Integrating Large Language Models and Image-Based Techniques for Radiotherapy Toxicity Prediction

Choi, Min Seo

Department of Medicine
Graduate School
Yonsei University

Integrating Large Language Models and Image-Based Techniques for Radiotherapy Toxicity Prediction

Advisor Kim, Jin Sung

A Dissertation Submitted
to the Department of Medicine
and the Committee on Graduate School
of Yonsei University in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy in Medical Science

Choi, Min Seo

June 2025

Integrating Large Language Models and Image-Based Techniques for Radiotherapy Toxicity Prediction

This Certifies that the Dissertation of Choi, Min Seo is Approved

`

Committee Chair	Kim, Jihun
Committee Member	Kim, Hojin
Committee Member	Kim, Hwiyoung
Committee Member	Veeraraghavan, Harini
Committee Member	Kim, Jin Sung

Department of Medicine
Graduate School
Yonsei University
June 2025

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Dr. Jin Sung Kim, for his guidance and mentorship during my Ph.D. His insights into the field of medical physics and his dedication to helping his students grow as a better researcher, presenter, and aspiring medical physicist - have been truly invaluable. I am extremely grateful to my thesis committee: Dr. Jihun Kim, Dr. Hwiyoung Kim, Dr. Hojin Kim, and Dr. Veeraraghavan for their guidance throughout my graduation. Their expertise has been instrumental in shaping and enhancing this project.

During my Ph.D, I was extremely fortunate to learn from many inspiring mentors. I am deeply indebted to Dr. Harini Veeraraghavan for her unwavering support, sharp insights, and for being an endless source of inspiration. I am also grateful to Dr. Jiang for his expertise and detailed supervision as well as kind words of encouragement. I would also like to thank Dr. Jeesuk Chang for giving me the opportunity to be involved in translational research projects, which gave me valuable insight into their clinical impact. His consistent kindness and enthusiasm for exploring new ideas greatly enriched my research experience. I am grateful to Dr. Jihun Kim who went above and beyond, dedicating his time to my development and teaching me how to become an independent researcher. I would like to acknowledge Dr. Jaehee Chun for encouraging me to keep learning new skills and for all the advice sessions – always seeming to have the perfect answer to everything.

I would like to thank all my friends at the Medical Physics and Biomedical Engineering Lab. I will always remember our late-night discussions about life in Shinchon and the memories we made along the way. It has been a privilege to share this journey through graduate school with all of you. During my time as a visiting researcher, I met many incredible friends at Memorial Sloan Kettering Cancer Center who helped me stay grounded and reminded me not to give up. A special thank you to my favorite medical physicist couple, Sebastian and Ingrid, for always checking in on me and making sure I was doing okay, and to Donghoon for his constant encouragement and thoughtful guidance on my career and future.

Last but not least, very special thanks to my family: thank you for always being there for me, supporting me, and believing in me even when I doubted myself. Looking back, I think one of the most valuable lessons I have learned through this journey is that even tasks that seem impossible to tackle alone can be overcome with the love and support of those around you. I want to dedicate this thesis to everyone who helped me reach the end of this journey.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Integrating Large Language Models and Image-Based Techniques for Radiotherapy Toxicity Prediction

Radiation Therapy (RT) is an important treatment modality for patients with thoracic cancer along with concurrent chemotherapy and surgery. Although RT aims to precisely target tumors, nearby normal tissues may still receive substantial radiation doses, leading to RT-induced toxicities such as esophagitis, cardiac toxicity, and pneumonitis, which can adversely affect patients' quality of life. With the growing number of long-term survivors, reducing treatment-related toxicities has become a key priority in the planning of radiation therapy.

Management of such toxicities can be addressed at various stages of the RT planning workflow. During the simulation phase, patients undergo a CT scan that serves as the basis for treatment planning. However, respiratory motion can introduce imaging artifacts and lead to discrepancies between the planned and delivered doses. Breath-hold techniques are commonly employed to reduce motion-related variability, but the current standard methods often require patients to hold their breath for extended periods, which can be difficult for some individuals. In the planning stage, accurate delineation of organs-at-risk (OARs) and tumor volumes is critical to ensure accurate RT planning, yet remains a significant challenge. Manual contouring is labor-intensive and prone to interobserver variability, making it a potential bottleneck in the clinical workflow. Finally, in the period between planning and treatment delivery, incorporating patient-specific toxicity prediction models can provide valuable decision support, enabling clinicians to better anticipate and mitigate potential adverse effects. However, as individual responses to RT vary, building these prediction models can be challenging.

Therefore, the aim of this thesis is to develop novel methods to address key challenges in the management of RT–induced toxicity. The thesis is structured into three chapters, each presenting a distinct contribution toward improving various aspects of toxicity management in the RT workflow. We first start with clinically implementing a novel breath-hold technique called continous positive airway pressure. Its clinical feasibility was tested on patients with breast cancer who underwent RT and geomtrically and dosimetrically compared against conventional methods including free-breathing and deep inspiration breath-hold.

In the subsequent chapter, we explore the role of a deep learning-based automated segmentation algorithm as a tool for streamlining RT planning while ensuring the accuracy needed to reduce errors that may impact RT-induced toxicity. Our deep learning algorithm was applied to retrospective breast cancer patient data, where we evaluated the geometric accuracy of the segmentations. These results were compared with the conventional atlas-based segmentation method to assess improvements in precision and efficiency.

In the final chapter, we focus on the development of a multi-modal prediction model for RT-induced esophagitis in patients with esophageal cancer. This chapter introduces an innovative approach by integrating both imaging and clinical data. We employ a pretrained image encoder to extract relevant features from medical images, alongside a large language model to incorporate clinical information, marking a shift away from traditional image-only prediction models. This multi-modal framework aims to enhance the accuracy and clinical utility of predicting RT-induced esophagitis, ultimately offering a more comprehensive tool for personalized patient care.

Key words : Deep learning, large language model, multi-modal, prediction

I. INTRODUCTION

1.1. Background and motivation

1.1.1 Radiation therapy

Cancer is a leading cause of mortality in the world, constituting about 50% of the annual cancer mortality among new cases[1]. Radiation therapy (RT) is a important treatment modality for cancer, often used alongside chemotherapy and surgery. RT uses ionizing radiation to damage the DNA of cancer cells, either directly by interacting with cellular DNA or indirectly by generating free radicals through the ionization or excitation of water molecules within the cell[2]. This damage ultimately leads to tumor cell death. Among the different types of radiation used in therapy, photon beams such as X-rays and gamma rays remain the most commonly employed. More recently, particle radiation including electrons, protons, and neutrons has been introduced for specific clinical applications, providing additional options for tailored treatment approaches[2]. RT can be used to treat a wide variety of cancer sites, with common targets including the breast, lung, prostate, head and neck, and skin[1]. It is estimated that approximately 50% of all cancer patients receive RT, with 40% of these cases delivered with curative intent[3].

RT is a collaborative effort involving a multidisciplinary team within radiation oncology. The workflow for RT consists of several key stages which include simulation, treatment planning, plan approval and quality assurance, and finally radiotherapy delivery (Fig 1). After initial consulation with the physician, patients first undergo simulation imaging, such as computed tomography (CT) and/or magnetic resonance imaging (MRI). The simulation scan acquired at this stage serves as the foundation for radiation therapy planning. Therefore, maintaining image quality by reducing artifacts caused by patient motion, positioning, or other factors is essential for accurate target delineation and dose calculation.Using the anatomic information available from the simulation scans, target volumes and normal tissues are segmented. Based on the defined structures and treatment constraints specific for each disease site, a personalized radiation dose distribution is planned with the goal of delivering an effective dose to the tumor while minimizing exposure to surrounding healthy tissue. The treatment plan then undergoes a series of checklists and quality assurance procedures by medical physicists to ensure it can be safely and accurately delivered on the treatment machine, before being administered to the patient in multiple treatment fractions.

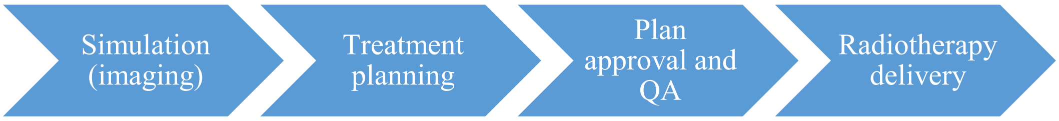


Fig 1. Overall workflow of radiation therapy; Abbreviations- QA: quality assurance

1.1.2 Radiation therapy-induced toxicity

Although the primary goal of RT is to deliver an effective dose to the tumor while minimizing exposure to surrounding healthy tissues, some degree of irradiation to normal tissues is often unavoidable. When healthy tissue is exposed to ionizing radiation, it can result in DNA damage, potentially leading to radiation-induced toxicity. This toxicity can manifest as acute side effects, typically occurring within 1–2 weeks after treatment, or as late effects that may appear months or even years later. Both forms of toxicity can significantly affect a patient’s quality of life, ranging from mild discomfort to severe, long-lasting complications.With advances in cancer treatment leading to more long-term survivors, modern RT planning has increasingly emphasized minimizing these RT-induced toxicity to improve long-term patient outcomes.

In cancers of the thoracic region, such as breast, lung, and esophageal cancer, common radiation-related toxicities include radiation pneumonitis, which may present with CT abnormalities, cough, or shortness of breath[4], where severe cases (grade 3 or higher) can require supplemental oxygen. Another possible toxicity is cardiotoxicity which is known to be a significant source of mortality for cancer survivors [5], which may lead to heart failure (e.g. myocardial fibrosis, valvular heart disease) or ischemic coronary artery disease [6]. Last but not least, another major thoracic toxicity is esophagitis, which refers to inflammation of the esophagus. Symtoms can include pain and difficulty swallowing. Although these toxicities vary in severity, they can significantly impact patient well-being. Therefore, early identification and mitigation during treatment planning are crucial to minimizing side effects and improving clinical outcomes.

1.1.3 Toxicity Prediction

One approach to managing and potentially reducing RT-induced toxicity is through a pretreatment prediction model. Numerous studies across various cancer types have explored the use of patient scans and clinical characteristics to predict toxicity before treatment begins. The clinical significance of such a model lies in its ability to support decision-making. By using pretreatment data, clinicians can identify patients at higher risk of toxicity early on, allowing for timely interventions and treatment adjustments. This, in turn, could lead to more personalized care, improving both treatment outcomes and patient well-being (Fig 2).

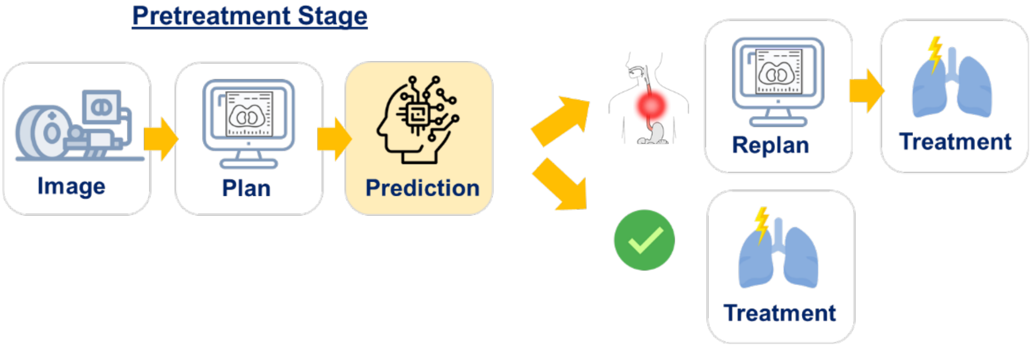


Fig 2. The clinical implication of pretreatment prediction model for a personalized plan

There are a number of different approaches available for RT-induced toxicity. Normal Tissue Complication Probability (NTCP) models are one of the first methods developed in order to estimate the risk of RT-induced toxicity. NTCP models are essentially mathematical modelling utilizing clinical and dose distribution information, tuning the model to best fit the training cohort [7]. The dose distribution informations are typically based on dose volume histograms (DVH) often utilizing dose parameters such as the mean organ dose or volume parameters (e.g. V20). NTCP models have previously been established for a variety of disease sites [8–10] and has been demonstrated to assist as a tool for determining the best treatment plan by comparing the risk of each plan [11,12]. The most common modelling methods for NTCP is the Lyman-Kutcher-Burman (LKB) and relative seriality (RS) models [13,14].

Voxel based analysis (VBA) is another category of prediction model used in radiation oncology. VBA is a method that utilizes the 3D dose distributions of patient population with and without toxicity in order to determine the heterogeneous dose sensitivities occuring with the organ of interest. VBA consists of two main processes which are 1) spatial normalization to common reference frame followed by 2) statistical analysis of the dose response between the group with and without adverse events. The spatial normalization step involves deformable image registration of all patients in the poluation into a single reference patient, typically picked as a patient with a typical individual anatomy. This is an important step as it forms the basis of all upcoming analysis involved in VBA. Most common models include demons and B-spline algorithms [15,16]which are iterative registration models and many studies utilized software packages found in SyN and elastix [17,18]. Thorax, head and neck as well as prostate applications have been reported by previous research [19–22].

Imaging-based prediction models using deep learning and radiomics have emerged as a key area of research in radiation oncology, aiming to enhance treatment personalization and outcome prediction. These models typically utilize RT treatment planning data, including CT scans, dose distributions, and sometimes contours, to predict clinical endpoints such as toxicity or treatment response. Studies have employed various input configurations to optimize model performance. Some models use a single modality, such as CT or dose distribution, while others combine multiple inputs for improved accuracy. For instance, one study demonstrated enhanced predictive power by integrating CT and dose [23], while another showed that incorporating CT, dose, and contours outperformed single-input benchmarks [24]. Beyond input selection, model architecture and feature extraction strategies play a crucial role. Feature extraction can occur at different stages, early or late in the model pipeline, with either single-layer or multi-layer approaches. The choice of feature extraction method impacts the model’s ability to capture spatial and dosimetric patterns relevant to treatment outcomes. Additionally, hybrid approaches combining handcrafted radiomic features with deep learning-based feature extraction are being explored to further improve robustness and interpretability.

1.1.4 Current limitations and motivations

In order to build a robust prediction model, there are several sources of errors that should be addressed. One source of error could be related to target and OAR segmentation during the planning stage. Accurate delineation of OARs and tumor forms the basis of RT planning as it ultimately determines the radiation dose distribution that the patients are going to receive and therefore is crucial for ensuring accurate treatment planning for the patient. However, manual delineation of these structures can be time-consuming and resource intensive. Furthermore, manual delineation has been reported to be associated with inter-observer variations which could be another added source of errors especially if involving multi-centers or multi-observers in a study. AI can be used for patient evaluation in the initial phase of the RT planning. RT planning is made up of many steps that require human input and many areas have been shown to benefit from automation through AI.

Another potential source of error is the simulation imaging. Even though patients are immobilized during simulation scans and treatment, tumor motion can still occur due to internal movements, such as swallowing or breathing. Since the simulation scan forms the basis of the patient's RT planning this could be problematic. In order to mitigate this, deep inspiration breath-Hold (DIBH) is the most commonly used technique for managing breathing during RT for thoracic cancers. This technique requires patients to voluntarily hold their breath during treatment. However, some patients may have reduced lung capacity, and maintaining a prolonged breath-hold can be difficult, making DIBH challenging for certain individuals. Failure to obtain good breath-hold can lead to unreliable simulation scan which can not only impact the reliability of the prediction model but the entire planning as well as delivery.

Secondly, most deep learning-based prediction models developed to date primarily rely on single-modality inputs. However, there is a growing trend toward multi-modal integration to enhance predictive performance. Beyond imaging data, other valuable patient information exists in the form of clinical texts, such as clinical notes, charts, and laboratory test results. The potential for leveraging such clinical data has become increasingly feasible with the advancement of large language models (LLMs), which have significantly improved the ability to encode and interpret textual information. Several studies have demonstrated the benefits of integrating clinical text with imaging through image-to-text alignment, enhancing conventional image-based models in applications such as segmentation and survival prediction [25,26].

1.2. Aims and objectives

The primary aim of this thesis is to explore and evaluate innovative imaging techniques alongside advanced AI-driven technologies with the goal of enhancing clinical workflow efficiency and improving the accuracy of patient-specific predictions. By integrating these approaches, the ultimate objective is to develop a comprehensive AI-based framework capable of reliably predicting RT-induced toxicity. To systematically address this aim, the study is structured around the following three key objectives:

1. To develop a deep learning-based auto-segmentation model and assess its clinical feasibility for radiotherapy.

2. To clinically evaluate a novel breath-hold technique and compare its effectiveness with conventional methods.

3. To develop a multimodal model for predicting adverse events by integrating imaging features and clinical data using deep learning and LLMs.

1.3. Thesis structure

In Chapter 2, we present the development of a deep learning–based autosegmentation model using convolutional neural network (CNN) for both tumor structures and organs at risk (OARs), with the aim of observing its feasibility within the RT workflow. We evaluated and compared conventional and commercially available atlas-based segmentation methods based on image registration algorithm with deep learning–based approaches for automatic delineation of key structures in breast RT planning, including clinical target volumes (CTVs), OARs, and heart substructures. The model's performance was assessed against physician-drawn ground truth contours, highlighting its potential role in supporting automated planning for breast RT.

Chapter 3 explores the clinical evaluation of breathing management techniques aimed at reducing radiation therapy–induced toxicity. We begin by assessing the heart-sparing potential of a novel breath-hold technology using continuous positive airway pressure (CPAP), a device traditionally used for sleep apnea. Given its novel application in RT, evaluating its feasibility and reproducibility is essential. In this study, we assessed inter- and intra-fractional reproducibility using real patient data by measuring the heart-to-target distance and quantifying lung displacement along three spatial directions. We report on its real-world implementation at our institution across a large patient cohort. Subsequently, we compare CPAP with the current standard, deep inspiration breath-hold (DIBH), through radiation therapy planning analyses to assess its clinical feasibility.

Chapter 4 presents the development of a multi-modal deep learning framework for predicting acute esophagitis in esophageal cancer patients. The model integrates imaging features and clinical variables to enhance predictive performance by leveraging complementary data modalities. Specifically, a transformer architecture was employed as the image encoder to capture spatial and contextual information from imaging inputs, while a large language model was utilized as the text encoder to process structured and unstructured clinical data. This chapter details the data preprocessing pipelines, model architecture, training strategy, and evaluation metrics used to rigorously assess model robustness and clinical applicability.

Finally, Chapter 5 summarizes the key findings and research contributions of the study, highlighting their impact on toxicity management in radiation therapy. It also discusses limitations and suggests future directions for further development, including improving model performance and enhancing clinical application.

