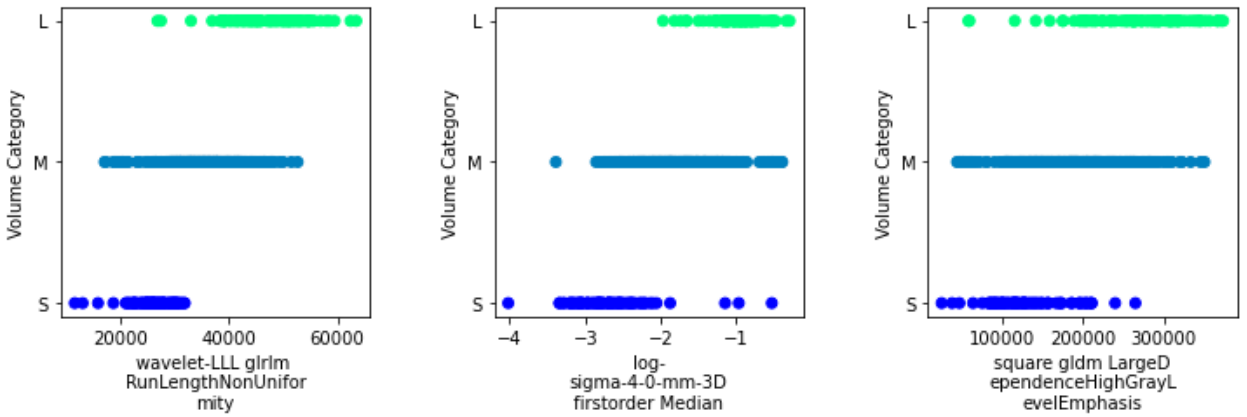

Figure 3-6: Distribution of the top 3 Features vs Volume Category. The values of the features are displayed for each volume category. There is much overlap in values between categories with a general trend of smaller values for small breasts and large values for larger breasts and medium breasts see a wide range of values throughout.

The first plot in Figure 3-6 for wavelet-LLL GLRLM run length non-uniformity suggests that the uniformity of the dose distribution decreases as the size of the treated volume increases, although M treatment volumes display varying degrees of uniformity both high and low. A similar observation can be made about the second plot in Figure 3-6 by looking at Log-Sigma 4.0mm 3D First Order Median. Median intensities are generally higher for large breasts, but there is a wide range across the sizes. Square GLDM Large Dependence High Gray Level Emphasis in Figure 3-6 shows a large degree of overlap between the volumes. Suggesting that each volume category can see a wide variation in the size and uniformity of the high dose region.

3.5.5 Variation of Treatment Volume Bin Sizes

When binning the treatment volumes into S, M, and L, the edge for the L bin was varied by ±10% as a test for sensitivity. The test was run over 100 iterations with varying random states giving a

mean and standard deviation in the accuracy which can be seen in Table 3-3. Changes in the accuracy and the order of the important features were observed. The first value tested was the median treatment volume plus one standard deviation giving greater than or equal to 2266cc as the edge for the L volume. At this level, the top 3 features are able to predict the L treatment volume with mean accuracy of 80%. Increasing the volume cut-off by 10% (2493cc) shifted accuracy to 79%. Decreasing the volume cut-off by 10% (2039cc) shifted it to 90%. The top two features switched for the smaller volume cut-off as show in Table 3-3.

Table 3-3: Results of Varying the Large Treatment Volume Bin Edge

	Large Bin Edge Value		
	2266 (cc)	**2493 (cc)**	**2039 (cc)**
Accuracy to predict L volume (mean %) ± std	80 ± 9	79 ± 11	90 ± 6
Top 3 Features	• Log-Sigma 4.0mm 3D First Order Median • Wavelet-LLL GLRLM Run Length Non-Uniformity • Square GLDM Large Dependence High Gray Level Emphasis	• Log-Sigma 4.0mm 3D First Order Median • Wavelet-LLL GLRLM Run Length Non-Uniformity • Square GLDM Large Dependence High Gray Level Emphasis	• Wavelet-LLL GLRLM Run Length Non-Uniformity • Log Sigma 3mm 3D First Order Mean • Square GLDM Large Dependence High Gray Level Emphasis

3.6 Discussion

The conventional measures of hot spots in the dose cloud through DVH parameters (V105, V107, D2cc) were not predictive of treatment volume. Using DVH features alone resulted in a very low prediction accuracy of 36%. This opposes the view that hot spots measured by the DVH could be used to stratify patients based on treated volume, or breast size. This could be because limiting these values within a certain threshold is part of the treatment planning objectives (such as <10cc receives 107% of the prescribed dose). This could involve increasing dose to healthy tissue, or accepting lower dose elsewhere. The conscious effort to reduce these hot spots reduces the relationship with treated volume and one would expect that applying these limits will impact the dose in other ways, which this study attempts to identify.

Wavelet GLRLM run length non-uniformity was the most predictive feature of treated volume with 67% accuracy. A grey level run is a set of consecutive pixels or voxels having the same grey level value. The length is the number of points in the run. This feature assesses the distribution of runs over the Gy values, has low values when the runs are equally distributed along the Gy levels. By examining Figure 3-6 it can be seen that, given a value between 2,750 and 4,000, there would be nearly an equal probability of guessing whether a treatment volume fell into the M or L categories. If larger treatment volumes simply had longer run lengths due to having more voxels, or a dose uniformity that was inversely related with volume, separation between volumes would be more distinct with higher run length values; however, this was not the observed behaviour. Although median values differed between volume categories, a wide variation in the uniformity of the run lengths was seen in each volume. Making it difficult to discern breast size based on this feature alone.

To increase prediction accuracy almost 30%, wavelet GLRLM run length non-uniformity can be combined with one or two other measures. Log-sigma 3D first order median describes the median voxel intensity in the ROI after a Laplacian of Gaussian (LoG) filter is applied. From Figure 3-6 it can be seen this value is generally higher for larger volumes, but M volumes see a wide distribution of this value. GLDM counts the number of voxels with neighbours of the same intensity. GLDM large dependence high grey level emphasis, emphasizes voxels that have high intensities with high numbers of neighbours with the same values. The random forest model was able to use these 3 features to achieve higher accuracy than using either a single feature, or all features. In contrast, simply using hotspot values from the DVH were not predictive. Inhomogeneity generally seems to increase with breast size, but medium breasts can see a wide range in the uniformity of the dose distribution and a single measure of uniformity was insufficient to classify breast size.