2. The Role of Deep Learning-based Automated Segmentation Model in Radiation Therapy

2.1. Background

Breast cancer is one of the most common cancer in the world, being the leading cauase of cancer deaths. As a treatment modeality for breast cancer, RT is being increasingly utilized. In the era of three-dimensional conformal and intensity-modulated RT, precise delineation of the clinical target volume (CTV) and organs at risk (OARs) is essential, as inaccuracies can result in excessive radiation to normal tissues and potentially increase treatment-related toxicity. Furthermore, as many breast cancer patients live for decades following radiation therapy, they remain at risk for long-term adverse effects, such as lymphedema, radiation pneumonitis, hypothyroidism, and cardiotoxicity, which can significantly impact their quality of life. However, there are limitations in the RT planning workflow related to autosegmentation that still require improvement, particularly the burden of manual segmentation and the inter-observer variability resulting from the subjective nature of contouring by individual physicians.

Conventionally, manual segmentation had been the standard approach for RT planning; however, it is time-intensive and susceptible to inter- and intra-observer variability. To address these limitations, auto-segmentation has gained considerable attention for its potential to improve efficiency and consistency in clinical workflows. Earlier efforts to automate segmentation primarily relied on atlas-based auto-segmentation (ABAS), which became the conventional approach and is supported by several commercially available solutions for use in cancer sites including the head and neck, prostate, breast and lung cancer. Despite its usage, ABAS has some limitations, including suboptimal contouring for low-contrast structures, slow image registration, and the frequent need for manual corrections to enhance segmentation accuracy. More recently, advancements in computational power and reductions in financial barriers have shifted research focus toward deep learning-based auto-segmentation (DLBAS). Studies have demonstrated the potential of deep learning algorithms to enhance auto-segmentation accuracy for breast cancer and other anatomical sites, further supporting their integration into clinical practice.

Beyond routine clinical operations, deviations from RT protocols are associated with increased risks of treatment failure and patient mortality. A critical and longstanding issue contributing to such deviations is inter-observer variability in target and OAR contouring, which has remained a major focus of research and quality assurance efforts in the field. To date, multiple studies have demonstrated significant IOVs in delineating target volumes, including clinical target volumes (CTVs) and organs-at-risk (OARs), in various types of cancers, both within and outside clinical trials [27–30]. Efforts to reduce interobserver variability (IOV) in contouring have focused on strategies such as site-specific atlases, consensus guidelines, trial-specific protocols, education, audits, and peer review. Benchmark studies, or dummy runs, are commonly used at the start of clinical trials or during individual case reviews as part of radiation therapy quality assurance (RTQA). While these methods have improved IOV, their limitations highlight the need for alternative approaches. DL-based auto-contouring has shown time-saving benefits and reduced

IOV in head and neck, prostate, and breast cancers [31]. Unlike static guidelines or atlases, it offers interactive adjustments, making it easier to adapt contours to individual anatomy and positioning [27]. However, its clinical utility within RTQA programs remains unstudied.

As more DL based tools are being implemented in radiation oncology, few studies have evaluated their real-world clinical utility. To address challenges in RT contouring, we investigated DLBAS in two contexts: routine clinical RT planning and multi-center collaborative settings. Both efforts have centered on breast cancer application. First, we conducted a single-center study comparing the accuracy of DLBAS to commercial ABAS and manual contours in RT planning for patients with breast cancer where the aim was to demonstrate the feasibility of using an automated contouring tool in a clinical workflow. The second part of the study was conducted in collaboration with the Korean Radiation Oncology Group (KROG), where we assessed DLBAS's impact on IOV across 31 institutions. Here, we conducted a two-phase study comparing IOV with and without the aid of DL-generated contours, visually assessing both the extent and location of variation.

2.2. Clinical Evaluation of Deep Learning-based Segmentation Model

2.2.1 Materials and Methods

A retrospective dataset from 62 patients with breast cancer who received breast-conservation surgery and RT between 2016 and 2019 at Yonsei Cancer Center (Seoul, South Korea) were included in this study. Each patient's data included plan CT scans as well as manual ground truth contour delineated by a single experienced radiation oncologist following the ESTRO guidelines. The list of contours included the clinical target volumes and lymph nodes (e.g. axillary, internal mammary and supraclavicular) as well as the OARs (lungs, esophagus, spinal cord and thyroid) including the heart substructures (atria, ventricles and right coronary artery (RCA) and left anterior descending artery (LAD)).

In this study, we developed a 3D fully convolutional DenseNet (FCDN) segmentation model (Fig 3). The FCDN architecture is made up of dense blocks that resemble the residual blocks in a U-Net architecture. Following the convolution layer, the transition down layers consist of BN, RELU, 1 × 1 convolution, dropout ($p = 0.2$), and a 2 × 2 max pooling operation. The skip connection components represent the concatenation of the feature maps from the down-sampling path with those in the up-sampling path, thereby ensuring a high-resolution output. Finally, the transition up (TU) layers consist of 3 × 3 deconvolutions with a stride of 2 to progressively recover spatial resolution. The model was trained for 200 epochs using 35 patients for training and 13 patients for validation.

For baseline comparison, we used two commercial atlas based ABAS systems, Mirada's Workflow Box (Mirada Medical, UK) and MIM Maestro (MIM Software, USA), to automatically segment target structures. The atlas libaries were made using the same training data as for DLBAS. In MIM, the first step in building an atlas was to assign a randomly selected reference or "template" subject. The remaining subjects were registered to the template one by one, along with the expert contours. Although MIM offers a tool to edit the registration alignment, in order to obtain a nonbiased auto-segmentation and keep the experimental settings as consistent as possible, we did not intervene during registration and segmentation. The final step was the segmentation process itself. In MIM, under the "Atlas Segment" tool, we selected the contours and ran the segmentation with the following default settings: Number of Match = 1, Mirroring Enabled and Multicontour finalisation method = Majority Vote. Next, because a single atlas segmentation was selected, the algorithm automatically searched for the atlas subject that best matched the input CT. Then, expert contours of the atlas subject were deformed, registered, and transferred to the input CT, based on intensity and a freeform cubic spline interpolation [32].

In Mirada, a workflow that linked the atlas created by the user and the segmentation operation was created that simply required selecting the input CT and assigning it to the workflow in a single click. As it functions like a black box, it is not possible to change settings in WFB. Also, unlike MIM, WFB does not require the assignment of template patients or any further user intervention. The construction of the library simply involved selecting CT scans and their corresponding structures. Once every subject was added, the atlas files were uploaded to the WFB server.

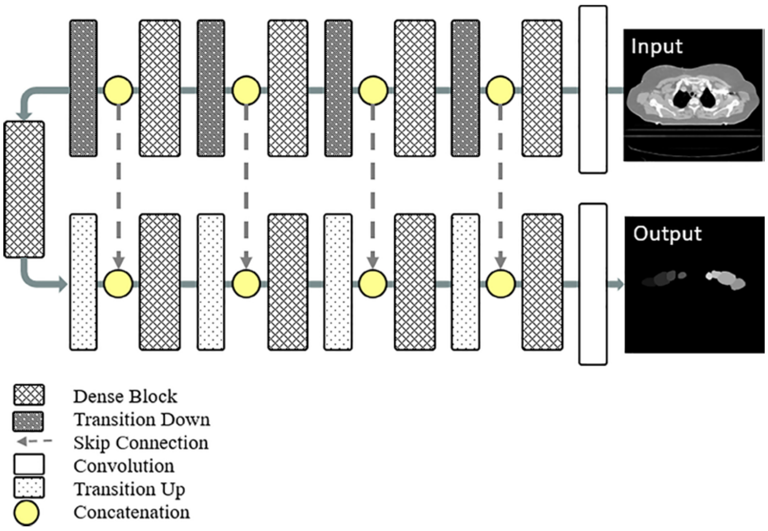


Fig 3. The architecture of the proposed DL segmentation model

The auto-segmentation was quantitatively assessed with 14 test patients using the Dice Similarity Coefficient (DSC) defined as $2 * |A \cap B| / (|A| + |B|)$ and 95% Hausdorff Distance (HD95) defined as $H(A,B) = \max\{h(A,B), h(B,A)\}$ where A and B are two different point sets. In this study, the manual contours created by a single expert radiation oncologist served as the ground truth, with which the ABAS and FCDN contours were compared.

A pairwise t-test was conducted to determine if there was a statistically significant difference between the results from the different software packages. Since there are three segmentation methods to compare, we adopted Bonferroni correction to address the multiple-comparison correction [33] with n = 3 and the alpha value adjusted to 0.0167 (0.05/3). A p-value of less than

0.0167 was determined to be a rejection of the null hypothesis and therefore a statistically significant result.

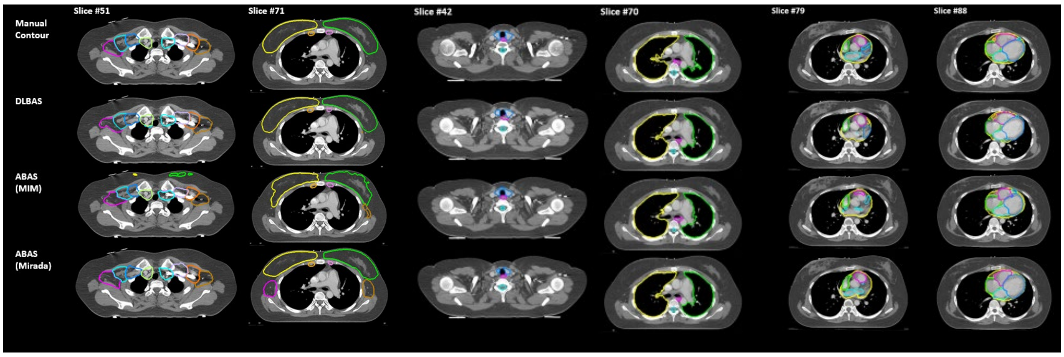


Fig 4. Examples of a) CTV, b) OAR, and c) heart segmentation results of DLBAS based on FCDN and ABAS by MIM and Mirada compared against ground-truth manual contours

2.2.2 Results

Fig 4 shows segmentation examples from DLBAS and ABAS. In terms of quantiataive accuracy, among 14 CTV structures, DLBAS achieved the highest average DSCs in 11, with significant differences in left and right AXL3 and IMN. HD95 comparisons indicate that DLBAS had smaller surface discrepancies across most CTVs, except for the SCL nodes (Fig. 5A).

ABAS and DLBAS performed similarly for OARs (Fig. 5B), with Mirada's ABAS achieving the highest DSC and lowest HD95 for the lungs and spinal cord, showing significant differences in lung segmentation. DLBAS exhibited larger inter-subject variations in the left lung and spinal cord (Fig. 5B) but outperformed for the thyroid and esophagus, with significant differences in esophagus segmentation against MIM's ABAS.

For heart structures, DLBAS had the highest DSC in five of seven structures, with significantly better results for the heart and right ventricle. HD95 comparisons further confirmed DLBAS's superiority with lower surface distances and smaller inter-subject variations (Fig. 5C). Artery segmentation (RCA and LAD) was suboptimal across all methods, with Mirada's ABAS failing to contour RCA in most cases.

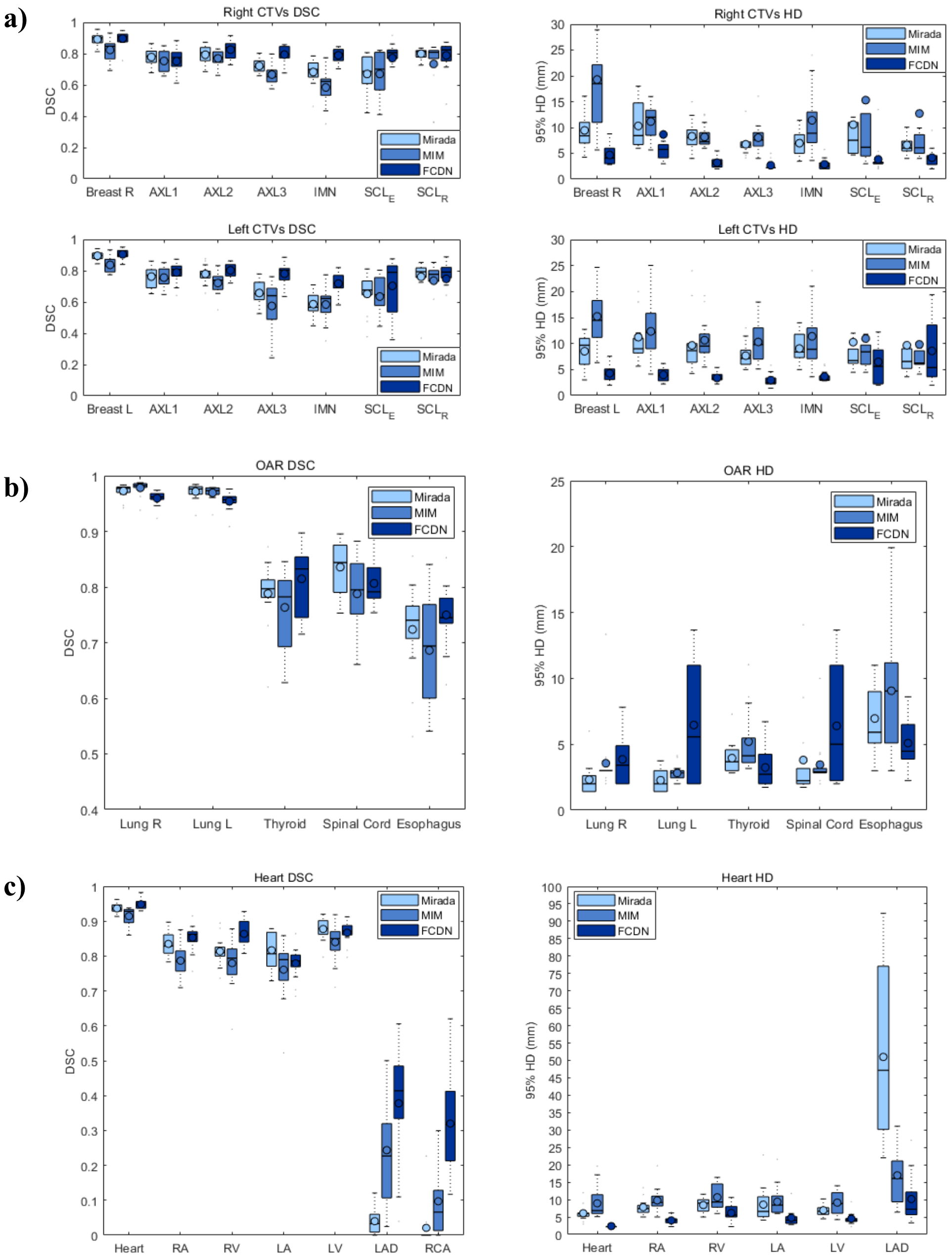


Fig 5. Box-plots of Dice Similarity Coefficients (DSC) and 95% Hausdorff Distance (HD95) in the a) CTVs, b) OARs, and c) Heart structures obtained from Mirada, MIM, and DLBAS based on FCDN using the manual contours as reference.

2.3. The Role of Automated Segmentation Models in Multi-center Study

2.3.1 Materials and Methods

This study utilized two retrospective datasets from left-sided breast cancer patients: Case 1 was a patient data with T1cN1M0 (1.5 cm, triple-negative, grade 3) post-breast-conserving surgery, and Case 2 was a patient data with T3N1M0 (9.5 cm, luminal A type, grade 2) post-mastectomy with implant-based reconstruction. These cases, along with CT and MRI scans and clinical information for radiation therapy, were distributed to 31 institutions across South Korea for analysis by participating investigators.

The study was conducted in two phases: In Phase 1, participants were asked to contour from scratch without the assistance of auto-segmentation (Fig 6). The European Society for Radiation and Oncology consensus guideline [34,35] was suggested to aid the contouring of CTVs (CTV axillary levels 1, 2, and 3 [CTVn_L1, 2, 3], intramammary node [CTVn_IMN], supraclavicular node [CTVn_SCL or CTVn_L4], and CTV [CTVp_breast]); however, clinical discretion was allowed based on their experience and knowledge. The planning target volume (PTV) was generated using a non-isotropic geometrical expansion based on the participants' institutional policy. OARs included the heart, contralateral breast (CLB), thyroid, esophagus, spinal cord, left and right lungs (Lung R, L), and left anterior descending artery (LAD).

Auto-contour sets containing target CTVs and OARs were generated on the test cases (i.e., Cases 1 and 2) using a previously published in-house DL model that has been used in clinics since 2020 [36]. The DL model used in this study had previously been tested on both internal and external cohorts of breast cancer patients, demonstrating robust performance in left-sided, right-sided, and bilateral breast cancer. The model was chosen because of the limited availability of contour tools that encompass all the CTVs used in this study. Six months after phase 1, the same participants took part in phase 2. In Phase 2, participants were instructed to use auto-contour sets, but were given flexibility to deviate from them if they disagreed with their quality or found them uncomfortable to use. Phase 1 focused on measuring participants' manual contouring, while Phase 2 assessed changes in the final segmentation after auto-segmentation was applied.

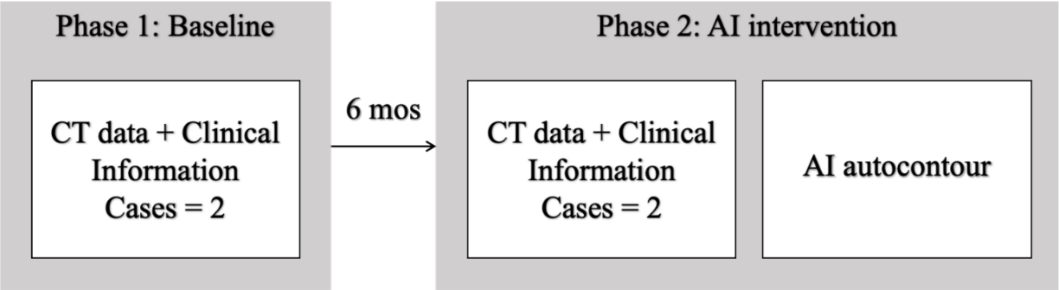

Phase 1: Baseline
CT data + Clinical Information Cases = 2
6 mos
Phase 2: AI intervention
CT data + Clinical Information Cases = 2
AI autocontour

Fig 6. Overall Study Design. Abbreviations: CT = computed tomography; mos = months; AI = artificial intelligence.

Contour discrepancies were measured using DSC, surface DSC, and HD, comparing all observer pairs and the consensus contour. A 3D heatmap visually confirmed the adjustments, with each point on the participant's surface compared to the reference surface. Adjustments were represented by a color map, ranging from red (maximum outward expansion of 10 mm) to blue (maximum inward shrinkage of 10 mm). For qualitative evaluation, a questionnaire was sent to all observers in the study with the following questions: "How much time did it take to complete the contours for each phase?" (Options: "<30 min," "30-60 min," ">1 h"); "How would you rate the auto-contour quality?" (5-point scale: 1 = not usable, 5 = no edits needed); "Do you think auto-contouring will help reduce IOV in the future?" (5-point scale: 1 = strongly disagree, 5 = strongly agree); and "How much auto-contour did you use in Phase 2?" (5-point scale: 1 = not at all, 5 = very much).