There were many lower importance features which were sensitive to changes in the analysis parameters. Use of these features is difficult from a reproducibility standpoint. Slight changes in the bin edges for classifying the volume categories caused the order of these less important features to change; however, use of the GLRLM grey level non-uniformity, plus an additional two of these less important features resulted in higher prediction accuracy than using the entire feature set. This also had the benefit of reducing the complexity of the model and increasing interpretability. Future studies could combine these results with treatment outcomes and investigate their relationship. The utility of these studies would be the identification of prognostic dosiomic features to help guide treatment planning to optimize outcomes.

Because the voxel intensities used in dosiomic feature extraction are related to radiation units of Gy, dosiomic feature extraction is more robust when compared with radiomics. Radiomics relies on grey level voxel intensities that are arbitrary and subject to change based on image reconstruction and processing parameters while dose intensities are related to physical units. Dosiomics can still suffer from variation due to parameters such as how the voxels are resampled, their size, or many other possible extraction parameters. Great improvements in the reproducibility of dosiomic studies can be achieved by standardizing extraction parameters. Also, a robust method of feature reduction should be applied. With the large number of features that can be extracted compared to the size of most data sets, many studies can suffer because of the curse of dimensionality, where the data becomes sparse due to too many features. The important features are sensitive to how feature reduction was achieved, and which type of prediction is desired. No single set of dosiomic features will be important for all types of outcomes prediction. This is an advantage in terms of flexibility in predicting outcomes, but it is a disadvantage in terms of study reproducibility. Dosiomics could become a powerful tool in personalized medicine, but much work needs to be done to standardize the method.

3.7 Conclusion

This study demonstrated that dosiomic features can be used as input into a random forest model to stratify patients into three breast-size pools. Permutation feature importance showed that features calculated from the GLRLM, mean dose, and the GLCM are most strongly associated with treated volume and are able to achieve a prediction accuracy of about 93%. Features that directly relate to volume such as surface area, mesh volume, first order energy, etc. were not included in the

analysis. Instead, the focus was to identify patterns contained in the dose distribution that relate to breast size. This study also showed that conventional measures of hot spots from the DVH (V105, 107, D2cc) were ineffectual when used as input features in a random forest classifier to predict breast size, only achieving about 36% accuracy.

Feature reduction is an important step in dosiomic analysis. Data sets from which dosiomic features are extracted are likely to be highly dimensional, having few samples in relation to the number of features that can be extracted and suffer from the curse of dimensionality. This study showed a method of feature reduction with hierarchical clustering and selecting the feature with the highest variance after min-max scaling, to select uncorrelated features for use in classification. The balanced accuracy as a function of the number of features was used as an indication of when the model begins to over-fit. The number of features was reduced from 1428 to 50 using this method.

There is a wealth of information hidden in the distribution of the radiotherapy dose cloud. With the use of modern analysis methods such as dosiomics and machine learning, we can uncover patterns in information that was historically intractable. This study aimed to test the application of dosiomics to whole-breast radiotherapy plans. Further research studies which can link the results of these type of analyzes with treatment outcomes could help to optimize cancer treatments and help to usher in the era of personalized medicine. All these techniques are digital, and in the future will be able to be automated, providing clinical professionals with tools to support their data-driven decision-making process, and free up resources for more human-centered care. Dosiomics is a relatively new area of research which combines the study of radiomic features with the calculated

radiation dose distribution, considering the complicated relationship between radiation and tissue. A thorough understanding of the relationship between these topics will no doubt help to improve cancer treatments.

CHAPTER 4

CONCLUSIONS

4.1 Contributions and Summary of Results

4.1.1 Patient Quality of Life

Patient reported outcome surveys, and radiotherapy dose clouds are complex and informative datasets containing a wealth of valuable information. Currently, the standard analysis used to investigate these data sets are limited summaries of the information therein to inform clinical decisions. With the advancement of ML and data science, these tools can be used to gain deeper insights and value can be extracted from this data. This thesis investigated how data science and ML can be applied to patient reported outcome surveys and radiotherapy dose clouds to gain new clinical insights that reach beyond what was once obtainable from standard analysis.

Patient reported outcome instruments are important tools that physicians can use to understand the impact treatments are having on patients. Not only is it important to treat a disease but having a high QoL after treatment is also important. If a patient's cancer goes into remission, but they can't eat or swallow, then it is questionable whether the treatment should have been conducted. Patient reported outcomes not only consider the physical effects of treatment, but a patient's interpretation of those effects and how it impacts their life. Aspects such as the ability to function in society, and being able to go for a meal with friends and family are incredibly important parts of life. If these aspects are inhibited, it can lead to high rates of depression or other complications. In order to capture these effects post treatment, clinicians administer PRO instruments to identify areas of concern that can be addressed. It is also important that the questionnaire burden remains low to

not require a great deal of time or effort for the patient. If the survey is too long, patients will be reluctant to complete them, and the integrity of the answers could be compromised as they begin to rush through them.

We examined two QoL instruments, the M.D. Anderson symptom inventory head and neck module (MDASI-HN), and the M.D. Anderson dysphagia inventory (MDADI). The MDASI-HN measures the severity of symptoms related to head and neck cancer treatments. The MDADI is a self-reported QoL questionnaire designed to measure impact on patient's ability to function post treatment. It does not measure the level of dysphagia a patient experiences but the impact on activities such as going out to eat, and a patient's social life. These two instruments were compared after having a cohort of 156 patients complete them, post-radiotherapy.

Our study identified a group of patients who experienced no symptoms as reported by the MDASI-HN but reported some level of severe symptoms on the and MDADI. This suggests that their ability to function is being affected through factors other than symptoms recorded on the MDASI-HN. These patients were hierarchically clustered and patterns in their responses were discovered. Most of these patients reported taking longer to eat and being asked 'why can't you eat that'. Through sensitivity/specificity analysis it was determined that the closest symptom causing this QoL impact was having dry mouth. Dry mouth is caused by xerostomia which is characterized by inhibited saliva production. Xerostomia and dysphagia (difficulty swallowing) are closely related but they are different ailments. Xerostomia can cause dysphagia, but difficulty swallowing can also be caused by other factors. Since these patients did not rate their xerostomia as moderate to severe on the MDASI-HN, they may have been overlooked in terms of requiring referral to specialists if they

had completed only the MDASI-HN. The question scale relative to the number of patients missed was examined for the MDASI dry mouth item. It was determined that if clinicians flagged patients who rated dry mouth as ≥2 as having a moderate/severe response, most of the patients missed by MDASI-HN would be captured. Although the MDASI-HN may miss a few swallowing/eating items, the intervention is the same, which is referral to a nutritionist or speech/language pathologist. Reducing the moderate/severe cut-off for this item would allow clinicians to reduce the questionnaire burden and administer only the MDASI-HN with the confidence that most patients would be flagged if they were experiencing detrimental symptoms impacting their quality of life.

Overall, the MDASI-HN is an effective patient reported outcome instrument, but some patients were missed from our cohort. A possible method of capturing these missed patients would be to include a small number of items from the MDADI. Including MDADI items; Taking longer to eat, being asked why can't you eat this, and limiting food intake would help to identify patients who are having difficulty, despite reporting no symptoms on the MDASI-HN. Another method to capture more patients would be to modify the moderate/severe threshold value down to two on the MDASI-HN dry mouth item. This would be an efficient way of capturing more patients, without having to add extra survey items. The study also showed that patients having dry mouth experienced the greatest impact on their quality of life.

4.1.2 Dosiomics

Dosiomics involves the extraction of statistical features from radiotherapy dose distributions. It expands on radiomics which extracts features from medical imaging data such as computed tomography (CT), magnetic resonance imaging (MRI), mammograms, and x-rays. All these modalities generate a great deal of data in the form of pixels and voxels. These imaging data can be comprised of slices along a central axis which are then combined into a 3D volume. The features can be shape features such as volume, maximum and minimum diameters, as well as higher order features which describe the distribution of voxel intensities throughout the region of interest. The relationship between dosiomic features and breast volume in whole breast radiotherapy cancer treatments was examined. A literature survey has shown numerous challenges to treatment planning associated with large breast size, including increased planning time, less dose homogeneity through the breast tissue, higher skin dose, increased dry and wet desquamation, and poor cosmesis. The assumption is that inhomogeneity in the dose causes hot spots in certain regions of the breast, resulting in toxicities. The relationship between the dose distribution and breast size was examined using dosiomics.

A data set of 631 whole breast radiotherapy treatment plans, planned with 3D conformal radiation therapy (3DCRT) was collected from the Tom Baker Cancer Center. Voxels to receive 50% or greater of the prescribed dose were used to represent the treated volume. Other studies have attempted to classify breast volume based on many different metrics; however, there is no consensus in the literature as to what metric to use to measure breast volume. The breast tissue extends into the armpit and cannot be discerned by bra cup size, visual inspection through photographs, or many other of the attempted methods. In preparation for breast radiotherapy, the

breast tissue is palpated by physicians or therapist and typically marked with a wire which shows visible on the CT scan which is used to align the edge of the treatment beam. This irradiated volume of tissue is a good representation of breast volume; however, a small region of chest wall can be included in the dose distribution. Since this research is interested in how the dose distribution changes with breast size, this volume is included as it is part of the treated volume.

Patients were separated into small (S), medium (M), and large (L) treatment volumes. A total of 1462 dosiomic features were extracted from the 3D dose cloud in this analysis. They were hierarchically clustered, and a single feature was selected from each cluster to reduce correlation and redundancy. The number of features to use in the analysis was determined by sequentially increasing the size of the feature set used in a random forest classifier. The number of features used to train the model was varied from one until the entire set was used, and the prediction accuracy was plotted as a function of the number of features. At first the prediction accuracy increased sharply as the number of features were increased. The point at which the accuracy curve begins to level off is when the model is potentially using the extra features to overfit. At this point the number of features was cut off at 50, and the classification accuracy was used as a benchmark for comparison with further feature reduction.

A random forest classifier was used to predict breast volume from the extracted features. Permutation feature importance determine the predictive capability of each feature. Single features were permuted in the set and the drop in performance was measured. Features that caused the greatest drop in performance when permuted had the highest importance. Wavelet-LLL GLRLM run length non-uniformity, was the most important feature and represents the distribution of run

lengths of given voxel intensities across the dose cloud. The analysis showed that there was some relationship between uneven distribution of runs of voxel intensity to increased treated volume size; however, each volume category saw a wide range of values. This relationship was compared to conventional measures of dose hot spots such as the volume receiving 105% of the prescribed dose (v105), the volume receiving 107% of the prescribed dose (v107), and the dose to the hottest 2cc of tissue (d2cc), calculated from the dose volume histogram (DVH). It was discovered that conventional hotspot calculations were insufficient and stratifying treated volumes based on size and were only able to achieve a prediction accuracy of 36%. A group of three dosiomic features was able to achieve a superior prediction accuracy of 93%, even surpassing the accuracy of using all 1462 features (83%). Concurrent feature reduction, and an increase in prediction accuracy was achieved. It was discovered that a single dosiomic feature was never as accurate as when using a combination. A single dosiomic feature may be able to measure the existence of inhomogeneity, similar to conventional DVH measures, but combining more features revealed how the inhomogeneity occurs in such a way as to allow machine learning algorithms to be able to make better predictions on the volume category.

Utilizing the power of computation and combining several features in a machine learning model shows how these tools are able to perform at a higher level and take into account intricacies that are difficult for humans on their own. The large amount of information about the dose structure that can be obtained with dosiomics can help find relationships between the dose and clinically relevant outcomes. Dosiomics will be a powerful tool to analyze dose distributions and radiotherapy treatment outcomes, which will help researchers and clinicians move past the limitations of DVH analysis.