We included both two-dimensional (2D) and three-dimensional (3D) heat maps to visualize interobserver agreement and areas of manual edits with respect to the edited auto-contour (reference contour). A radiation oncologist with 9 years of experience edited the auto-contour sets, and an independent panel of three radiation oncologists it as a reference. The 2D heatmap shows variations among observers, with values ranging from 0 to 31 (Supplementary Fig. 3). The areas with the greatest and least overlap are indicated in red and blue, respectively. A three-dimensional heatmap was created to show the average adjustment of the participants projected on the reference shape of each OAR. The nearest point on the participant's 3D surface was determined using reference 3D surface. Subsequently, we determined whether the point was outside or inside the closed reference surface. Depending on the degree of adjustment, each point was represented by a colour map ranging from red (i.e., maximum outward expansion of 10 mm) to blue (i.e., maximum inward shrinkage of 10 mm).

2.3.2 Results

A total of 31 and 30 institutions participated in phases 1 and 2, respectively. Participants contoured 15 structures in two cases, resulting in 930 and 870 paired comparisons per structure. Phase 2 showed improved IOV and better alignment with the consensus contour (Table 1). Surface DSC was lower than DSC for CTVs but higher for smaller structures like the thyroid and LAD. HD decreased across all structures, with LAD DSC improving from 0.44 to 0.61. CTVs showed greater IOV than OARs, except for the LAD. The DL model closely matched the consensus but had lower similarity for structures like the LAD (DSC 0.19 vs. consensus 0.50) and spinal cord (DSC 0.66 vs. consensus 0.76).

Phase 2 demonstrated stronger interobserver agreement than Phase 1, as indicated by smaller blue regions in the CTVs (Fig. 7). The percentage of high-agreement areas in the CTV breast increased from 8.7% to 16.2% in case 1 and from 9.9% to 25.0% in case 2. Similar trends were observed in CTVn_L2 and CTVn_SCL (Fig 7). In case 2, fewer observers included the central portion of the breast implant in Phase 2 compared to Phase 1. While OARs also showed improvement, the changes were less pronounced (Supplementary Fig. 6).

Three-dimensional heatmaps revealed fewer user contour edits in Phase 2, with reduced red and blue regions indicating outward and inward modifications (Fig. 8). Contours in Phase 2 showed closer alignment with the reference, particularly in CTVn_L1 and CTVn_IMN. In Phase 1, physicians more often extended CTVn_L1 toward the skin, but this was significantly reduced in

Phase 2. Heatmaps also showed greater variation in CTVn_L2 and CTVp_breast starting points in Phase 1.

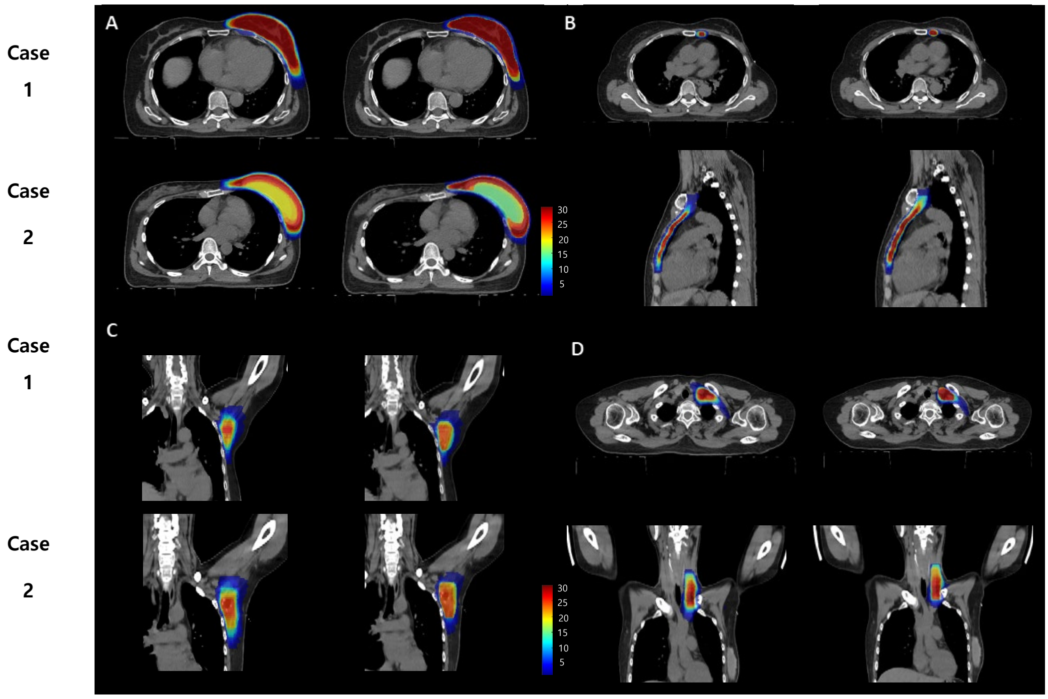


Fig 7. CT with contour variation heatmaps overlaid for Cases 1 and 2 in Phases 1 and 2. (A) CTVp_breast, (B) CTVn_ IMN, (C) CTVn_L1, (D) CTVn_SCL. Abbreviations: CT = computed tomography; CTVp_breast = clinical target volume; CTVn_IMN = intramammary node; CTV

Table 1. Quantitative evaluation through interobserver comparison (a) and with reference to consensus contour (b). Abbreviations: P1=Phase 1, P2=Phase 2, DSC = dice similarity coefficient; SD = standard deviation; HD = Hausdorff distance; CTVn_L1 = CTV axillary level 1; CTVn_L2 = CTV axillary level 2; CTVn_L3 = CTV axillary level 3; CTVn_IMN = intramammary node; CTVn_SCL = supraclavicular node; CTVp_breast = clinical target volume; CLB = contralateral breast; Lung R = right lung; Lung L = left lung; LAD = left anterior descending artery

	(a) Interobserver Comparison						(b) Comparison to consensus contour					
	DSC (± SD)		Surface DSC (± SD)		HD (± SD)		DSC (± SD)		Surface DSC (± SD)		HD (± SD)	
	P1	P2	P1	P2	P1	P2	P1	P2	P1	P2	P1	P2
CTVn_L1	0.58 ± 0.13	0.67 ± 0.22	0.43 ± 0.14	0.55 ± 0.25	42.96 ± 29.29	31.65 ± 33.63	0.64 ± 0.12	0.73 ± 0.17	0.44 ± 0.14	0.58 ± 0.18	27.93 ± 18.27	19.89 ± 23.64
CTVn_L2	0.51 ± 0.17	0.70 ± 0.25	0.44 ± 0.16	0.64 ± 0.25	46.22 ± 35.59	24.83 ± 24.20	0.52 ± 0.20	0.65 ± 0.21	0.46 ± 0.19	0.60 ± 0.20	56.12 ± 32.07	36.45 ± 27.58
CTVn_L3	0.47 ± 0.14	0.55 ± 0.23	0.45 ± 0.14	0.54 ± 0.23	41.47 ± 38.70	28.72 ± 28.12	0.49 ± 0.18	0.61 ± 0.20	0.47 ± 0.18	0.60 ± 0.21	35.56 ± 30.95	21.31 ± 23.80
CTVn_IMN	0.52 ± 0.13	0.61 ± 0.16	0.65 ± 0.16	0.72 ± 0.16	35.46 ± 29.96	18.78 ± 19.67	0.49 ± 0.15	0.64 ± 0.15	0.59 ± 0.19	0.74 ± 0.19	31.74 ± 27.27	13.84 ± 15.66
CTVn_SCL	0.51 ± 0.14	0.62 ± 0.20	0.40 ± 0.14	0.52 ± 0.21	48.26 ± 32.26	38.47 ± 34.94	0.30 ± 0.14	0.32 ± 0.13	0.29 ± 0.11	0.34 ± 0.10	96.30 ± 38.43	105.48 ± 42.84
CTVp_breast	0.75 ± 0.12	0.80 ± 0.13	0.59 ± 0.16	0.72 ± 0.20	22.52 ± 16.91	16.01 ± 20.22	0.73 ± 0.13	0.81 ± 0.14	0.64 ± 0.16	0.79 ± 0.18	18.02 ± 11.95	11.99 ± 17.85
Heart	0.90 ± 0.05	0.95 ± 0.03	0.68 ± 0.14	0.82 ± 0.12	16.36 ± 12.21	8.38 ± 6.06	0.92 ± 0.05	0.95 ± 0.02	0.73 ± 0.16	0.84 ± 0.11	12.36 ± 8.84	7.21 ± 4.54
Contralateral breast	0.81 ± 0.06	0.89 ± 0.10	0.61 ± 0.17	0.79 ± 0.21	21.70 ± 15.67	15.22 ± 25.78	0.84 ± 0.06	0.92 ± 0.08	0.67 ± 0.19	0.87 ± 0.18	15.57 ± 11.16	9.17 ± 20.11
Thyroid	0.75 ± 0.12	0.79 ± 0.12	0.86 ± 0.11	0.89 ± 0.11	9.72 ± 15.83	7.03 ± 15.28	0.79 ± 0.10	0.82 ± 0.08	0.90 ± 0.09	0.92 ± 0.07	8.10 ± 12.29	5.47 ± 10.98
Esophagus	0.77 ± 0.06	0.81 ± 0.07	0.89 ± 0.07	0.91 ± 0.06	31.62 ± 59.29	8.52 ± 17.51	0.81 ± 0.05	0.83 ± 0.04	0.92 ± 0.05	0.94 ± 0.04	15.70 ± 39.07	5.18 ± 10.86
Spinal cord	0.68 ± 0.12	0.79 ± 0.14	0.80 ± 0.13	0.89 ± 0.11	112.88 ± 111.66	33.93 ± 64.35	0.76 ± 0.09	0.82 ± 0.07	0.87 ± 0.09	0.94 ± 0.10	65.09 ± 86.21	18.27 ± 44.42
Lung R	0.97 ± 0.01	0.98 ± 0.01	0.95 ± 0.03	0.97 ± 0.03	3.86 ± 2.50	2.56 ± 1.80	0.97 ± 0.01	0.98 ± 0.01	0.96 ± 0.02	0.97 ± 0.02	2.76 ± 1.18	1.81 ± 0.86
Lung L	0.87 ± 0.01	0.98 ± 0.02	0.95 ± 0.03	0.97 ± 0.03	3.63 ± 1.98	2.35 ± 1.50	0.97 ± 0.01	0.98 ± 0.01	0.96 ± 0.02	0.97 ± 0.01	2.86 ± 0.84	2.15 ± 0.80
LAD	0.44 ± 0.21	0.61 ± 0.18	0.71 ± 0.21	0.80 ± 0.15	52.14 ± 59.74	14.08 ± 15.30	0.50 ± 0.16	0.59 ± 0.01	0.77 ± 0.14	0.81 ± 0.10	39.03 ± 51.87	12.96 ± 13.48

In Phase 1, most participants took 30–60 minutes to complete contours, while in Phase 2, over half finished in under 30 minutes (Fig. 9A). For target structures, minor edits were most common (53.2%), followed by major edits (36.2%), mostly acceptable (6.4%), and not usable (4.3%) (Fig.9B). Observers were generally satisfied with auto-contours for OARs, though none rated them as perfect. AI-generated contours were positively correlated with the contour assessment score (r = 0.88) and future usefulness score (r = 0.82) (Fig. 9C).

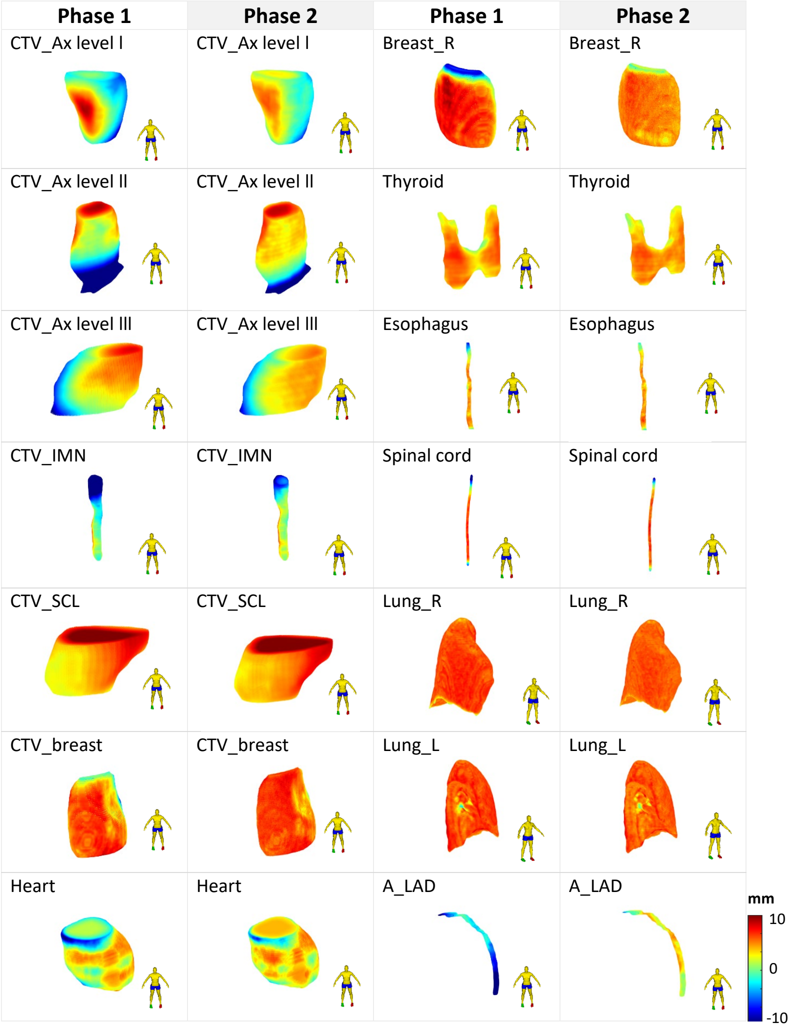


Fig 8. A three-dimensional projection of average adjustments for Phases 1 and 2 compared to the reference contour. Adjustments are on the scale of -10mm to 10mm where positive indicates an outward adjustment.

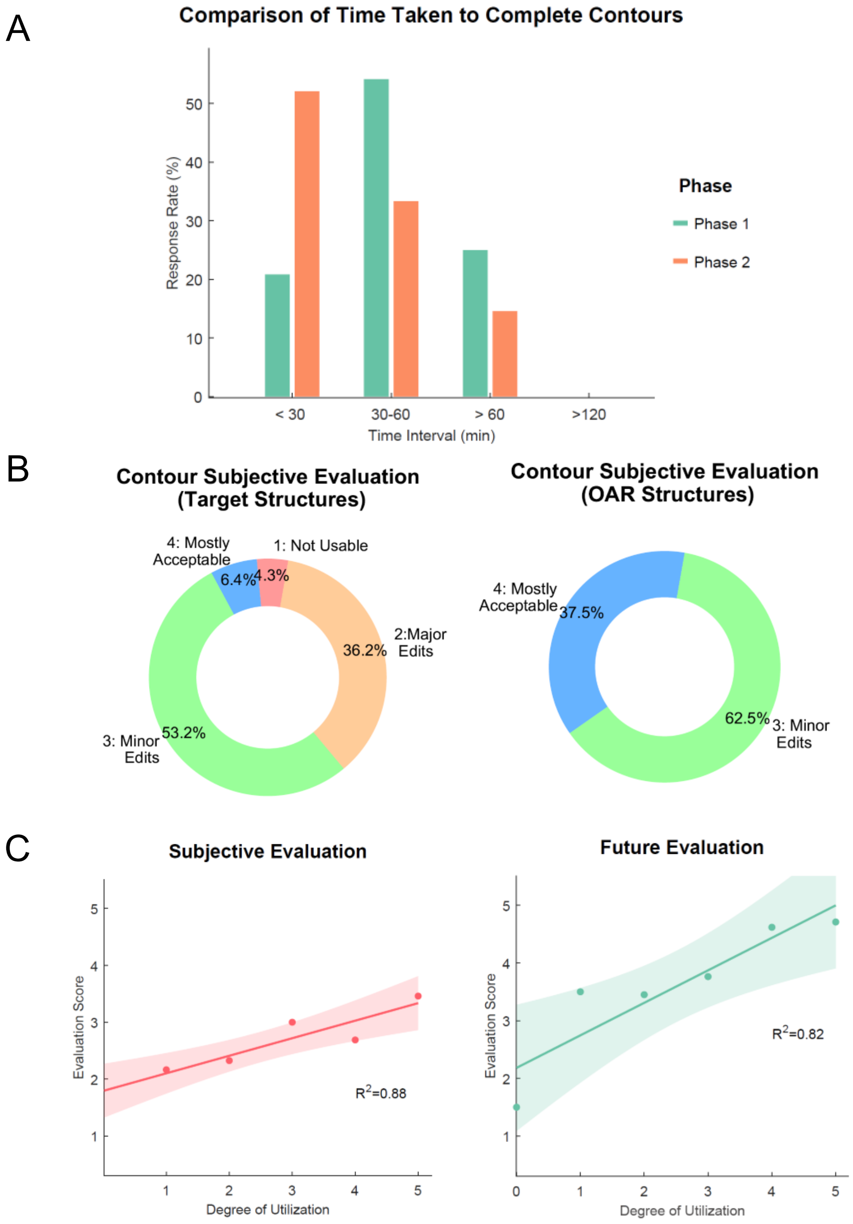


Fig 9. (A) Time comparison, (B) Subjective evaluation, (C) Left: Auto-contour usage vs. evaluation scores (Cases 1 & 2); Right: Auto-contour usage vs. future applicability scores

2.4. Discussion and Conclusion

In this chapter, we examined the impact of deep learning-based auto-segmentation (DLBAS) on both routine clinical workflows and multi-center RTQA. To the best of our knowledge, our studies were one of the first attempts to report on the use of DLBAS in breast cancer radiation therapy planning that supports the integration of DLBAS in radiation oncology.

In the first part, we demonstrated the efficacy of DLBAS by evaluating its performance across various structures (CTVs, OARs, and heart substructures) in a direct comparison with commercial ABAS solutions. While ABAS showed acceptable performance, DLBAS provided more robust and reliable segmentation with results closely aligned to the ground truth. Our results highlight three key findings for DLBAS. First, DLBAS effectively learns the characteristics of complex and low-contrast anatomy, making it ideal for CTV delineation, where expert knowledge is often crucial. In contrast, ABAS, based on landmark detection, has limitations in this regard. Second, DLBAS showed significantly smaller contouring differences with the ground-truth (measured in millimeters by HD95), underscoring its clinical impact. HD95 is a more reliable metric than DSC, as it is less influenced by contour volume size and better reflects the accuracy of contour outlines. We believe DLBAS could reduce contouring time, particularly for CTVs and heart substructures. Third, DLBAS demonstrated superior robustness on non-contrast CT samples, with smaller DSC discrepancies compared to ABAS, indicating that DLBAS is less reliant on input type and more adaptable to diverse clinical protocols.

DLBAS shows great potential in overcoming challenges in radiation therapy planning, particularly in addressing the consistency issues of manual contouring across institutions and large-scale studies. Variations in target volume delineation, as seen in clinical trials, can result in safety concerns and treatment toxicity. For example, a study in South Korea found significant heterogeneity in breast IMRT plans across institutions. By generating consistent contours, DLBAS could mitigate these issues. However, implementing DLBAS requires initial investments in model development, patient data collection, and expert labeling. Once established, DLBAS can greatly reduce the time spent on repetitive contouring tasks, freeing up resources for other clinical activities.

Our study also demonstrated a significant reduction in IOV in OARs and various CTVs through targeted interventions. In particular, we assessed IOV in contouring CTVs and OARs and showed that DLBAS improved inter-observer variations across structures, increasing the average DSC from 0.69 to 0.77. Auto-segmentation saved time and provided a quality benchmark but raised concerns about generalizability, manual corrections, and user acceptance. A key aspect of our approach was the local assessment of contour editing, combined with the classification of observers' contouring preferences and styles. By aligning participants' contours with auto-contour reference shapes in both 2D and 3D, we identified common anatomical regions where adjustments were frequently made. Additionally, cluster analysis based on the volume of different CTVs allowed us to group participants according to their contouring styles. This insight is valuable for providing feedback on unacceptable protocol deviations and can aid local investigators in refining the contouring process for enrolled patients.

In conclusion, our studies have successfully demonstrated the feasibility of incorporating DLBAS solutions into the RT planning workflow. We assessed the performance of DLBAS in comparison to conventional contouring methods, and our findings clearly highlight its superiority in terms of accuracy, efficiency, and consistency. By leveraging deep learning algorithms, DLBAS not only enhances the precision of contour delineation but also streamlines the RT planning process, offering a significant advantage over traditional manual approaches. Furthermore, our comprehensive evaluation of the impact of DLBAS on interobserver variability across multiple institutions reveals that the adoption of deep learning-based auto-contouring technology leads to a substantial improvement in contour agreement. This improvement was observed both qualitatively and quantitatively for critical structures such as OARs and CTVs. By reducing interobserver variability, DLBAS not only enhances the reliability of RT plans but also ensures more consistent and personalized treatment planning, paving the way for more accurate and effective radiation therapy in clinical practice.

3. Clinical Evaluation of Breathing Management Techniques for Toxicity Reduction

3.1. Background

Radiation therapy (RT) is a critical component in breast cancer treatment. However, it can expose the heart to radiation, increasing the risk of cardiac diseases. Minimizing cardiac doses without compromising therapeutic benefits is a priority in breast cancer RT. Deep inspiration breath-hold (DIBH) is the standard technique for cardiac sparing in left-sided breast cancer, as it increases the distance between the heart and the radiation field [37]. However, DIBH requires additional time and resources and may introduce intrafraction organ motion uncertainties [38].

Recently, continuous positive airway pressure (CPAP), originally used for sleep apnea, has shown potential in reducing lung tumor motion during RT [39]. Studies on CPAP in three-dimensional conformal RT (3D CRT) for left-sided breast cancer have demonstrated reduced cardiac radiation exposure by allowing a more caudal heart displacement [40,41]. A recent prospective study further confirmed CPAP's feasibility for motion management and its dosimetric benefits.

However, unresolved challenges remain, limiting its clinical adoption as a standard cardiac-sparing technique in breast cancer RT. Therefore, the goal of this study was to determine the inter- and intrafractional reproducibility of using CPAP as a breath-hold managing tool for volumetric modulated arc therapy (VMAT) for left-sided breast cancer. We also investigated the feasibility of CPAP as a heart sparing technique in more than 200 consecutive patients with left-sided breast cancer who underwent adjuvant RT between June 2020 and January 2021.

3.2. Materials and Methods

3.2.1 Study design and patient characteristics

This retrospective study was approved by the institutional review board of Yonsei Cancer Center (2020-4417-001). The need for written informed consent was waived owing to the retrospective nature of the study. To explore the potential of CPAP from various aspects, the study was conducted to assess 3 main factors: (1) interfractional reproducibility, (2) intrafractional reproducibility, and (3) real-world application. The overall study design and data set used for each part are displayed in Table 2.

Table 2. Overview of data sets used in the study; Abbreviations: 4D CT = 4-dimensional computed tomography; BC = breast cancer; CBCT = cone beam computed tomography; FB = free-breathing; LC = lung cancer; pCT = planning computed tomography; RWD = real-world data set.

Study name	Dataset1 (N=20): BC, pCT, and CBCT	Dataset 2 (N=20): LC and 4D CT	Dataset 3 (N=237): BC and RWD
Interfractional reproducibility	✓		
Dosimetric and volumetric comparison with FB-based plans	✓		
Intrafractional reproducibility		✓	
Evaluation of continous positive airway pressure implementation into routine practice			✓

3.2.2 Assessment of interfractional reproducibility

The first part of this study evaluated the interfractional reproducibility of continuous CPAP in increasing the distance between the heart and the planning target volume (PTV) for left-sided breast cancer patients undergoing VMAT. Data from 20 patients were analyzed using CT and daily cone beam CT (CBCT) across 15 treatment fractions. Initial simulations were conducted with free breathing (FB) and CPAP, with CPAP pressure starting at 12 cmH_2O and later increased to 15 cmH_2O to enhance effectiveness.

Heart position reproducibility was assessed using the minimum heart distance (MinHD), defined as the shortest perpendicular distance between the heart and the PTV along a virtual posterior field edge (Fig. 10). MinHD measurements were taken from planning CT (pCT) and daily CBCT images, with MinHD error calculated as the difference between the two. The PTV was contoured according to ESTRO guidelines, and findings aimed to determine the consistency of CPAP's heart-sparing effects throughout treatment.

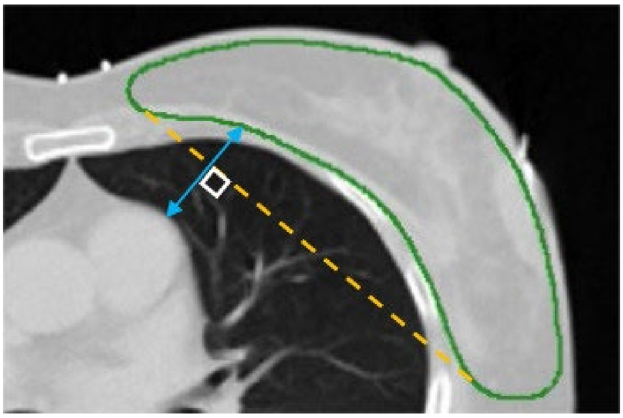


Fig 10. Schematic of MinHD (blue arrow) measuring the distance from heart to PTV (green). The line is perpendicular to the virtual field edge (yellow dotted), posteriorly tangent to the PTV.

3.2.3 Dosimetric and volumetric comparisons with FB

We then compared CPAP-based and FB-based VMAT plans to assess CPAP's heart-sparing capability. Experienced dosimetrists created new FB-based plans using the same dosimetric criteria as the original CPAP-based plans. Cardiac substructures, including the left anterior descending artery (LAD), left atrium, and left ventricle, were segmented using auto-contouring software (Aview; Coreline Soft, Seoul, South Korea). Plans were generated in RayStation (version 5; RaySearch Laboratories, Stockholm, Sweden), with 30% covering only the left breast, 25% as simultaneous integrated boost plans, 30% including regional node irradiation, and 15% targeting the supraclavicular node, internal mammary lymph node (IMN), and whole breast. In over 90% of cases, the prescribed dose was approximately 40.05 Gy for 95% of the PTV, delivered using two beams with start and stop angles of 300° to 170° and 170° to 300°, respectively.

The dosimetric metrics included the mean heart dose (MHD), heart volumes receiving 1, 2.5, 5, and 10 Gy (V1, V2.5, V5, V10), mean lung dose, and lung volumes receiving 5 to 40 Gy (V5–V40). Additional measures for cardiac substructures included LAD maximum dose (Dmax) and mean dose (Dmean) and V5 for the left ventricle and left atrium. Dose differences between CPAP and FB-based plans were analyzed using a two-tailed t-test with 95% significance.

3.2.4 Assessment of intrafractional reproducibility

Next, we assessed CPAP reproducibility during treatment by analyzing 4D CT scans from 20 female lung cancer patients who underwent external RT, as these scans are not routinely performed for left-sided breast cancer patients. Intrafractional PTV motion was measured using mutual information-based rigid image registration with Insight ToolKit, comparing the 0% phase (reference) to 10%-90% phases. Breast target movement was analyzed in the craniocaudal (CC), anteroposterior (AP), and mediolateral (ML) directions, while diaphragm motion in the CC direction was manually measured for each patient.

3.2.4 Assessment of feasibility of CPAP in routine practice

We investigated the feasibility of using CPAP in routine clinical practice using data set 3. This data set included 237 patients with left-sided breast cancer who underwent adjuvant RT between June 2020 and January 2021, excluding the patients in data set 1. The number of patients who completed CPAP-based treatment and the reasons for exclusion were retrospectively identified. The patients who received CPAP-based VMAT were divided into 2 subgroups based on whether or not they received IMN irradiation (IMNI), and the dosimetric parameters (MHD, V1, V2.5, V5, and V10) were examined.

3.3. Results

3.3.1 Interfractional reproducibility

Starting with the inter-fractional reproducibility, Fig 11 shows box plots of the minimum heart distance (MinHD) from planning CT (MinHDpCT, asterisks) and daily CBCTs (MinHDCBCT, box plots), with blue and orange boxes representing high and low pressure groups, respectively. The difference between MinHDpCT and the median MinHD from 15 daily CBCTs was under 1 cm, indicating reproducible heart sparing, and pressure type did not affect MinHD.

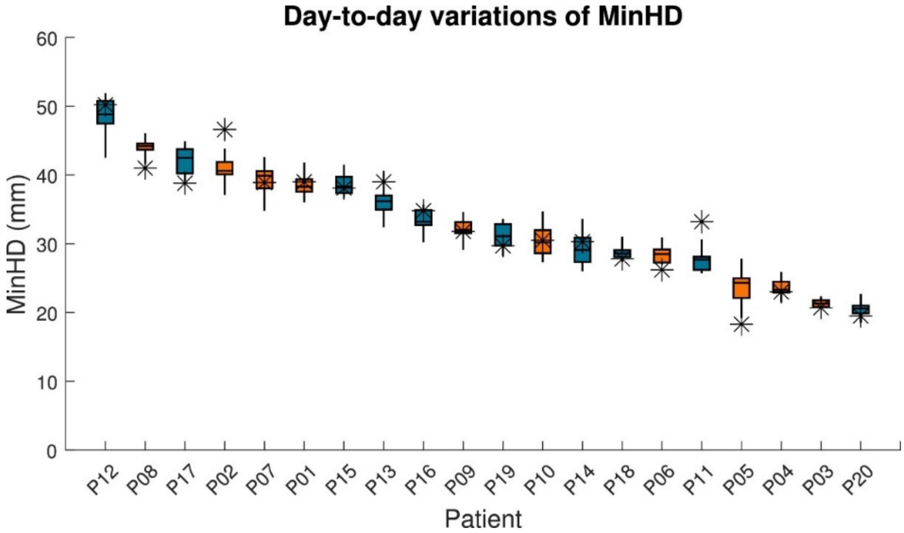


Fig 11. Box plots of MinHD: MinHDpCT (baseline CT) shown as asterisks; MinHDCBCT (daily CBCTs) as box plots. Blue and orange indicate high and low pressure groups, respectively. Abbreviations: CBCT = cone beam computed tomography; CT = computed tomography; MinHD = minimum heart distance

In terms of intrafraction motion, Fig 12 shows mean intrafraction breast motion: 2.5 ± 2.0 mm in the craniocaudal (CC), 1.8 ± 1.4 mm in the anteroposterior (AP), and 0.5 ± 0.5 mm in the ML directions (Fig 12). The greatest motion and interpatient variation occurred in the CC direction, with outliers at 8 mm (CC) and 6 mm (AP) for patient 16. Diaphragm motion ranged from 8 to 24 mm, with no clear correlation to PTV motion amplitude.

3.3.2 Dosimetric and volumetric comparison

The mean dose of the whole heart was significantly smaller for the CPAP plan than for the FB plan in both the IMNI and no IMNI groups (Table 3). The heart V1, V2.5, and V5 were also significantly smaller in the CPAP-based plan, whereas there was no significant reduction in V10.

No significant differences in the mean lung dose were found between CPAP and FB. Further, the average lung volumes receiving 5 to 20 Gy (ie, V5 to V20) were comparable between FB and CPAP. As for the cardiac substructures, the maximum doses of the LAD, left ventricle V5, and mean dose of the left ventricle were significantly reduced in CPAP.

When comparing CPAP-based plans to FB-based plans, CPAP resulted in slightly smaller heart volumes but significantly larger lung volumes. Additionally, lung expansion increased, on average, 1.7 times less in the low-pressure group (P1-10) compared to the high-pressure group (P11-20).

3.3.3 Intrafractional reproducibility

The mean intrafraction breast motion for the total cohort was 2.5 ± 2.0, 1.8 ± 1.4, and 0.5 ± 0.5 mm in the CC, AP, and ML directions, respectively (Fig 12). The greatest movement was observed in the CC direction, which also had the greatest interpatient variations. Although most patients had motion <5 mm, there were outliers in the CC and AP directions at 8 and 6 mm, respectively, for patient 16 of the lung patient cohort. For the diaphragm, relatively large motions ranging from 8 to 24 mm were observed. There was no clear correlation between the motion amplitude of PTV and diaphragm motion. For instance, although P01 and P16 both had small diaphragm motions, 1 had small PTV motion amplitude, whereas the other had a large value.

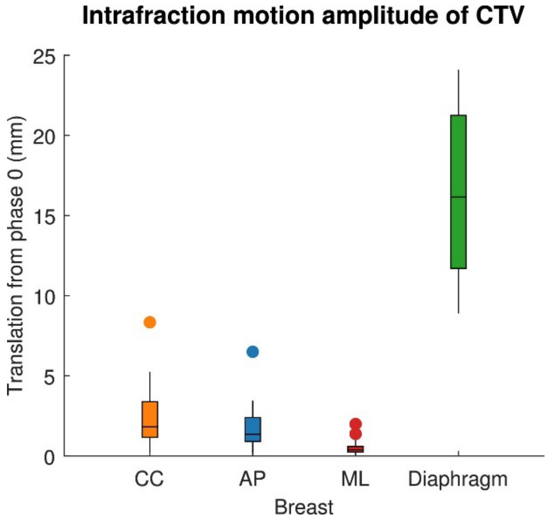


Fig 12. Comparison of the intrafraction motion amplitudes between the breast PTV and lung diaphragm. Compared with the diaphragm, the movements in the CC, AP, and ML directions were relatively small, implying that CPAP managed to keep the breast motion minimal despite the large movement of the diaphragm. Dots indicate the datapoints that exceed 1.5 times the interquartile range. Abbreviations: AP = anteroposterior; CC = craniocaudal; CPAP = continuous positive airway pressure; ML = mediolateral; PTV = planning target volume.

Table 3. Comparison of average OAR volumes and dosimetric parameters between the FB- and CPAP-based VMAT plans.

	IMNI			No IMNI		
	FB	**CPAP**	**p-value**	**FB**	**CPAP**	**p-value**
Heart						
Volume (cc)	517.8	513.6	0.8	497.7	477.9	0.1
Dmean (Gy)	2.6	2.0	**< 0.01**	1.6	1.3	**< 0.01**
V1 (%)	90.3	82.2	**< 0.01**	66.8	53.8	**< 0.01**
V2.5 (%)	34.1	20.7	**< 0.01**	13.1	7.5	**< 0.01**
V5 (%)	8.4	3.8	**< 0.05**	2.0	1.2	**< 0.05**
V10 (%)	1.7	0.5	0.1	0.1	0.1	0.6
Ipsilateral lung						
Volume (cc)	1049.5	1560.0	**< 0.01**	1153.4	1709.6	**< 0.01**
Dmean (Gy)	6.9	6.5	0.1	4.8	4.6	0.2
V5 (%)	35.9	34.0	0.1	25.0	24.3	0.5
V10 (%)	21.9	21.0	0.2	14.8	13.9	0.2
V20 (%)	10.1	9.8	0.6	6.0	5.4	0.1
LAD						
Dmax (Gy)	8.5	6.0	**< 0.01**	10.1	5.7	**< 0.05**
Left atrium						
Volume (cc)	54.9	49.8	**< 0.05**	52.2	44.3	**< 0.05**
Dmean (Gy)	1.5	1.5	0.6	0.9	0.8	0.4
V5 (%)	0.6	0.4	0.7	0	0	-
Left ventricle						
Volume (cc)	141.1	127.6	**< 0.05**	130.3	116.5	**< 0.05**
Dmean (Gy)	2.1	1.7	**< 0.01**	1.7	1.2	**< 0.01**
V5 (%)	1.4	0.3	**< 0.05**	0.8	0.0	0.1

3.3.4 Feasibility of CPAP in routine practice

The CPAP-based RT was successfully acclimated in 221 of the 237 patients (93%). The use of CPAP was not evaluated in 7 patients owing to low compliance (n = 5) and old age (n = 2). CPAP was not used in 8 patients because 2D fluoroscopy (n = 3) or 3D-CT/planned dose distributions (n = 5) showed no benefit. Only 1 patient failed to tolerate breathing with CPAP. Among the patients for which CPAP was used, 116 patients were treated with IMNI. The mean dose, V1, V2.5, V5, and V10 of the heart in the 221 patients were 1.6 ± 0.7 Gy, 62.5 ± 32.7%, 14.8 ± 14.7%, 2.6 ± 4.0%, and 0.4 ± 0.9%, respectively. All dosimetric parameters were higher in the IMNI group than in the no IMNI group (Fig 13). The median MHDs of the IMNI and no IMNI groups were 1.90 and 0.8 Gy, respectively.