CHAPTER 5

FUTURE WORK

Patient reported outcome measures are valuable tools with both clinical and research applications. An interesting area of research would be to combine PRO instruments with dosiomic analysis. The irradiated treatment volume could be used to relate to not only measurable clinical outcomes, but patient reported outcomes. Dosiomic features could be extracted from the dose cloud and machine learning algorithms could be used to discover patterns between irradiated tissue and patient experience. This could show if there are aspects of the dose that have a greater impact on a patient's quality of life, not only regarding the success of their treatment, but how the patient interprets their QoL afterwards. This could be valuable in influencing treatment decisions in finding a balance between disease control and patient comfort. Particularly for patients that could survive for many years post treatment, as they will have to live with the side effects long-term.

The study in chapter 3 of this thesis helps to inform further research that could utilize treatment outcomes with the dosiomic features for whole breast radiotherapy. Previous studies have examined radiomic features, and the patterns present in imaging data in relation to treatment outcomes, but very few studies have involved radiotherapy dose distribution data, and no studies up until this point have been conducted for breast radiotherapy plans. The 3D dose distribution considers the complex interactions of radiation with the molecules of tissue. Features of this data can be used to examine the relationship between the dose cloud and treatment outcomes. Dosiomics, unlike radiomics, is related to a physical quantity. While a range in the values of

Hounsfield units in a CT scan can be associated with a particular tissue, the radiation units of Gy represent the amount of energy in joules per kilogram delivered to the tissue, and could have greater relevance to treatment outcomes, but there is still difficulty in the reproducibility of features based on how voxels are resampled and other extraction parameters. Finding a way to standardize these parameters will help to greatly improve reproducibility and advance the study of the radiotherapy dose cloud. Once the physical nature of important dosiomic features can be understood, this knowledge could pave the way to optimizing treatment plans for outcome and provide decision support tools for physicians and benefiting patients with higher quality data driven treatment plans.

Bibliography

[1] S. Canada, “Canadian Cancer Statistics,” 2019.

[2] Government of Canada, “Cancer-specific stats 2020,” 2020.

[3] D. R. Brenner *et al.*, “Projected estimates of cancer in Canada in 2020,” vol. 192, no. 9, pp. 199–205, 2020.

[4] D. Hanahan and R. A. Weinberg, “Review Hallmarks of Cancer : The Next Generation,” *Cell*, vol. 144, no. 5, pp. 646–674, 2011.

[5] J. E. Talmadge and I. J. Fidler, “AACR centennial series: the biology of cancer metastasis: historical perspective.,” *Cancer Res.*, vol. 70, no. 14, pp. 5649–69, Jul. 2010.

[6] Canadian Cancer Society, “Cancer Treatment,” 2021. [Online]. Available: https://www.cancer.ca/en/cancer-information/diagnosis-and-treatment/treatment/?region=ab. [Accessed: 27-Aug-2021].

[7] E. Vlashi, W. H. McBride, and F. Pajonk, “Radiation responses of cancer stem cells,” *J. Cell. Biochem.*, vol. 108, no. 2, pp. 339–342, 2009.

[8] National Cancer Institute, “Brachytherapy to Treat Cancer,” *About Cancer*, 2019. [Online]. Available: https://www.cancer.gov/about-cancer/treatment/types/radiation-therapy/brachytherapy. [Accessed: 27-Aug-2021].

[9] E. Basch *et al.*, “Symptom Monitoring With Patient-Reported Outcomes During Routine Cancer Treatment: A Randomized Controlled Trial,” *J. Clin. Oncol.*, vol. 34, no. 6, pp. 557–565, Feb. 2016.

[10] M. A. Bakitas *et al.*, “Early versus delayed initiation of concurrent palliative oncology care: Patient outcomes in the ENABLE III randomized controlled trial,” *J. Clin. Oncol.*, vol. 33, no. 13, pp. 1438–1445, 2015.

[11] W. C. Röntgen, “On a New Kind of Rays,” *Nature*, vol. 53, no. 1369, pp. 274–276, Jan. 1896.

[12] G. Miller, *X-Rays and Radium in the Treatment of Diseases of the Skin*. Philidelphia, New York: Lea and Febinger, 1921.

[13] V. Despeignes, “Observation concernant un cas de cancer de l’estomac traite par les rayons Rontgen,” *Lyon Med*, vol. 82, pp. 428–430, 1896.

[14] L. Redniss, *Radioactive: Marie & Pierre Curie. A tale of love and fallout*. New York: Harper Collins, 2010.

[15] F. M. Khan, “Physics of Radiation Therapy Third Edition,” *J. Am. Med. Assoc.*, p. 1138, 2003.

[16] C. Roy, “1000 Curie Cobalt-60 Units for Radiation Therapy,” *Nature*, vol. 12, no. 2, pp. 1035–1036, 1951.

[17] Y. Xiao *et al.*, “Flattening filter-free accelerators : a report from the AAPM Therapy Emerging Technology Assessment Work Group,” vol. 16, no. 3, pp. 12–29, 2015.

[18] D. Herbert *et al.*, “Methodological issues in radiation dose–volume outcome analyses:,” pp. 2109–2127, 2002.

[19] B. Méry *et al.*, “Correlation between anthropometric parameters and acute skin toxicity in breast cancer radiotherapy patients: A pilot assessment study,” *Br. J. Radiol.*, vol. 88, no. 1055, pp. 1–6, 2015.

[20] E. Haciislamoglu *et al.*, “Dosimetric comparison of left-sided whole-breast irradiation with

3DCRT, forward-planned IMRT, inverse-planned IMRT, helical tomotherapy, and volumetric arc therapy," *Phys. Medica*, vol. 31, no. 4, pp. 360–367, 2015.

[21] S. H. Lee *et al.*, "Multi-view radiomics and dosiomics analysis with machine learning for predicting acute-phase weight loss in lung cancer patients treated with radiotherapy," *Phys. Med. Biol.*, vol. 65, no. 19, p. 195015, Sep. 2020.

[22] C. C. Society, "Canadian Cancer Society's Advisory Committee on Cancer Statistics. Canadian Cancer Statistics 2016," *Can. Cancer Soc.*, pp. 1–132, 2016.

[23] W. H. Westra, "The changing face of head and neck cancer in the 21st century: The impact of hpv on the epidemiology and pathology of oral cancer," *Head Neck Pathol.*, vol. 3, no. 1, pp. 78–81, 2009.

[24] A. . El-Naggar, "Cellular and Molecular Pathology of Head and Neck Tumors.," in *Head and Neck Cancer Multimodality Management*, J. Bernier, Ed. Springer, Newy York, NY, 2011, pp. 57–79.

[25] M. A. Tyler *et al.*, "Long-term quality of life after definitive treatment of sinonasal and nasopharyngeal malignancies," *Laryngoscope*, vol. 130, no. 1, pp. 86–93, 2020.

[26] S. Grant *et al.*, "Single-item discrimination of quality-of-life–altering dysphagia among 714 long-term oropharyngeal cancer survivors: Comparison of patient-reported outcome measures of swallowing," *Cancer*, vol. 125, no. 10, pp. 1654–1664, 2019.

[27] P. Dirix, S. Nuyts, and W. Van Den Bogaert, "Radiation-induced xerostomia in patients with head and neck cancer: A literature review," *Cancer*, vol. 107, no. 11, pp. 2525–2534, 2006.

[28] A. A. Forastiere, R. S. Weber, and A. Trotti, "Organ preservation for advanced larynx cancer: Issues and outcomes," *J. Clin. Oncol.*, vol. 33, no. 29, pp. 3262–3268, 2015.

[29] D. J. Adelstein *et al.*, "An intergroup phase III comparison of standard radiation therapy and two schedules of concurrent chemoradiotherapy in patients with unresectable squamous cell head and neck cancer," *J. Clin. Oncol.*, vol. 21, no. 1, pp. 92–98, 2003.

[30] V. Budach *et al.*, "Hyperfractionated accelerated chemoradiation with concurrent fluorouracil-mitomycin is more effective than dose-escalated hyperfractionated accelerated radiation therapy alone in locally advanced head and neck cancer: Final results of the Radiotherapy Coo," *J. Clin. Oncol.*, vol. 23, no. 6, pp. 1125–1135, 2005.

[31] B. Lacas *et al.*, "Role of radiotherapy fractionation in head and neck cancers (MARCH): an updated meta-analysis," *Lancet Oncol*, vol. 18, no. 9, pp. 1221–1237, 2017.

[32] B. Fisher *et al.*, "Lumpectomy Compared with Lumpectomy and Radiation Therapy for the Treatment of Intraductal Breast Cancer," *N. Engl. J. Med.*, vol. 328, no. 22, pp. 1581–1586, 1993.

[33] M. E. Taylor *et al.*, "Factors influencing cosmetic results after conservation therapy for breast cancer," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 31, no. 4, pp. 753–764, 1995.

[34] J. R. Owen *et al.*, "Eff ect of radiotherapy fraction size on tumour control in patients with early-stage breast cancer after local tumour excision : long-term results of a randomised trial," pp. 467–471, 1999.

[35] J. A. Julian *et al.*, "Long-Term Results of Hypofractionated Radiation Therapy for Breast Cancer," pp. 513–520, 2010.

[36] J. S. Haviland *et al.*, "The UK Standardisation of Breast Radiotherapy (START) trials of radiotherapy hypofractionation for treatment of early breast cancer : 10-year follow-up results of two randomised controlled trials," *Lancet Oncol.*, vol. 14, no. 11, pp. 1086–1094, 2013.

[37] B. D. Smith *et al.*, “Radiation therapy for the whole breast: Executive summary of an American Society for Radiation Oncology (ASTRO) evidence-based guideline,” *Pract. Radiat. Oncol.*, vol. 8, no. 3, pp. 145–152, 2018.
[38] D. Press, “Hypofractionated whole breast radiotherapy : current perspectives,” pp. 363–370, 2015.
[39] I. Ratosa, A. Jenko, and I. Oblak, “Breast size impact on adjuvant radiotherapy adverse effects and dose parameters in treatment planning,” *Radiol. Oncol.*, vol. 52, no. 3, pp. 233–244, 2018.
[40] S. De Langhe *et al.*, “Factors modifying the risk for developing acute skin toxicity after whole-breast intensity modulated radiotherapy,” *BMC Cancer*, vol. 14, no. 1, pp. 1–9, 2014.
[41] A. M. Moody *et al.*, “The influence of breast size on late radiation effects and association with radiotherapy dose inhomogeneity,” *Med. Dosim.*, vol. 20, no. 1, p. 68, 1995.
[42] N. R. Gustafson, T. Burrier, B. Butler, A. Hunzeker, N. Lenards, and L. Culp, “Correlation of hot spot to breast separation in patients treated with postlumpectomy tangent 3D-CRT using field-in-field technique and mixed photon energies,” *Med. Dosim.*, vol. 45, no. 2, pp. 134–139, 2020.
[43] P. Ciammella *et al.*, “Toxicity and cosmetic outcome of hypofractionated whole-breast radiotherapy: Predictive clinical and dosimetric factors,” *Radiat. Oncol.*, vol. 9, no. 1, pp. 1–10, 2014.
[44] M. Avanzo *et al.*, “Electron Density and Biologically Effective Dose (BED) Radiomics-Based Machine Learning Models to Predict Late Radiation-Induced Subcutaneous Fibrosis,” *Front. Oncol.*, vol. 10, no. April, pp. 1–9, 2020.
[45] R. Hannan *et al.*, “Hypofractionated whole-breast radiation therapy: does breast size matter?,” *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 84, no. 4, pp. 894–901, Nov. 2012.
[46] H.-G. Menzel, “The International Commission on Radiation Units and Measurements,” *J. ICRU*, vol. 10, no. 1, p. NP.2-NP, Apr. 2010.
[47] R. F. Mould, *A Century of X-Rays and Radioactivity in Medicine*. London: Institute of Physics Publishing, 1993.
[48] I. M. C. Columbia University, “Radiation Induced Dysphagia.” [Online]. Available: https://www.entcolumbia.org/health-library/radiation-induced-dysphagia. [Accessed: 08-Aug-2021].
[49] I. Fragkandrea *et al.*, “Radiation induced pneumonitis following whole breast radiotherapy treatment in early breast cancer patients treated with breast conserving surgery: A single institution study,” *Hippokratia*, vol. 17, no. 3, pp. 233–238, 2013.
[50] G. Chaushu *et al.*, “Salivary flow and its relation with oral symptoms in terminally III patients,” *Cancer*, vol. 88, no. 5, pp. 984–987, 2000.
[51] Mayo Clinic, “Dysphagia,” 2019. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/dysphagia/symptoms-causes/syc-20372028. [Accessed: 07-Apr-2021].
[52] A. Y. Chen *et al.*, “The development and validation of a dysphagia-specific quality-of-life questionnaire for patients with head and neck cancer: The M. D. Anderson Dysphagia Inventory,” *Arch. Otolaryngol. - Head Neck Surg.*, vol. 127, no. 7, pp. 870–876, 2001.
[53] S. C. Darby *et al.*, “Radiation-Related Heart Disease: Current Knowledge and Future Prospects,” *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 76, no. 3, pp. 656–665, 2010.
[54] I. Meattini, M. Guenzi, A. Fozza, and C. Vidali, “Overview on cardiac , pulmonary and cutaneous toxicity in patients treated with adjuvant radiotherapy for breast cancer,” *Breast*