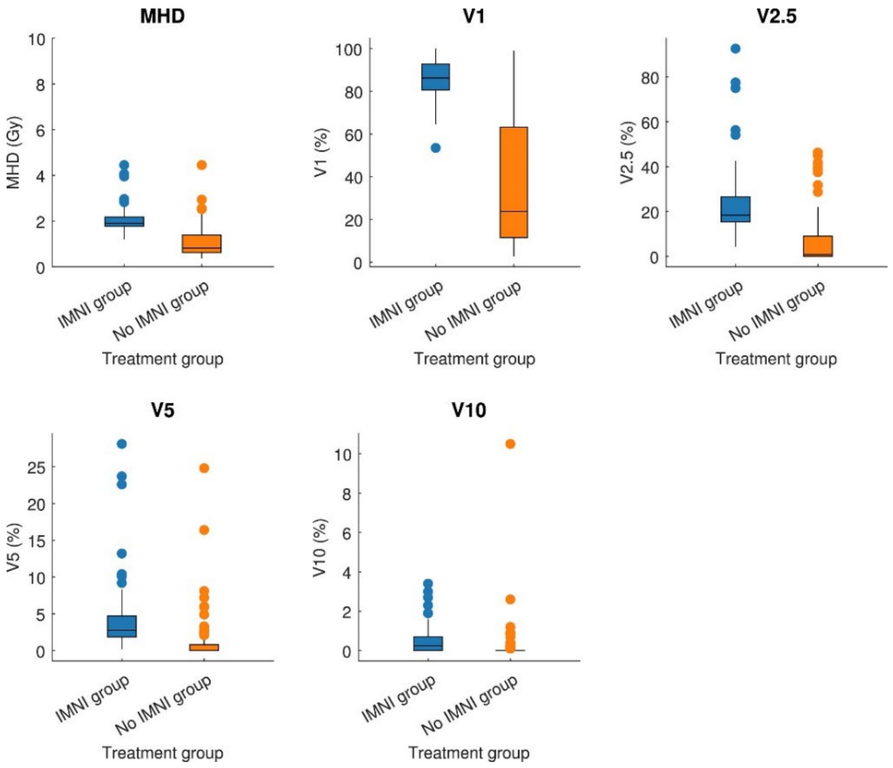


Fig 13. Dose summary of the 221 patients who underwent CPAP according to the IMNI group. Although there are differences depending on the presence of IMN, the overall median MHD was observed to be less than 2 Gy, which is comparable to the MHD of data set 1, confirming the practicality of CPAP across a large cohort. The difference between the IMN group and the no IMN group was the highest for V1, which then decreased gradually as the dose threshold increased (ie, V2.5, V5.0, V10). Dots indicate the datapoints that exceed 1.5 times the interquartile range. Abbreviations: CPAP = continuous positive airway pressure; IMNI = internal mammary lymph node irradiation; MHD = mean heart dose.

3.4. Discussion and Conclusion

This study demonstrated that CPAP has good intra- and intersession reproducibility and is a feasible heart-sparing technique during RT in patients with breast cancer who will undergo DIBH. To the best of our knowledge, this is the first study to evaluate the reproducibility and the largest study to evaluate the heart-sparing capability of CPAP-based VMAT for left-sided breast cancer treatment.

The efficacy of breath management techniques is greatly reliant on their reproducibility. The measures for evaluating positional reproducibility for DIBH have been extensively reported [42–46]. Comsa et al[45] reported a maximum heart shift of 6.2 mm with respect to the chest wall using CBCT images. Another study found a median interfraction heart shift of 1, 0, and 1 mm in the CC, AP, and ML, respectively[46]. Our results (average MinHD error of 2 mm) are comparable to those reported for DIBH, confirming CPAP's ability to keep the heart position fairly consistent.

The findings provide instrumental evidence supporting the clinical effectiveness of CPAP for heart sparing in left-sided breast RT. The subcentimeter variations of the heart position observed in this study indicate that the reduced heart dose from CPAP is consistently maintained during the treatment. In addition, the wide range of CPAP pressure (7-20 cmH2O) in previous studies underlines the need to establish the optimal CPAP [39–41,47–49]. In our study, the stability of the breathing pattern did not significantly differ according to the CPAP (≥15 cmH2O vs <15 cmH2O). This indicates that the guiding principle of “as high as achievable” does not increase the motion amplitude and position instability, although our study did not test whether these similar findings on reproducibility would be observed at <12 cmH2O.

The CPAP-based plans generated more favorable dosimetric outcomes than FB for dose/volume parameters across all structures investigated. As radiation exposure to cardiac substructures is correlated with subsequent cardiac morbidity, our findings that CPAP reduces the dose to the whole heart and the substructures have important clinical implications. However, the clinical relevance of these differences is yet to be elucidated. In previous literature, the effect of CPAP on target volume coverage and doses to organ-at-risks was investigated in 3D-CRT using partial wide tangents, VMAT, and proton beam therapy in conventional or hypofractionated regimens[50]. Compared with corresponding FB conditions, the MHD was lower in every technique in CPAP, especially when combined with 3D-CRT, where 50% reduction in the MHD was observed. This is comparable to the findings by Allen et al,[40] who showed that the MHD was decreased from 3.0 Gy in FB to 1.6 Gy in CPAP with 3D-CRT. The absolute benefit of CPAP in the present study is smaller than in previous reports, and this could be because of the already reduced MHD from VMAT.

The intrafractional motion for CPAP is yet to be thoroughly evaluated, although there have been a few studies focusing on DIBH. Another study investigated intrafraction tumor motion in early-stage breast cancer using a fiducial marker in FB and reported movements of 1.0 ± 0.9, 1.8 ± 1.5, and 1.3 ± 1.2 mm in the CC, AP, and ML directions, respectively[51]. Our study had comparably small motions (2.5 ± 2.0, 1.8 ± 1.4, 0.5 ± 0.5 mm in the CC, AP, and ML directions, respectively). The lung diaphragm moved about 17.6 ± 6.2 mm on average during the same 4D CT, indicating that increased airway pressure by CPAP does not increase the breast motion, although it may increase the lung diaphragm motion. Our study had an outlier patient with a large upper chest movement regardless of CPAP use (6.5 and 8.3 mm in the AP and CC directions; Video E1). Therefore, possible CPAP-related increases in breast motion should be assessed in advance in every patient.

We achieved a high completion rate of CPAP at 93% in its real-world application in our institution. This is substantially higher than what has been reported for DIBH by Rice et al, [52] in which approximately 43% of 272 patients included in the study were unable to complete DIBH treatment owing to having no obvious advantage, inability to demonstrate good breath-hold, and anxiety. Unlike DIBH, CPAP is less dependent on the patients' ability to hold breath; thus, it is more applicable to a wider range of patients and has greater benefits for future clinical implementation. Our outcomes further confirmed the dosimetric advantages of using CPAP. The MHDs for the 216 patients who did and did not undergo IMNI (data set 3) were 2.0 ± 0.5 Gy and 1.1 ± 0.7 Gy, respectively, comparable to the results from data set 1 (2.0 ± 0.5 Gy and 1.3 ± 0.4 Gy). Furthermore, CPAP may have an advantage over DIBH with respect to treatment time because of high duty cycle in CPAP-based RT. A detailed comparison between DIBH and CPAP should be conducted in a future study.

The following study limitations should be addressed in future studies. First, this study incorporated 3 different types of data sets (data sets 1, 2, and 3), rather than a single common data set. Although the reason this was done was to investigate all possible aspects of CPAP within the limits of using retrospective data, a follow-up study prospectively preparing a common data set may be useful to obtain more in-depth evidence of CPAP applications in breast RT. Second, this study did not include the dosimetric comparison of interfraction CBCT data to further confirm the reproducibility of CPAP with respect to safety. Further studies should investigate the dosimetric changes across different treatment settings. Lastly, our MinHD measurement can be deemed subjective and prone to reproducibility issues, because a slight shift of the positioning could result in a different measurement. We believe that the reproducibility can be further improved if the volume-wise distance can be measured between the heart contour and the PTV contour, instead of a 2D-based approach. Unfortunately, this could not be done in the current study due to the unavailability of contours on individual CBCT scans. In future studies, approaches such as deep learning-based automatic segmentation could be implemented to generate high-quality contours of the heart and PTV on CBCTs to enable 3D minimum heart distance measurement.

In conclusion, this study demonstrates that CPAP is a reproducible and effective heart-sparing technique for left-sided breast cancer radiation therapy, with comparable reproducibility to DIBH. CPAP significantly reduces radiation exposure to cardiac structures, offering favorable dosimetric outcomes over free-breathing and contributing to improved heart protection during treatment. With a high completion rate of 93% in real-world clinical practice, CPAP presents a more accessible alternative to DIBH, particularly for patients who struggle with breath-holding. These findings support the potential of CPAP for broader clinical adoption, though further studies comparing CPAP and DIBH in terms of treatment time and long-term clinical outcomes are needed.

4. Multi-modal Toxicity Prediction Model using a Large Language Model and Image-Based Techniques

4.1. Background

Esophageal cancer (EC) is the sixth leading cause of cancer-related mortality worldwide, with a five-year survival rate of approximately 20% across all stages [53]. Radiation therapy (RT) is one of the treatment modalities for EC, often combined with chemotherapy, which has been shown to improve survival compared to surgery alone [54]. Radiation-induced acute esophagitis (AE) is one of the common complications occurring in patients with EC who receive RT. The symptoms of AE include difficulty swallowing and pain, which may, in turn, significantly impact patients' quality of life during and after treatment. The risk of AE has been reported to increase with higher radiation doses and concurrent chemotherapy [55]. Given the significant impact of AE on patients' quality of life, accurately predicting its occurrence during radiation therapy is essential for effective management and intervention.

Previous efforts have been made to predict AE, although most studies have focused on its occurrence in non-small cell lung cancer (NSCLC). Conventionally, normal tissue complication probability (NTCP) models have been developed for AE prediction [9,56–58] focusing on dosimetric and clinical variables, with common predictors being mean esophageal dose and concurrent chemotherapy. Similarly, machine learning models have been employed to capture more complex representations of patient data, leveraging a broader range of predictors, including dose volume histogram (DVH)-derived parameters and patient-specific characteristics [59,60]. However, a recent study evaluating 35 clinical and dosimetric variables, including conventional metrics like esophageal mean dose and V20–V60, found no reliable predictors for AE [61]. To overcome these limitations, imaging-based approaches have been explored, including radiomics, dosiomics, and deep learning (DL) models [62–64]. The most recent study in esophageal cancer, combining radiomics, dosiomics, and DL in a hybrid approach[64].

Prior studies have explored esophagitis prediction using either imaging data or clinical variables, but none have investigated the potential of integrating both modalities. Imaging data capture rich anatomical and pathological information, whereas clinical variables provide essential context about patient characteristics and treatment factors. A multimodal approach that combines these information may enhance predictive performance. Recent advances in large language models (LLMs) have opened new opportunities for multimodal learning by effectively integrating structured text and visual data, showing promising results in tasks such as medical image segmentation [25] and survival prediction [26]. However, the added complexity of such models may pose challenges in clinical settings, where there is an emphasis on the need for clinical models to be simple and reproducible.

In this study, we aimed to develop a robust multimodal model for predicting AE in patients with EC by integrating imaging features and clinical data. To address the trade-off between performance and model complexity, we developed a multimodal prediction model based on context-aware 3D Swin transformer model that leverages LLM with multi-modal alignment. The goal was to assess whether these multimodal approaches offer improved predictive performance over image-only models. Specifically, we sought to answer two key questions: (1) What is the most effective strategy for combining image-based and clinical information in a multimodal prediction framework? and (2) How does the predictive performance of such models vary across different demographic cohorts?

4.2. Materials and Methods

4.2.1 Data characteristics

This study included two independent datasets: an internal dataset from Memorial Sloan Kettering Cancer Center (MSKCC) and an external test dataset from Gangnam Severance Hospital. The internal dataset was retrospectively collected and comprised of 217 patients with esophageal cancer who previously underwent RT treatment at MSKCC between 2009 and 2022. The dataset was further divided into training (N=197) and internal test cohorts (N=20 for internal test) using a stratified random split. All patients received intensity-modulated radiation therapy (IMRT) or volumetric modulated arc therapy (VMAT) with concurrent chemotherapy. The external test set consisted of 20 esophageal cancer patients treated with Tomotherapy at Gangnam Severance Hospital in 2024. AE was labeled as a binary outcome based on the presence or absence of grade ≥2 toxicity, according to the CTCAE grade system [65].

The data included in this study was imaging data as well as clinical data. The imaging data included planning CT scans, dose distributions, and pre-treatment gross tumor volume (GTV) contours. The baseline clincal data was also collected from the electronic medical records from each institution by radiation oncologists. Patient characteristics are summarized in Table 2. To evaluate potential differences between the training and test cohorts, we performed statistical analyses using appropriate tests based on data type. Specifically, we applied independent t-tests for normally distributed continuous variables, Mann-Whitney U tests for non-normally distributed continuous variables, and Fisher’s Exact or Chi-Square tests for categorical variables, where p-value <0.05 was considered a threshold for statistical significance.

Table 4. Patient Characteristics. P-values were calculated using (a) t-test, (b) Mann-Whitney U, (c) Fisher's Exact (<5 count), and (d) Chi-Square test.

	Train/Val (N=197)	Internal Test (N=20)	External Test (N=20)	Training vs test p-value	Training vs external test p-value	Test vs external test p-value
Age	66.11 ± 10.99	65.05 ± 10.23	65.05 ± 11.05	0.66 [a]	0.68 [a]	0.99 [a]
Subsite				0.64 [d]	0.56 [d]	0.64 [d]
Lower	161	17	9			
Middle	26	2	8			
Upper	10	1	3			
Esophagitis Status						
Positive	94	11	12			
Negative	103	9	8			
Smoking				1.00 [d]	0.09 [d]	0.26 [d]
Smoker (current and previous)	135	14	9			
Non-smoker	62	6	11			
Stage				0.18 [d]	0.21 [d]	0.18 [d]
1	10	-	2			
2	53	3	3			
3	126	4	11			
4	8	13	4			
Sex				0.26 [c]	0.42 [d]	0.13 [c]

Female	45	2	7			
Male	152	18	13			
Diabetes				0.76 [c]	1.00 [c]	0.70 [c]
Positive	34	4	3			
Negative	163	16	17			
Pre-existing heart disease				0.75 [c]	0.05 [c]	0.05 [c]
Positive	32	4	0			
Negative	165	16	20			
Concurrent Chemo				1.00 [c]	0.0001 [c]	0.11 [c]
True	197	20	16			
False	-	-	4			
RT Intent				0.18 [d]	0.12 [d]	0.64 [d]
Preoperative	124	9	8			
Definitive	73	11	8			
Others	-	-	4			
Dose delivered (cGy)	4964.16 ± 233.36	4903.50 ± 400.07	5237.62 ± 552.29	0.89 [b]	0.94 [b]	0.92 [b]
Histology				1.00 [d]	<0.001 [c]	<0.001 [c]
Adenocarcinoma	145	15	-			
Squamous cell carcinoma	52	5	19			
Others	-	-	1			

4.2.2 Clinical feature selection

To identify the most relevant clinical variables as predictors for AE, we utilized a multivariable logistic regression model with Least Absolute Shrinkage and Selection Operator (LASSO) (Fig 13). The following clinical features were included: age, stage, smoking, delivered dose, histology, diabetes, preexisting heart disease, radiotherapy intent, Karnofsky Performance Scale, tumor location as well as the induction chemotherapy status. Among these features, the continuous clinical variables were scaled using the Standard Scaler, and categorical variables were one-hot encoded before normalization to ensure compatibility with the regression model. LASSO penalizes the regression coefficients of the variables, allowing for the selection of variables with non-zero coefficients [7]. To enhance model robustness, this procedure was repeated across 1,000 bootstrap samples, each utilizing an 80/20 random train-test split. Variables with non-zero coefficients were retained, and those selected in more than 50% of the bootstrap iterations were included in the final model. Finally, the clinical features that were selected in over 50% of the bootstrap iterations were selected for inclusion for the final model training and testing.

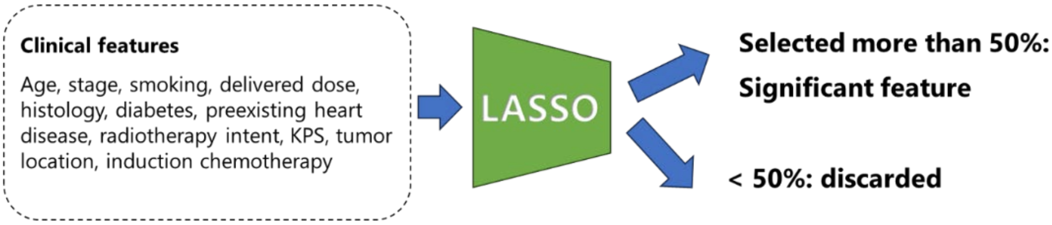

Clinical features
Age, stage, smoking, delivered dose, histology, diabetes, preexisting heart disease, radiotherapy intent, KPS, tumor location, induction chemotherapy
LASSO
Selected more than 50%:
Significant feature
< 50%: discarded

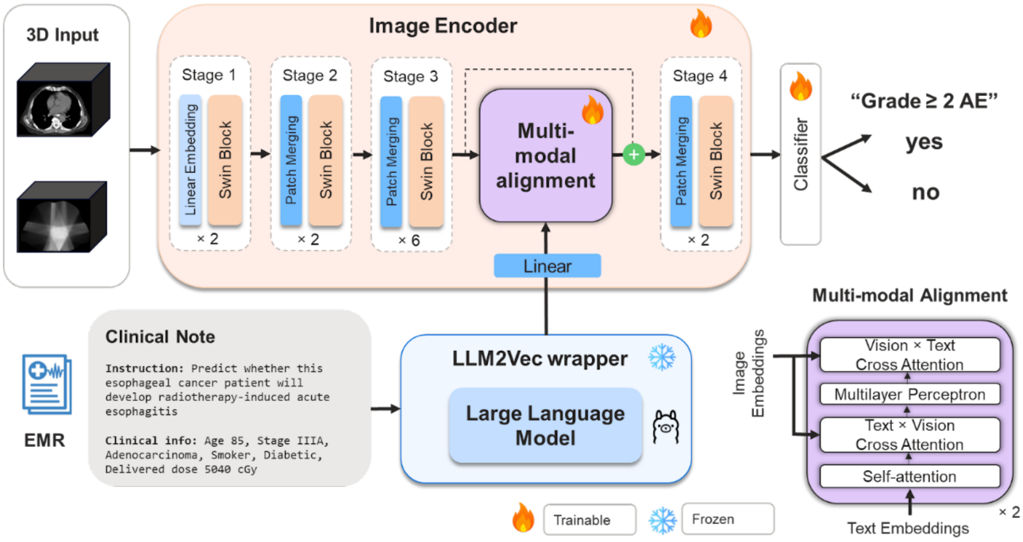
Fig 14. Overview of Feature Selection using Least Absolute Shrinkage and Selection Operator

4.2.3 Input homogeneity testing using t-SNE

To evaluate the similarity of the input data used in this study, including CT, dose maps, clinical text, we used an input data visualiztion technique called t-Distributed Stochastic Neighbor embedding (t-SNE). t-SNE is a nonlinear dimension reduction technique that can represent high dimensional data into a new location in a lower dimensional data [66]. To investigate potential data inhomogeneity between the training and testing cohorts (internal and external), which could influence model performance, we compared the distributions of CT images, dose maps, and clinical text between the training and internal test sets, as well as between the training and external test sets.

4.2.4 Prediction model framework

Our prediction framework consists of three key components: (1) an image encoder, (2) a text encoder, and (3) an attention-based multimodal alignment module (Fig 15). The image inputs are 3D CT or dose, and the text input is the patient's clinical note that consists of an instruction and a query of the patient characteristics. Starting with the image encoder, we fine-tuned a Swin transformer [67], which was pre-trained with self-supervised learning on 3,643 CT images of various cancer types (head and neck, kidney, and lung cancers) and COVID-19 [68]. The architecture utilizes the Swin-S model, a variant of the Swin base model, with the number of channels in the hidden layer of the first stage set to 96. The model consists of four stages with layer configurations of {2, 2, 18, 2} and a total of 50 million parameters. The embedding size is 768, the window size is 4x4x4, and the patch size is 2. For the text encoder, we utilized a frozen LLM2Vec encoder with the LLM model (Meta-Llama-3-8B-Instruct) [69,70]. LLM2Vec, pre-trained with masked next token prediction and unsupervised contrastive learning on English Wikipedia [20], adds bidirectional attention to decoder-only LLMs, enabling full-context aware text embedding. Lastly, vision and text features were aligned using the Segment Anything Model's two-way transformer module [71].



Instruction: Predict whether this esophageal cancer patient will develop radiotherapy-induced acute esophagitis

Clinical info: Age 85, Stage IIIA, Adenocarcinoma, Smoker, Diabetic, Delivered dose 5040 cGy

Fig 15. Overview of the 3D multimodal esophagitis prediction framework: (1) Swin Transformer image encoder, (2) LLM2Vec language encoder with frozen LLaMA 3-8B, and (3) cross-attention-based multimodal alignment.

4.2.5 Training details

The models were trained on an NVIDIA RTX A100 80GB GPU. Each approach was compared against its vision-only baseline model. Training was done using 10-fold cross-validation, and the model with the highest validation accuracy was locked and used for testing. The model was fine-tuned using the MONAI library. A 3D patch with dimensions of 128x128x128 was extracted around the GTV during the data loading process. To ensure sufficient context around the tumor, a margin of 70 voxels was added outside the tumor boundary for CT images, and a margin of 50 voxels was used for dose data in the x, y, and z directions. For the combined CT+dose model, we used a common margin of 50. Image intensity values were scaled between -1000 and 1000 Hounsfield Units for CT and between 0 and 60 Gy for dose. The model utilized cross-entropy loss with the Adam optimizer, using a learning rate of 1e-5 for image only model and 1e-4 for image+LLM model.

4.2.6 Model evaluation

To evaluate model performance, we used the area under the curve (AUC), specificity (TP/(TP+FN)), and sensitivity (TN/(TN+FP)), where classification outcomes were defined as true positives (TP), false negatives (FN), true negatives (TN), and false positives (FP). The optimal classification threshold for specificity and senstivitiy calculation was established by maximizing Youden’s index in the internal validation. The best model was selected based on the highest overall performance metrics (AUC, specificity, and sensitivity) observed across all 10 folds of the validation set. The checkpoint corresponding to the highest AUC was then used for subsequent analysis on both the internal and external test sets. Furthermore, to better understand the model’s decision-making process, we applied Gradient-weighted Class Activation Mapping (Grad-CAM) to our models which highlighted the regions of the input that contribute most to the model’s prediction.

4.2.7 Ablation experiments

Our study uses multiple inputs including CT, dose and clinical text and the way of combining these are diverse. Therefore we conducted different approahces and experiments to come up with the best way for our data and our task. To gain a deeper understanding of the model's strategies, and to assess the impact of each hyperparameter, the following subanalysis was conducted:

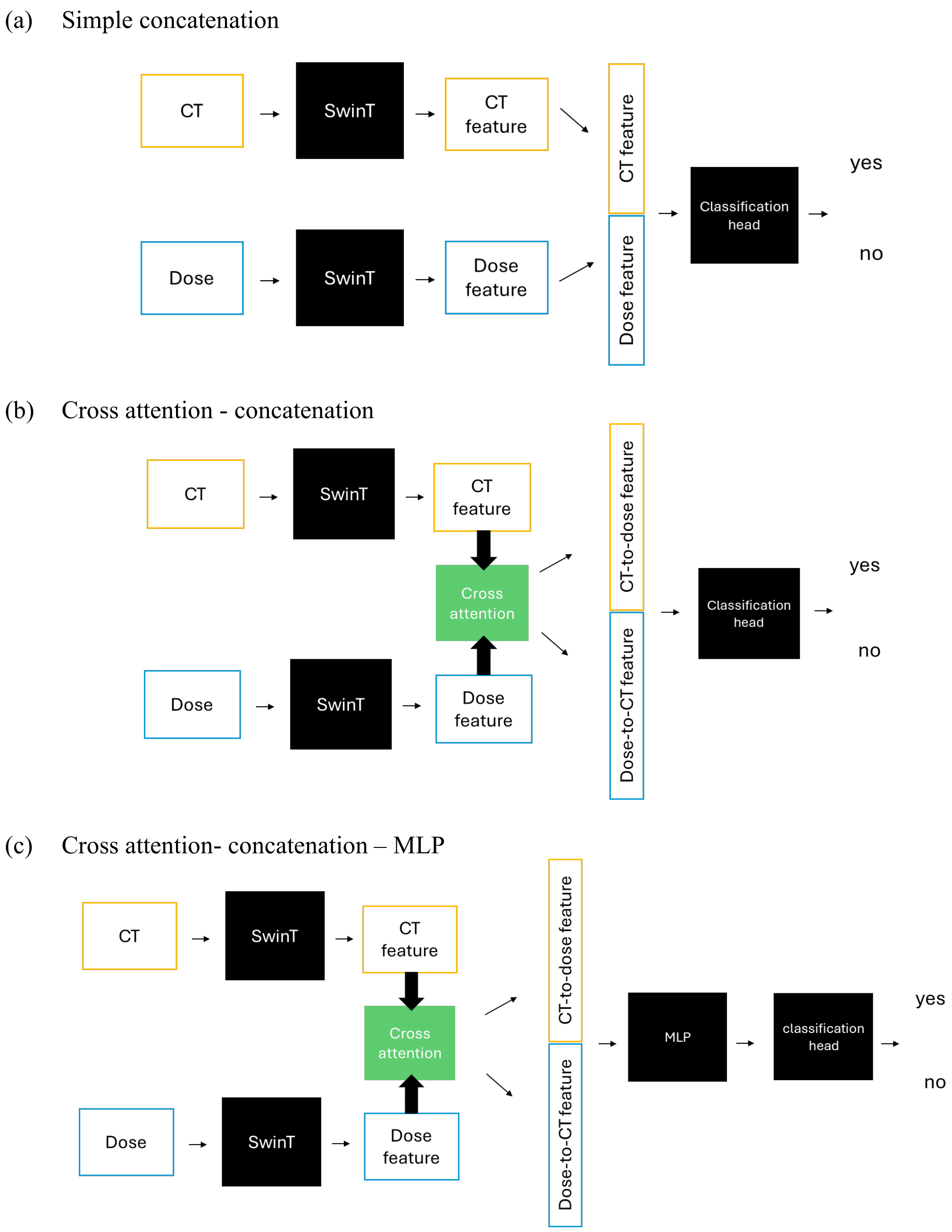
(1) Image-Text concatenation strategies: To evaluate the effectiveness of our proposed multi-modal alignment approach, we compared it with a conventional strategy where image and textual features are simply concatenated prior to the classification layer, followed by a fully connected layer for final prediction[72]. We conducted this comparison using the internal test set to determine whether the cross-attention mechanism leads to improved performance over the naive concatenation method. Both methods use the text features produced by the LLM2Vec text encoder.

(2) CT-Dose concatenation stratigies: Our dataset includes spatially aligned CT and dose distribution images from the same patient. To combine these inputs, we explored several strategies, ranging from simple to more complex. First, we used basic concatenation (Figure 16a), where CT and dose features from the transformer blocks are flattened and directly concatenated. Next, we applied bi-directional cross-attention, generating two sets of features: CT-to-dose and dose-to-CT (Figure 16b). We then tested two variants comparing simple concatenation versus adding an MLP after concatenation (Figure 16c). Finally, we evaluated whether summing the cross-attended features instead of flattening and concatenating improved performance (Figure 16d).

(3) Effect of margin size on model performance: We compared test accuracy using two different margin sizes, 20 and 50 voxels, to evaluate whether focusing the model on a smaller, more localized region around the tumor is more beneficial than providing a broader, more global field of view. This analysis was motivated in part by differences in dose distribution patterns, as tomotherapy scans typically exhibit more refined dose gradients compared to IMRT scans (Fig 16). We assessed the impact of these margin sizes on both internal and external test performance to determine which provided more informative input for the model.

(4) Prompt variation with and without histology: The primary difference in clinical information between the training and external cohorts lies in histology. Korean patients, who make up the external cohort, more commonly present with squamous cell carcinoma, whereas adenocarcinoma is more prevalent among U.S. patients in the training cohort. To assess whether this discrepancy impacts model performance, we tested the effect of removing histology information from the text prompt on the external test set.

(5) Impact of training size on model performance: A key challenge in developing deep learning models for clinical applications is limited data availability. Ideally, models should maintain robust performance even with smaller datasets. To assess whether incorporating clinical text enhances model accuracy under data constraints, we randomly subsampled the training set to 10%, 20%, and 50% of its original size and evaluated performance on the internal test set.



(a) Simple concatenation

(b) Cross attention - concatenation

(c) Cross attention- concatenation – MLP

(d) Cross attention – summation

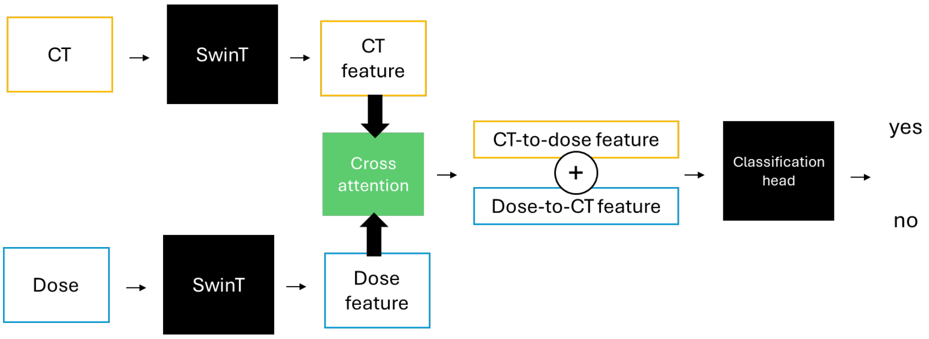


Fig 16: Comparison of multi-modal image input concatenation methods. (a) simple concatenation strategy, (b), (c), (d) combination using cross attention between CT and dose (CT-to-dose, dose-to-CT). Abbreviations: SwinT: swin transformer. MLP: multi layer perceptron

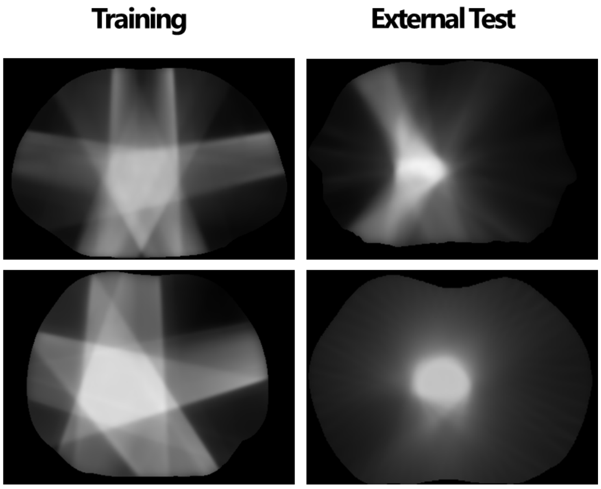


Fig 17. Visual comparison of dose maps of the training (IMRT) and external test (Tomotherapy).

4.3. Results

4.3.1 Feature selection

In the current data cohort, 13 clinical features were collected and subjected to a LASSO-based feature selection pipeline. Variables that were selected in ≥50% of the bootstrap iterations included, in descending order of selection frequency, squamous cell histology, delivered radiation dose, smoking status, clinical stage III or higher, age, and diabetes. Among these, squamous cell histology, delivered dose, smoking status, and stage III or higher were selected in 100% of the iterations, demonstrating exceptionally strong and consistent associations with the outcome of interest within this cohort. In contrast, age and diabetes were selected in a slightly lower proportion of bootstrap samples but still surpassed the inclusion threshold. The final set of selected features was retained for subsequent analyses and served as the clinical input for the multimodal analysis performed in this study. These features were incorporated into the modeling framework alongside imaging and dosimetric data to comprehensively evaluate their combined predictive value (Fig. 18).

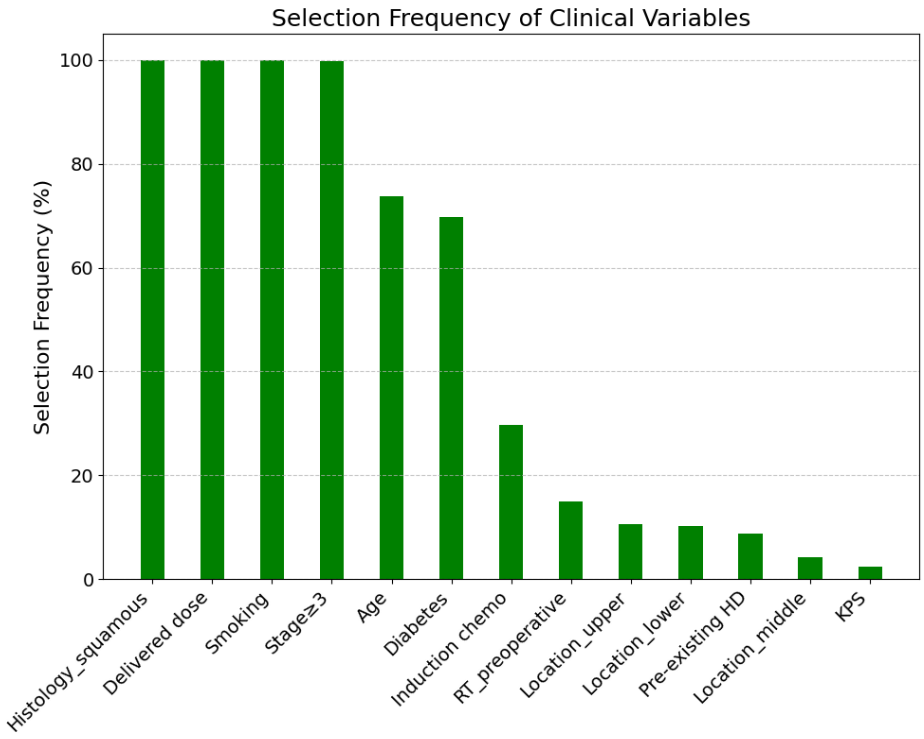


Fig 18: LASSO regression feature selection results, sorted by selection percentage

A comparative analysis of clinical characteristics between the internal (MSKCC) and external (Yonsei) cohorts revealed several key differences, as shown in Fig. 19. Notably, the distributions of smoking status, cancer stage, and histology varied between the two cohorts. The MSKCC cohort had a higher proportion of smokers, while the Yonsei cohort included more patients diagnosed with stage IV disease. The most pronounced difference was observed in histology: adenocarcinoma was more prevalent in the MSKCC cohort, whereas squamous cell carcinoma was more common in the Yonsei cohort. This difference in histological distribution was statistically significant, with a p-value less than 0.05.

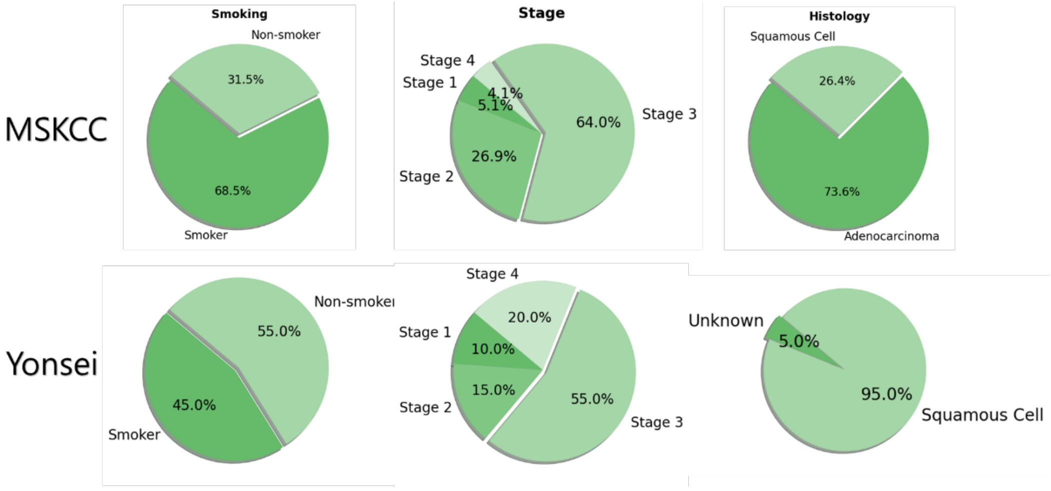


Fig 19: Difference between smoking, stage and histology of the training cohort vs external test cohort

4.3.2 t-SNE visualization

Next, we focused on evaluating the underlying data characteristics of the three modalities included in this study- CT images, dose distributions, and clinical variables using t-SNE visualization. The t-SNE plots revealed minimal differences in data distribution between the training and internal test sets across all modalities, as evidenced by the similar spread and well-mixed data points (Fig. 20). This suggests that there is no significant domain shift between the internal training and test datasets. In contrast, a greater variance was observed when comparing the training data with the external test set collected from an external institution. The t-SNE plots for all three modalities demonstrated clear differences, with the external test data forming tight, distinct clusters that showed minimal overlap with the training data. This effect was especially pronounced in the CT and clinical variable plots, where the external cohort clustered in separate regions, indicating notable differences in the underlying data distributions. These findings suggest that the external dataset has distinct characteristics compared to the internal training set, which may have implications for model generalizability.