Cancer, vol. 24, no. 1, pp. 52–62, 2017.
[55] M. Koc, P. Polat, and S. Suma, "Effects of tamoxifen on pulmonary fibrosis after cobalt-60 radiotherapy in breast cancer patients," vol. 64, pp. 171–175, 2002.
[56] C. Xiao *et al.*, "Quality of Life and Performance Status From a Substudy Conducted Within a Prospective Phase 3 Randomized Trial of Concurrent Standard Radiation Versus Accelerated Radiation Plus Cisplatin for Locally Advanced Head and Neck Carcinoma: NRG Oncology RTOG 012," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 97, no. 4, pp. 667–677, 2017.
[57] "Patient Reported Outcomes: Understanding what happens to patients who are treated with radiotherapy," *Canadian Partnership for Quality Radiotherapy*, 2018. [Online]. Available: http://www.cpqr.ca/programs/patient-reported-outcomes/. [Accessed: 10-Sep-2020].
[58] C. J., O. L., and H. S.J., "A systematic review of the impact of routine collection of patient reported outcome measures on patients, providers and health organisations in an oncologic setting," *BMC Health Serv. Res.*, vol. 13, p. 211, 2013.
[59] A. M. Association, "Overall Survival Results of a Trial Assessing Patient-Reported Outcomes for Symptom Monitoring During Routine Cancer Treatment," vol. 318, no. 2, pp. 197–198, 2017.
[60] L. Y. Yang, D. S. Manhas, A. F. Howard, and R. A. Olson, "Patient-reported outcome use in oncology: a systematic review of the impact on patient-clinician communication," *Support. Care Cancer*, vol. 26, no. 1, pp. 41–60, 2018.
[61] D. I. Rosenthal *et al.*, "Measuring head and neck cancer symptom burden: the development and validation of the M. D. Anderson symptom inventory, head and neck module.," *Head Neck*, vol. 29, no. 10, pp. 923–31, Oct. 2007.
[62] C. S. Cleeland *et al.*, "Assessing symptom distress in cancer patients: The M. D. Anderson Symptom Inventory," *Cancer*, vol. 89, no. 7, pp. 1634–1646, 2000.
[63] B. L. T. Ramaekers *et al.*, "The impact of late treatment-toxicity on generic health-related quality of life in head and neck cancer patients after radiotherapy," *Oral Oncol.*, vol. 47, no. 8, pp. 768–774, 2011.
[64] J. E. van Timmeren, D. Cester, S. Tanadini-Lang, H. Alkadhi, and B. Baessler, "Radiomics in medical imaging—'how-to' guide and critical reflection," *Insights Imaging*, vol. 11, no. 1, 2020.
[65] V. Kumar *et al.*, "Radiomics: The Process and the Challenges," *Magn. Reson. Imaging*, vol. 30, no. 9, pp. 1234–1248, 2012.
[66] J. J. M. Van Griethuysen *et al.*, "Computational radiomics system to decode the radiographic phenotype," *Cancer Res.*, vol. 77, no. 21, pp. e104–e107, 2017.
[67] A. Vial *et al.*, "The role of deep learning and radiomic feature extraction in cancer-specific predictive modelling: A review," *Transl. Cancer Res.*, vol. 7, no. 3, pp. 803–816, 2018.
[68] A. S. Tagliafico, M. Piana, D. Schenone, R. Lai, A. M. Massone, and N. Houssami, "Overview of radiomics in breast cancer diagnosis and prognostication," *Breast*, vol. 49, pp. 74–80, 2020.
[69] K. Yu *et al.*, "Radiomic analysis in prediction of Human Papilloma Virus status," *Clin. Transl. Radiat. Oncol.*, vol. 7, pp. 49–54, 2017.
[70] S. Leger *et al.*, "CT imaging during treatment improves radiomic models for patients with locally advanced head and neck cancer," *Radiother. Oncol.*, vol. 130, pp. 10–17, 2019.
[71] X. Fave *et al.*, "Using Pretreatment Radiomics and Delta-Radiomics Features to Predict Non–Small Cell Lung Cancer Patient Outcomes," *Int. J. Radiat. Oncol.*, vol. 98, no. 1, p.