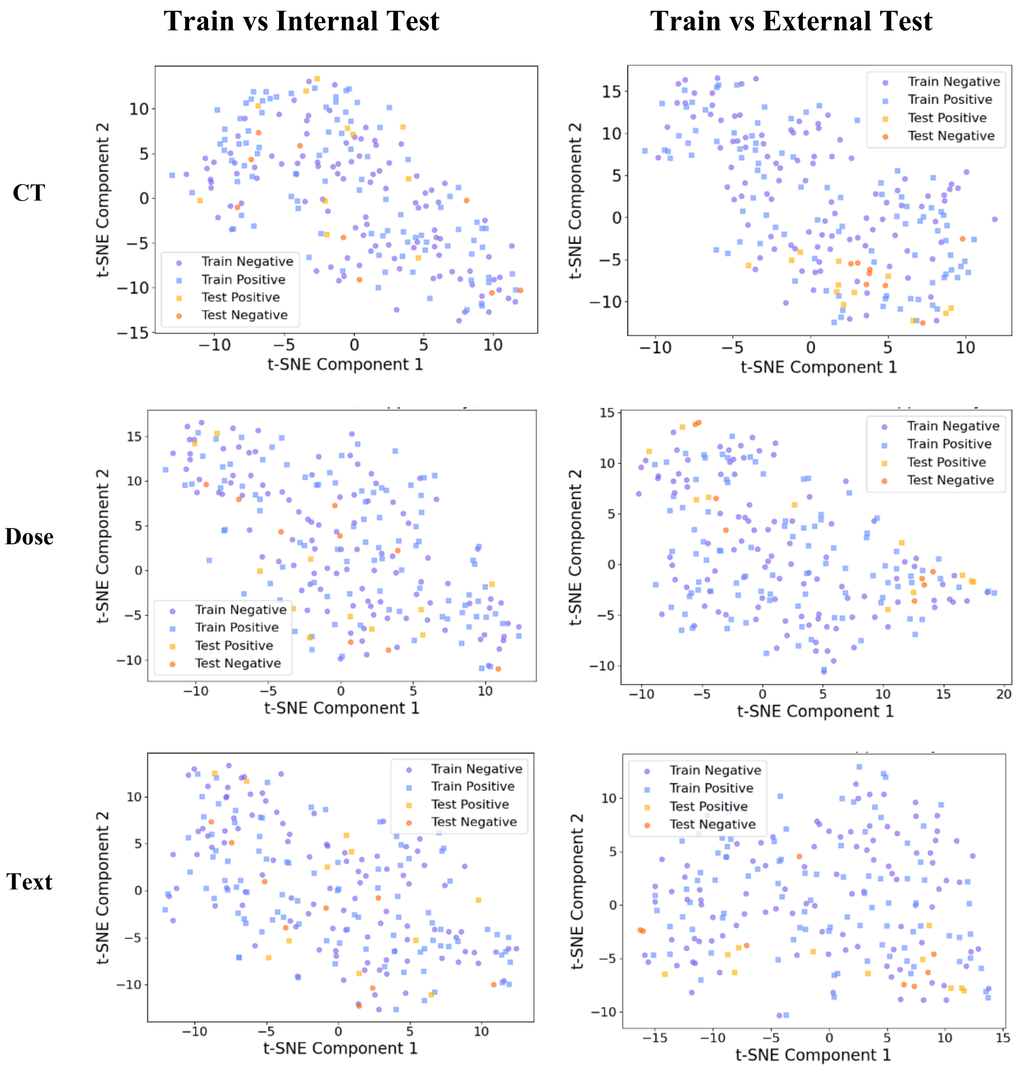


Fig 20: t-SNE plots illustrating the data distribution across the training set, internal test set, and external test set for each modality.

4.3.3 Quantitative evaluations

Table 5 summarizes the average performance metrics (AUC, specificity, and sensitivity) across 10-fold cross validation for all image only models versus their corresponding image + text models leveraging LLM and multi-modal alignment. In the internal validation, the LLM only model achieved an AUC of 0.69. Starting with the image only models, the AUC was the highest for CT+Dose model (AUC=0.77), followed by CT model (AUC=0.74) and Dose model (AUC=0.64). Overall, adding clinical information through LLM resulted in a performance increase of 0.01, 0.06 and 0.02 AUC for CT, Dose and CT+Dose models, respectively, The highest performance was observed for the model combining CT, Dose and Text, with an AUC of 0.78. Combining all three modalities had the highest SEN and SPE overall.

On the internal test set, the LLM-only model showed poor performance (AUC=0.53) (Table 6). The CT and CT+Text models performed similarly, while adding text slightly reduced the Dose model's performance (AUC dropped from 0.69 to 0.66). In contrast, the CT+Dose+Text model achieved the highest AUC of 0.80, showing an AUC improvement of 0.03 over CT+Dose. On the external test set, the LLM-only model again performed poorly (AUC=0.33). The CT-only model had the highest AUC (0.72), while CT+Text and the Dose models (with or without text) showed lower performance (AUCs of 0.32 and 0.38).

Table 5. Comparison of AUC, accuracy, specificity, and sensitivity for AE prediction across 10-fold cross validation. Arrows indicate changes relative to the image-only model (▲ for an increase, ▼ for a decrease). The highest values are highlighted in bold.

CT	Dose	Text	AUC	SEN	SPE
		✓	0.69 ± 0.10	0.60 ± 0.18	0.77 ± 0.16
✓			0.74 ± 0.10	0.69 ± 0.20	0.72 ± 0.08
✓		✓	0.76 ± 0.07	0.74 ± 0.19	0.81 ± 0.14
			▲0.02	▲0.05	▲0.09
	✓		0.64 ± 0.09	0.44 ± 0.22	0.80 ± 0.18
	✓	✓	0.70 ± 0.07	0.65 ± 0.16	0.77 ± 0.19
			▲0.06	▲0.21	▼0.03
✓	✓		0.77 ± 0.07	0.80 ± 0.14	0.75 ± 0.17
✓	✓	✓	0.78 ± 0.10	0.82 ± 0.13	0.74 ± 0.17
			▲0.01	▲0.02	▼0.01

Table 6. Comparison of AUC, accuracy, specificity, and sensitivity for AE prediction across for internal and external test sets. Arrows indicate changes relative to the image-only model (▲ for an increase, ▼ for a decrease). The highest values are highlighted in bold.

			Internal Test			External Test		
CT	**Dose**	**Text**	**AUC**	**SEN**	**SPE**	**AUC**	**SEN**	**SPE**
		✓	0.53	0.54	0.60	0.33	0.92	0.00
✓			0.74	0.64	0.67	0.72	0.66	0.63
✓		✓	0.74	0.64	0.67	0.60	0.08	0.88
			-	-	-	▼0.12	▼0.58	▲0.25
	✓		0.69	0.67	0.54	0.32	0.60	0.00
	✓	✓	0.66	0.63	0.67	0.38	0.60	0.40
			▼0.03	▼0.04	▲0.07	▲0.06	-	▲0.40
✓	✓		0.77	0.73	0.67	0.75	1.00	0.25
✓	✓	✓	0.80	0.64	0.67	0.63	0.17	0.88
			▲0.03	▼0.09	-	▼0.05	▼0.43	▲0.25

4.3.4 Ablation experiments

Image-Text concatenation strategies: We first compared interactive alignment and simple concatenation strategies to evaluate the effectiveness of our proposed multi-modal alignment approach using cross-attention (Table 7). The textual features were extracted using the LLaMA-LLM2Vec encoder. For input modalities, we evaluated two combinations: CT+Text and CT+Dose+Text. Using simple concatenation followed by a fully connected layer, the AUC for CT+Text was 0.60, while the addition of dose information resulted in a decreased AUC of 0.54. In contrast, our proposed method incorporating a LLM with multi-modal alignment via cross-attention demonstrated improved performance, achieving an AUC of 0.74 for CT+Text and 0.80 for CT+Dose+Text.

Table 7. Comparison of model performance of two image+text concatenation strategies 1) simple concatenation followed by fully connected (fc) layer, 2) the proposed multi-modal alignment using cross attention between image and text. The performance was testing using the internal test set.

	CT	Dose	Text	AUC	SEN	SPE
Simple concatenation + fc	✓		✓	0.60	0.72	0.67
	✓	✓	✓	0.54	0.54	0.44
Proposed: LLM + multi-alignment	✓		✓	0.74	0.64	0.67
	✓	✓	✓	0.80	0.64	0.67

Effect of margin size on model performance: Next, we compared models using smaller (20 voxels) vs. larger (50 voxels) in x,y and z around the GTV (Table 6) while training the model (Table 9). The larger margin, which covers a broader field of view including the whole thorax, performed better, with internal test AUCs of 0.72 vs. 0.64. The same pattern was observed in the external test set, with AUCs of 0.68 vs 0.57 for magins 50 and 20, respectively, with more balanced SEN and SPE. However, we observed a general decline in performance on the external test set compared to the internal test results.

Table 8. Comparison of using smaller margin vs bigger margin for CT-dose combined model

				Internal Test			External Test		
Margin	CT	Dose	Text	AUC	SEN	SPE	AUC	SEN	SPE
20	✓	✓		0.64	0.64	0.56	0.57	0.70	0.38
50	✓	✓		0.72	0.82	0.56	0.68	0.60	0.63

CT-Dose concatenation stratigies: Compared to the simple concatenation strategy, all cross-attention-based approaches demonstrated higher AUC values (Table 8). The best performance was observed when feature concatenation was applied immediately after the cross-attention module (Table 8b). Adding an MLP layer did not lead to any improvement. The summation-based aggregation method (Table 8d) also yielded a relatively high AUC, but its sensitivity and specificity were lower than those of methods (a–c). Among all approaches, the cross-attention method with concatenation achieved the highest specificity (0.67).

Table 9: Comparison of CT+dose concatenation strategies on internal test set

	CT	**Dose**	**Text**	**AUC**	**SEN**	**SPE**
(a) Simple concatenation	✓	✓		0.72	**0.82**	0.56
(b) Cross attention - concatenation	✓	✓		**0.77**	0.73	**0.67**
(c) Cross attention-concatenation – MLP	✓	✓		0.73	**0.82**	**0.67**
(d) Cross attention – summation	✓	✓		0.75	0.64	0.56

Prompt variation with and without histology: To account for differences between the training and external cohorts, we evaluated the effect of removing histology information from the text prompts (Table 10). On the internal test set, including histology led to a slight improvement in AUC (0.74 vs. 0.72), along with modest gains in sensitivity and specificity (0.64 and 0.67, respectively). On the external test set, the inclusion of histology similarly resulted in only a minimal AUC increase, with sensitivity and specificity remaining suboptimal regardless. Overall, the presence or absence of histology in the prompt had limited impact on performance

Table 10. Comparison of prompt with and without histology information

				Internal Test			**External Test**		
	CT	**Dose**	**Text**	**AUC**	**SEN**	**SPE**	**AUC**	**SEN**	**SPE**
With	✓		✓	0.74	0.64	0.67	0.60	0.08	0.88
Without	✓		✓	0.72	0.45	0.67	0.61	0.92	0.13

Impact of training size on model performance: Finally, we assessed how training size affects internal test AUC by comparing the performance of CT-only and CT+LLM models (Fig 22). Both models reached performance saturation with as little as 50% of the training data, suggesting that additional data beyond this point offers minimal gains. However, at reduced training sizes (20% and 10%), the CT+LLM model consistently outperformed the CT-only model, highlighting the added value of integrating clinical features via LLM in low-data settings. The AUC at 10% training size was slightly higher than at 20%, which we attribute to random variability, as subsets were selected randomly.

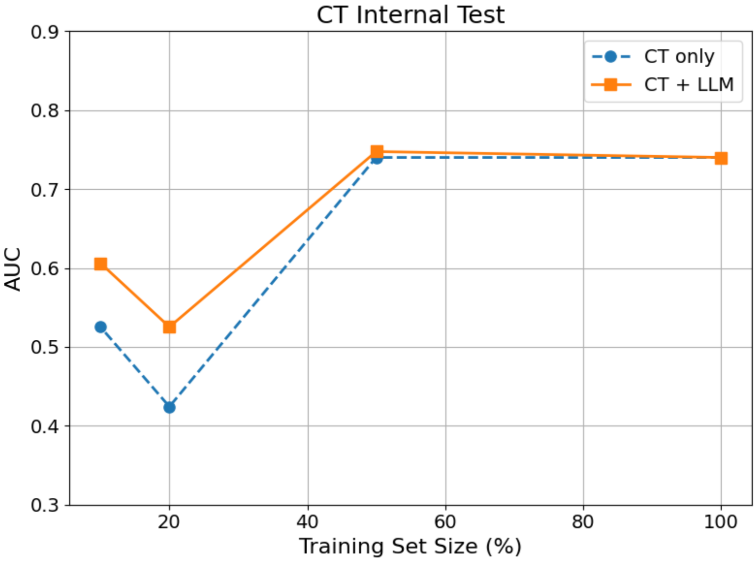


Fig 21. Effect of training size on internal test performance with data points at 50%, 20%, and 10% of training data and its associated AUC.

4.3.5 Grad-CAM visualization

Figure 21 shows Grad-CAM visualizations from three representative patients, overlaid on their CT images and dose distributions. In these heatmaps, red indicates areas of high model focus, while blue represents low focus. The visualizations reveal distinct attention patterns for the two input modalities: the CT-based model predominantly attends to the GTV, outlined in yellow, suggesting that morphological features within the tumor region are important for prediction. In contrast, the dose-based model places greater emphasis on high-dose regions or hotspots, indicating a possible link between predicted toxicity and localized radiation exposure. Despite these general trends, attention patterns varied across patients, reflecting the complex and individualized nature of model interpretation. These differences highlight the complementary value of CT and dose inputs and underscore the need for interpretability tools to better understand model behavior in clinical contexts.

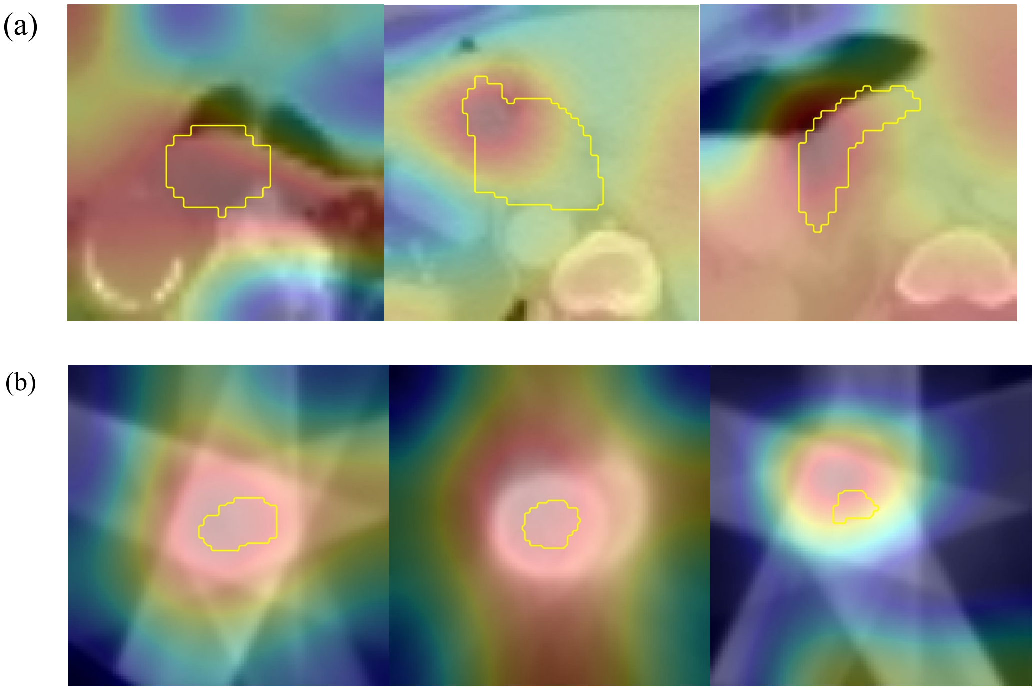


Fig 22: Grad-CAM visualizations of the image-based models for three representative patients: (a) CT-based model and (b) dose-based model. Heatmaps illustrate regions of model attention, with red indicating high focus and blue indicating low focus. The yellow contour outlines the gross tumor volume (GTV).

4.4. Discussion and Conclusion

In this study, we developed a multimodal prediction framework to estimate the risk of RT-induced acute esophagitis by integrating pre-RT imaging and clinical text data. To maximize representational capacity of the models, we used a pretrained transformer-based model for processing imaging inputs and an LLM-based text embedding model to represent relevant information from clinical notes. To the best of our knowledge, this is the first application of an LLM for aiding toxicity prediction for patients with esophageal cancer. Among the evaluated models, the CT + Dose + LLM model achieved the best performance, yielding AUCs of 0.78 on internal validation and 0.80 on internal testing, both outperforming image-only baselines and highlighting the added value of multi-modal input and clinical context in predictive modeling.

Previous work on acute esophagitis prediction has predominantly focused on patients with lung cancer. The methods can be divided into three, which are NTCP models, machine learning, and radiomics. Starting with the NTCP models, these models use simple logistic regression with clinical and dosimetric variables, Huang et al. achieved an AUC of 0.78 using the mean esophageal dose and concurrent chemotherapy as variables[56]. Similarly, Chen et al. reported a cross validation AUC of 0.79 using a multivariable logistic regression[58]. These models, while interpretable and clinically intuitive, are constrained by their reliance on a small number of hand-selected features. Machine learning approaches, in contrast, utilises a broader set of input variables. Luna et al. (2017) reported an AUC of 0.62 using a machine learning framework[60]; however, a subsequent study by the same group in 2020, which included a larger cohort, failed to identify consistent predictors of esophagitis and yielded lower AUCs ranging from 0.46 to 0.56[61]. These findings highlight the limitation of traditional clinical and dosimetric features on building robust predictive models. Lastly, radiomics-based models have recently gained interest, leveraging high-dimensional imaging features to enhance predictive performance compared to the previous methods[63,64,72]. Existing studies employed either radiomics alone or in combination with other modalities. The performance of our proposed model is comparable to these radiomics benchmarks, which report AUCs in the range of 0.74 to 0.82[64,73].