249, 2017.
[72] P. Real, *Computer analysis of images and patterns*, vol. 47, no. 1–2. 2019.
[73] L. Rossi *et al.*, "Texture analysis of 3D dose distributions for predictive modelling of toxicity rates in radiotherapy," *Radiother. Oncol.*, vol. 129, no. 3, pp. 548–553, 2018.
[74] B. Liang *et al.*, "Prediction of Radiation Pneumonitis With Dose Distribution: A Convolutional Neural Network (CNN) Based Model," *Front. Oncol.*, vol. 9, no. January, pp. 1–7, 2020.
[75] T. Adachi, M. Nakamura, T. Shintani, and T. Mistsuyoshi, "Prediction of Radiation Pneumonitis After Lung Stereotactic Body Radiation Therapy Using Dosiomics Features A Retrospective Multi Institutional Study.pdf."
[76] B. Liang *et al.*, "Dosiomics: Extracting 3D spatial features from dose distribution to predict incidence of radiation pneumonitis," *Front. Oncol.*, vol. 9, no. APR, pp. 1–7, 2019.
[77] A. Wu *et al.*, "Dosiomics improves prediction of locoregional recurrence for intensity modulated radiotherapy treated head and neck cancer cases," *Oral Oncol.*, vol. 104, 2020.
[78] T. S. Madhulatha, "AN OVERVIEW ON CLUSTERING METHODS," vol. 2, no. 4, pp. 719–725, 2012.
[79] S. Chormunge and S. Jena, "Correlation based feature selection with clustering for high dimensional data," *J. Electr. Syst. Inf. Technol.*, vol. 5, no. 3, pp. 542–549, 2018.
[80] C. A. Kumar, "Analysis of unsupervised dimensionality reduction techniques," *Comput. Sci. Inf. Syst.*, vol. 6, no. 2, pp. 217–227, 2009.
[81] D. Müllner, "Modern hierarchical, agglomerative clustering algorithms," no. 1973, pp. 1–29, 2011.
[82] C. Strobl, A. L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, "Conditional variable importance for random forests," *BMC Bioinformatics*, vol. 9, pp. 1–11, 2008.
[83] D. Jarrett, E. Stride, K. Vallis, and M. J. Gooding, "Applications and limitations of machine learning in radiation oncology," *Br. J. Radiol.*, vol. 92, no. 1100, p. 20190001, 2019.
[84] N. Altman and M. Krzywinski, "The curse(s) of dimensionality this-month," *Nat. Methods*, vol. 15, no. 6, pp. 399–400, 2018.
[85] M. Avanzo, L. Wei, M. Valli, and O. Morin, "Machine and deep learning methods for radiomics," *Med. Phys.*, vol. 47, no. 5, pp. e185–e202, 2020.
[86] L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, 1st ed. Boca Raton, Fl: Chapman & Hall/CRC Press, 1984.
[87] E. Scornet, "Trees, forests, and impurity-based variable importance," 2020.
[88] L. Breiman, "Random Forests," *Mach. Learn.*, pp. 5–32, 2013.
[89] R. Genuer, J. M. Poggi, and C. Tuleau-Malot, "Variable selection using random forests," *Pattern Recognit. Lett.*, vol. 31, no. 14, pp. 2225–2236, 2010.
[90] A. S. Garden *et al.*, "Preliminary results of Radiation Therapy Oncology Group 97-03: A randomized phase II trial of concurrent radiation and chemotherapy for advanced squamous cell carcinomas of the head and neck," *J. Clin. Oncol.*, vol. 22, no. 14, pp. 2856–2864, 2004.
[91] G. Calais *et al.*, "Randomized Trial of Radiation Therapy Versus," *J. Natl. Cancer Inst.*, vol. 91, no. 24, pp. 2081–2086, 1999.
[92] F. C. Holsinger and R. L. Ferris, "Transoral endoscopic head and neck surgery and its role within the multidisciplinary treatment paradigm of oropharynx cancer: Robotics, lasers, and clinical trials," *J. Clin. Oncol.*, vol. 33, no. 29, pp. 3285–3292, 2015.
[93] L. C. Peng *et al.*, "Prospective evaluation of patient reported swallow function with the Functional Assessment of Cancer Therapy (FACT), MD Anderson Dysphagia Inventory

(MDADI) and the Sydney Swallow Questionnaire (SSQ) in head and neck cancer patients," *Oral Oncol.*, vol. 84, no. July, pp. 25–30, 2018.

[94] J. Shah, "A century of progress in head and neck cancer," *J. Head Neck Physicians Surg.*, vol. 4, no. 2, p. 50, 2016.

[95] C. M. Nutting *et al.*, "Parotid-sparing intensity modulated versus conventional radiotherapy in head and neck cancer (PARSPORT): A phase 3 multicentre randomised controlled trial," *Lancet Oncol.*, vol. 12, no. 2, pp. 127–136, 2011.

[96] S. Rogers and B. Barber, "Using PROMs to guide patients and practitioners through the head and neck cancer journey," *Patient Relat. Outcome Meas.*, vol. Volume 8, pp. 133–142, 2017.

[97] B. Ojo, E. M. Genden, M. S. Teng, K. Milbury, K. J. Misiukiewicz, and H. Badr, "A systematic review of head and neck cancer quality of life assessment instruments," *Oral Oncol.*, vol. 48, no. 10, pp. 923–937, 2012.

[98] C. Holländer-Mieritz, J. Johansen, C. Johansen, I. R. Vogelius, C. A. Kristensen, and H. Pappot, "Comparing the patients' subjective experiences of acute side effects during radiotherapy for head and neck cancer with four different patient-reported outcomes questionnaires," *Acta Oncol. (Madr).*, vol. 58, no. 5, pp. 603–609, 2019.

[99] F. M. Fang, Y. T. Liu, Y. Tang, C. J. Wang, and S. F. Ko, "Quality of Life as a Survival Predictor for Patients with Advanced Head and Neck Carcinoma Treated with Radiotherapy," *Cancer*, vol. 100, no. 2, pp. 425–432, 2004.

[100] A. Eskander *et al.*, "Emergency department visits and unplanned hospitalizations in the treatment period for head and neck cancer patients treated with curative intent: A population-based analysis," *Oral Oncol.*, vol. 83, no. June, pp. 107–114, 2018.

[101] F. Denis *et al.*, "Randomized Trial Comparing a Web-Mediated Follow-up With Routine Surveillance in Lung Cancer Patients," vol. 109, no. October, pp. 1–8, 2017.

[102] M. Kamal *et al.*, "Modeling symptom drivers of oral intake in long-term head and neck cancer survivors," *Support. Care Cancer*, vol. 27, no. 4, pp. 1405–1415, 2019.

[103] S. N. Rogers, C. Semple, M. Babb, and G. Humphris, "Quality of life considerations in head and neck cancer: United Kingdom National Multidisciplinary Guidelines," *J. Laryngol. Otol.*, vol. 130, no. S2, pp. S49–S52, 2016.

[104] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[105] C. Goldsmith, J. Haviland, Y. Tsang, M. Sydenham, and J. Yarnold, "Large breast size as a risk factor for late adverse effects of breast radiotherapy: Is residual dose inhomogeneity, despite 3D treatment planning and delivery, the main explanation?," *Radiother. Oncol.*, vol. 100, no. 2, pp. 236–240, 2011.

[106] Y. S. Butler-xu, M. Marietta, and M. Mitchell, "The effect of breast volume on toxicity using hypofractionated regimens for early stage breast cancer for patients," *Advancesradonc*, vol. 4, no. 2, pp. 261–267, 2019.

[107] T. S. Trialists, "The UK Standardisation of Breast Radiotherapy (START) Trial A of radiotherapy hypofractionation for treatment of early breast cancer: a randomised trial," *Lancet Oncol.*, vol. 9, no. 4, pp. 331–341, 2008.

[108] J. Yarnold *et al.*, "Fractionation sensitivity and dose response of late adverse effects in the breast after radiotherapy for early breast cancer: Long-term results of a randomised trial," *Radiother. Oncol.*, vol. 75, no. 1, pp. 9–17, 2005.

[109] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC

plot when evaluating binary classifiers on imbalanced datasets," *PLoS One*, vol. 10, no. 3, pp. 1–21, 2015.
[110] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," *Proc. - Int. Conf. Pattern Recognit.*, pp. 3121–3124, 2010.
[111] D. Mason, Scaramallion, Rhaxton, and et al., "Pydicom: An open source DICOM library." 2020.
[112] I. Community, "IBSI Documentation," 2019.
[113] 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," *Phys. Med. Biol.*, vol. 60, no. 14, pp. 5471–5496, 2015.
[114] M. Chavent, R. Genuer, and J. Saracco, "Combining clustering of variables and feature selection using random forests," *Commun. Stat. Simul. Comput.*, vol. 50, no. 2, pp. 426–445, 2021.
[115] C. Strobl, A. L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC Bioinformatics*, vol. 8, 2007.

Table S 1: MDASI-HN questions with abbreviations

MDASI Question	MDASI Question Abbreviation
1. Your **pain** at its WORST?	MDASI_Pain
2. Your **fatigue (tiredness)** at its WORST?	MDASI_Fatigue
3. Your **nausea** at its WORST?	MDASI_Nausea
4. Your **disturbed sleep** at its WORST?	MDASI_Sleep
5. Your feeling of being **distressed (upset)** at its WORST?	MDASI_Distressed
6. Your **shortness of breath** at its WORST?	MDASI_ShortBreath
7. Your problem with **remembering things** at its WORST?	MDASI_Remembering
8. Your problem with **lack of appetite** at its WORST?	MDASI_Appetite
9. Your feeling **drowsy (sleepy)** at its WORST?	MDASI_Drowsy
10. Your having a **dry mouth** at its WORST?	MDASI_DryMouth
11. Your feeling **sad** at its WORST?	MDASI_Sad
12. Your **vomiting** at its WORST?	MDASI_Vomiting
13. Your **numbness or tingling** at its WORST?	MDASI_NumbTingling
14. Your problem with **mucus** in our mouth and throat at its WORST?	MDASI_Mucus
15. Your difficulty **swallowing/chewing** at its WORST?	MDASI_SwallowChew

16. Your **choking/coughing** (food/liquids going down the wrong pipe) at its WORST?	MDASI_ChokeCough
17. Your difficulty with **voice/speech** at its WORST?	MDASI_VoiceSpeech
18. Your **skin pain/burning/rash** at its WORST?	MDASI_SkinRash
19. Your **constipation** at its WORST?	MDASI_Constipation
20. Your problem with **tasting food** at its WORST?	MDASI_Taste
21. Your **mouth/throat sores** at their WORST?	MDASI_MouthSores
22. Your problem with your **teeth or gums** at its WORST?	MDASI_TeethGums
23. **General activity?**	MDASI_Activity
24. **Mood?**	MDASI_Mood
25. **Work (including work around the house)?**	MDASI_Work
26. **Relations with other people?**	MDASI_Relationships
27. **Walking?**	MDASI_Walking
28. **Enjoyment of life?**	MDASI_EnjoyLife

Table S 2: MDADI questions and abbreviations

MDADI Question	MDADI Question Abbreviation
E1. My swallowing ability limits my day-to-day activities.	MDADI_Activities
E2. I am embarrassed by my eating habits.	MDADI_Embarrassed
F1. People have difficulty cooking for me.	MDADI_Cooking
P2. Swallowing is more difficult at the end of the day.	MDADI_EndOfDay
E7. I do not feel self-conscious when I eat.	MDADI_NotSelfConscious
E4. I am upset by my swallowing problem.	MDADI_Upset
P6. Swallowing takes great effort.	MDADI_Effort
E5. I do not go out because of my swallowing problem.	MDADI_NotGoOut
F5. My swallowing difficulty has caused me to lose income.	MDADI_Income
P7. It takes me longer to eat because of my swallowing problem.	MDADI_LongerToEat
P3. People ask me, “Why can’t you eat that?”	MDADI_WhyNotEat
E3. Other people are irritated by my eating problem.	MDADI_OthersIrritated
P8. I cough when I try to drink liquids.	MDADI_LiquidsCough
F3. My swallowing problems limit my social and personal life.	MDADI_LimitSocial

F2. I feel free to go out to eat with my friends, neighbors, and relatives.	MDADI_FreeToGoOut
P5. I limit my food intake because of my swallowing difficulty.	MDADI_LimitFood
P1. I cannot maintain my weight because of my swallowing problem.	MDADI_MaintainWeight
E6. I have low self-esteem because of my swallowing problem.	MDADI_SelfEsteem
P4. I feel that I am swallowing a huge amount of food.	MDADI_HugeAmountOfFood
F4. I feel excluded because of my eating habits.	MDADI_Excluded