Our study provided a comprehensive comparison of predictive models developed using different input modalities, highlighting their relative strengths. Among the single modality models (CT only, dose only, and text only), the CT-based model demonstrated the highest performance (test internal AUC=0.74), followed by the dose-based model (internal test AUC=0.69), with the text-based model performing the lowest (internal test AUC=0.53). These findings suggest that not all modalities contribute equally to prediction accuracy, and imaging features from CT appear to be the most informative in this context. Furthermore, when applied to external test dataset with a completely different demographic (i.e. Korean cohort), our CT-based model had an AUC of 0.72, meaning that the CT model is more generalizable, whereas other modalities failed to generalize. Our CT-based model outperformed those reported in previous studies. For example, one study reported a test AUC of 0.691 for their best CT-based deep learning model[64]. One possible reason for the improved performance in our work may be the choice of model architecture and the use of a pretrained image encoder. Specifically, we employed a Swin Transformer model pretrained on a large-scale medical imaging dataset, which likely enabled the model to extract richer and more relevant features compared to non-pretrained CNN-based models. This approach proved effective even with our relatively small training cohort. However, their dose-based model achieved a higher internal test AUC of 0.76, compared to 0.69 in our dose-only model. This difference may be attributed to the variation in input design; Xie et al. used a mask that covered the entire esophagus, whereas our model focused only on the region surrounding the GTV.

Through extensive ablation studies, we demonstrated the critical role of concatenation and cross-attention strategies in multimodal modeling. Our framework incorporated a diverse range of modality combinations, including image-to-image pairings such as CT and dose distributions, as well as image-to-text integrations combining imaging with clinical information. We found that incorporating cross-attention mechanisms improved model performance compared to simple concatenation. The cross-attention allowed the model to effectively align and integrate complementary features within and across modalities, dynamically weighing their contributions. This was particularly beneficial not only for image-to-text fusion but also for image-to-image combinations, where capturing complex inter-modal relationships is essential. Without attention-based alignment, the model's ability to fully utilize the rich information from each modality was limited, restricting predictive accuracy and the overall effectiveness of multimodal fusion.

In addition to the performance improvement observed with the inclusion of the LLM compared to the image-only model, another advantage of using the LLM is its reduced dependency on dataset size. Our study also explored the impact of training dataset size on deep learning model performance in medical imaging. We found that performance gains plateaued at approximately 50% of the training data, suggesting that adding more data beyond this point yields minimal benefits. More notably, the LLM enhanced performance even with smaller datasets (10-20%) than the image-only model, indicating that the LLM provides valuable contextual information that compensates for the limited imaging data. Given the challenges associated with healthcare data collection, optimizing model performance with limited data is highly advantageous, which may streamline the model development and enable broader applications.

Understanding data distribution is as important as developing sophisticated models, especially ones involving multi-modal input. A rich and diverse representation is essential to ensure model robustness and generalization, especially when working with multimodal inputs. While integrating text-based clinical information improved performance on internal datasets, the model's effectiveness declined on external datasets. This drop is likely due to substantial differences in dose distributions and underlying clinical characteristics between the training and external test data, emphasizing the need for more robust data representations to better accommodate such variations.

This study has several limitations. First, the dataset used was relatively small, with limited variability in clinical language prompts. While our findings provide early evidence supporting the feasibility of using LLMs for classification tasks in radiation oncology, more extensive text ablation experiments are necessary to better understand the contribution of clinical context. Additionally, the current model does not address the issue of domain shift, as LLMs are not inherently robust to variations across cohorts. Thus, future studies should include larger and more diverse datasets to improve model generalization. To this end, we plan to expand our training data with both internal and external sources, enhancing representation and reducing bias. In parallel, we will explore subset training strategies to identify the minimal data requirements needed for stable performance, with the goal of reducing computational and annotation burdens. Another limitation lies in the lack of interpretability: we did not assess the individual contributions of each clinical variable provided to the LLM. Understanding the relative importance of these variables could improve transparency and user confidence in AI-driven predictions. Future work will also focus on optimizing the integration of LLMs into radiation therapy workflows, including the development of automated tools for data summarization, prompt engineering, and streamlined clinical data curation.

In conclusion, we developed a novel multimodal framework for predicting acute esophagitis that, to our knowledge, is the first to integrate large language models with vision-based architectures for this purpose. Our study highlights not only the promising potential of combining these modalities but also identifies critical areas for further refinement, including data diversity, model architecture, and hyperparameter optimization. Although our initial results are encouraging, more work remains to enhance the model's generalizability and predictive accuracy across varied and heterogeneous patient cohorts. Going forward, our goal is to evolve this framework into a robust, interpretable, and clinically applicable tool capable of providing reliable pre-treatment toxicity risk assessments for patients undergoing thoracic radiotherapy, thereby supporting personalized treatment planning and improving patient outcomes.

5. Conclusion and Future Work

The current study aimed to determine innovative imaging and AI technologies to enhance RT workflow across multiple stages and improve patient-specific prediction accuracy, aiming to develop a reliable AI-based framework for predicting RT-induced toxicity. Our key aims included the following:

1. To develop a deep learning-based auto-segmentation model and assess its clinical feasibility for radiotherapy.
2. To clinically evaluate a novel breath-hold technique and compare its effectiveness with conventional methods.
3. To develop a multimodal model for predicting adverse events by integrating imaging features and clinical data using deep learning and LLMs.

In Chapter 2, we proposed a deep learning–based automated segmentation framework to streamline treatment planning and reduce interobserver variability in multi-center RTQA. Compared to manual contours and conventional atlas-based segmentation (ABAS), the DLBAS method demonstrated greater consistency and robustness across most CTVs and normal organs. Our study was among the first to highlight DLBAS's potential in supporting a key step in the RT workflow, offering a more accurate and time-saving tool for structure definition that benefits downstream planning.

Beyond routine clinical use, we validated the tool's feasibility in a multi-center trial setting. User surveys confirmed both quantitative and qualitative improvements in segmentation consistency, time efficiency, and interobserver agreement. These enhancements support not only treatment quality and efficiency but also RT education and quality assurance. The improved consistency in defining target volumes and organs-at-risk suggests that deep learning models like DLBAS may help reduce protocol deviations in clinical trials.
In chapter 3, we implemented a novel techniques to improve simulation imaging and delivery accuracy by addressing motion-related artifacts and patient variability

In Chapter 3, we investigated the clinical feasibility of using CPAP to reduce motion artifacts caused by breathing, aiming to improve patient compliance and reduce toxicity compared to the conventional breath-hold technique (DIBH). Our study showed that CPAP provides sufficient heart-sparing, with sub-centimeter variation across treatment fractions and reproducible positioning in VMAT for left-sided breast cancer, along with a high compliance rate. These findings support CPAP as a practical and effective option for routine use in left-sided breast cancer radiation therapy. Clinically, this has two key implications: first, more reliable breath management during pre-treatment imaging enables the acquisition of a more consistent planning CT; second, using the same breath-hold technique during treatment helps ensure that the planned dose is delivered more accurately.

Finally, in Chapter 4, we developed multi-modal predictive models that integrate imaging and clinical data to support personalized, risk-adaptive decision-making. By extensively testing input combinations and employing innovative modeling strategies, we proposed a multi-modal framework that leverages the full range of pre-treatment information. Notably, incorporating clinical data through an LLM enhanced the performance of image-based models by providing rich contextual representation. This approach not only improves patient monitoring but also helps address data-dependency challenges by introducing an additional, complementary modality.

Future work will focus on expanding the toxicity prediction framework beyond esophagitis to other thoracic toxicities, such as radiation pneumonitis, cardiac toxicity, and fibrosis. Broadening the application across different cancer types will enhance clinical relevance. Further, improving the robustness and interpretability of large language model–based predictors through domain-specific fine-tuning will be essential for clinical integration. Exploring adaptive treatment strategies based on early toxicity risk, such as dynamic replanning, also presents a promising direction.

In conclusion, this thesis lays the foundation for the development of AI-driven tools that can improve the safety, precision, and personalization of radiation therapy. By addressing critical points in both the planning and delivery stages, our methods contribute to a more efficient RT workflow and hold significant potential to improve patient outcomes by enhancing the management of RT-induced toxicity. This work not only advances the technological integration of AI in clinical oncology but also sets the stage for more personalized, adaptive treatment strategies that could benefit a wide range of patients undergoing radiation therapy.

References

1. Cancer Statistics - NCI. April 2, 2015. Accessed May 19, 2025. https://www.cancer.gov/about-cancer/understanding/statistics

2. Baskar R, Lee KA, Yeo R, Yeoh KW. Cancer and Radiation Therapy: Current Advances and Future Directions. *Int J Med Sci*. 2012;9(3):193-199. doi:10.7150/ijms.3635

3. Citrin DE. Recent Developments in Radiotherapy. *New England Journal of Medicine*. 2017;377(11):1065-1075. doi:10.1056/NEJMra1608986

4. Wang K, Tepper JE. Radiation therapy-associated toxicity: Etiology, management, and prevention. *CA: A Cancer Journal for Clinicians*. 2021;71(5):437-454. doi:10.3322/caac.21689

5. Lee Chuy K, Nahhas O, Dominic P, et al. Cardiovascular Complications Associated with Mediastinal Radiation. *Curr Treat Options Cardio Med*. 2019;21(7):31. doi:10.1007/s11936-019-0737-0

6. Latrèche A, Bourbonne V, Lucia F. Unrecognized thoracic radiotherapy toxicity: A review of literature. *Cancer/Radiothérapie*. 2022;26(4):616-621. doi:10.1016/j.canrad.2021.10.008

7. Yorke ED. Modeling the effects of inhomogeneous dose distributions in normal tissues. *Seminars in Radiation Oncology*. 2001;11(3):197-209. doi:10.1053/srao.2001.23478

8. Gagliardi G, Lax I, Ottolenghi A, Rutqvist LE. Long-term cardiac mortality after radiotherapy of breast cancer--application of the relative seriality model. *Br J Radiol*. 1996;69(825):839-846. doi:10.1259/0007-1285-69-825-839

9. Huang EX, Bradley JD, El Naqa I, et al. Modeling the Risk of Radiation-Induced Acute Esophagitis for Combined Washington University and RTOG Trial 93-11 Lung Cancer Patients. *International Journal of Radiation Oncology*Biology*Physics*. 2012;82(5):1674-1679. doi:10.1016/j.ijrobp.2011.02.052

10. Niezink AGH, van der Schaaf A, Wijsman R, et al. External validation of NTCP-models for radiation pneumonitis in lung cancer patients treated with chemoradiotherapy. *Radiother Oncol*. 2023;186:109735. doi:10.1016/j.radonc.2023.109735

11. Langendijk JA, Lambin P, De Ruysscher D, Widder J, Bos M, Verheij M. Selection of patients for radiotherapy with protons aiming at reduction of side effects: The model-based approach. *Radiotherapy and Oncology*. 2013;107(3):267-273. doi:10.1016/j.radonc.2013.05.007

12. Van den Bosch L, Schuit E, van der Laan HP, et al. Key challenges in normal tissue complication probability model development and validation: towards a comprehensive strategy. *Radiotherapy and Oncology*. 2020;148:151-156. doi:10.1016/j.radonc.2020.04.012

13. Lyman JT. Complication Probability as Assessed from Dose-Volume Histograms. *Radiation Research*. 1985;104(2s):S13-S19. doi:10.2307/3576626

14. Källman P, Ågren ,A., and Brahme A. Tumour and Normal Tissue Responses to Fractionated Non-uniform Dose Delivery. *International Journal of Radiation Biology*. 1992;62(2):249-262. doi:10.1080/09553009214552071

15. Rueckert D, Aljabar P, Heckemann RA, Hajnal JV, Hammers A. Diffeomorphic registration using B-splines. *Med Image Comput Comput Assist Interv*. 2006;9(Pt 2):702-709. doi:10.1007/11866763_86

16. Wang H, Dong L, O'Daniel J, et al. Validation of an accelerated 'demons' algorithm for deformable image registration in radiation therapy. *Phys Med Biol*. 2005;50(12):2887. doi:10.1088/0031-9155/50/12/011

17. Marstal K, Berendsen F, Staring M, Klein S. SimpleElastix: A User-Friendly, Multi-Lingual Library for Medical Image Registration. In: ; 2016:134-142. Accessed September 10, 2024. https://www.cv-foundation.org/openaccess/content_cvpr_2016_workshops/w15/html/Marstal_SimpleElastix_A_User-Friendly_CVPR_2016_paper.html

18. Avants BB, Epstein CL, Grossman M, Gee JC. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Med Image Anal*. 2008;12(1):26-41. doi:10.1016/j.media.2007.06.004

19. Mylona E, Acosta O, Lizee T, et al. Voxel-Based Analysis for Identification of Urethrovesical Subregions Predicting Urinary Toxicity After Prostate Cancer Radiation Therapy. *International Journal of Radiation Oncology*Biology*Physics*. 2019;104(2):343-354. doi:10.1016/j.ijrobp.2019.01.088

20. Monti S, Xu T, Liao Z, Mohan R, Cella L, Palma G. On the interplay between dosiomics and genomics in radiation-induced lymphopenia of lung cancer patients. *Radiotherapy and Oncology*. 2022;167:219-225. doi:10.1016/j.radonc.2021.12.038

21. Cho Y, Kim Y, Chamseddine I, et al. Lymphocyte dynamics during and after chemo-radiation correlate to dose and outcome in stage III NSCLC patients undergoing maintenance immunotherapy. *Radiotherapy and Oncology*. 2022;168:1-7. doi:10.1016/j.radonc.2022.01.007

22. Palma G, Monti S, D'Avino V, et al. A Voxel-Based Approach to Explore Local Dose Differences Associated With Radiation-Induced Lung Damage. *International Journal of Radiation Oncology*Biology*Physics*. 2016;96(1):127-133. doi:10.1016/j.ijrobp.2016.04.033

23. Ibragimov B, Toesca D, Chang D, Yuan Y, Koong A, Xing L. Development of deep neural network for individualized hepatobiliary toxicity prediction after liver SBRT. *Med Phys*. 2018;45(10):4763-4774. doi:10.1002/mp.13122

24. Men K, Geng H, Zhong H, Fan Y, Lin A, Xiao Y. A Deep Learning Model for Predicting Xerostomia Due to Radiation Therapy for Head and Neck Squamous Cell Carcinoma in the RTOG 0522 Clinical Trial. *International Journal of Radiation Oncology*Biology*Physics*. 2019;105(2):440-447. doi:10.1016/j.ijrobp.2019.06.009

25. Oh Y, Park S, Byun HK, et al. LLM-driven multimodal target volume contouring in radiation oncology. *Nat Commun*. 2024;15(1):9186. doi:10.1038/s41467-024-53387-y

26. Dosik K Kyungwon and Lee, Yongmoon and Park, Doohyun and Eo, Taejoon and Youn, Daemyung and Lee, Hyesang and Hwang. LLM-guided Multi-modal Multiple Instance Learning for 5-year Overall Survival Prediction of Lung Cancer. MICCAI 2024 - Open Access. September 1, 2024. Accessed March 6, 2025. https://papers.miccai.org/miccai-2024/468-Paper2173

27. Vinod SK, Jameson MG, Min M, Holloway LC. Uncertainties in volume delineation in radiation oncology: A systematic review and recommendations for future studies. *Radiotherapy and Oncology*. 2016;121(2):169-179. doi:10.1016/j.radonc.2016.09.009

28. Caravatta L, Macchia G, Mattiucci GC, et al. Inter-observer variability of clinical target volume delineation in radiotherapy treatment of pancreatic cancer: a multi-institutional contouring experience. *Radiat Oncol*. 2014;9(1):198. doi:10.1186/1748-717X-9-198

29. Hong TS, Tome WA, Chappell RJ, Harari PM. Variations in target delineation for head and neck IMRT: An international multi-institutional study. *International Journal of Radiation Oncology*Biology*Physics*. 2004;60(1, Supplement):S157-S158. doi:10.1016/j.ijrobp.2004.06.073

30. Jansen EPM, Nijkamp J, Gubanski M, Lind PARM, Verheij M. Interobserver Variation of Clinical Target Volume Delineation in Gastric Cancer. *International Journal of Radiation Oncology*Biology*Physics*. 2010;77(4):1166-1170. doi:10.1016/j.ijrobp.2009.06.023

31. Poortmans PMP, Takanen S, Marta GN, Meattini I, Kaidar-Person O. Winter is over: The use of Artificial Intelligence to individualise radiation therapy for breast cancer. *Breast*. 2019;49:194-200. doi:10.1016/j.breast.2019.11.011

32. Fukumitsu N, Nitta K, Terunuma T, et al. Registration error of the liver CT using deformable image registration of MIM Maestro and Velocity AI. *BMC Medical Imaging*. 2017;17(1):30. doi:10.1186/s12880-017-0202-z

33. Rice TK, Schork NJ, Rao DC. Methods for Handling Multiple Testing. In: *Advances in Genetics*. Vol 60. Genetic Dissection of Complex Traits. Academic Press; 2008:293-308. doi:10.1016/S0065-2660(07)00412-9

34. Offersen BV, Boersma LJ, Kirkove C, et al. ESTRO consensus guideline on target volume delineation for elective radiation therapy of early stage breast cancer. *Radiotherapy and Oncology*. 2015;114(1):3-10. doi:10.1016/j.radonc.2014.11.030

35. Kaidar-Person O, Vrou Offersen B, Hol S, et al. ESTRO ACROP consensus guideline for target volume delineation in the setting of postmastectomy radiation therapy after implant-based immediate reconstruction for early stage breast cancer. *Radiother Oncol*. 2019;137:159-166. doi:10.1016/j.radonc.2019.04.010

36. Choi MS, Choi BS, Chung SY, et al. Clinical evaluation of atlas- and deep learning-based automatic segmentation of multiple organs and clinical target volumes for breast cancer. *Radiotherapy and Oncology*. 2020;153:139-145. doi:10.1016/j.radonc.2020.09.045

37. Lee HY, Chang JS, Lee IJ, et al. The deep inspiration breath hold technique using Abches reduces cardiac dose in patients undergoing left-sided breast irradiation. *Radiat Oncol J*. 2013;31(4):239-246. doi:10.3857/roj.2013.31.4.239

38. Stick LB, Vogelius IR, Risum S, Josipovic M. Intrafractional fiducial marker position variations in stereotactic liver radiotherapy during voluntary deep inspiration breath-hold. *Br J Radiol*. 2020;93(1116):20200859. doi:10.1259/bjr.20200859

39. Goldstein JD, Lawrence YR, Appel S, et al. Continuous Positive Airway Pressure for Motion Management in Stereotactic Body Radiation Therapy to the Lung: A Controlled Pilot Study. *International Journal of Radiation Oncology*Biology*Physics*. 2015;93(2):391-399. doi:10.1016/j.ijrobp.2015.06.011

40. Allen AM, Ceder YK, Shochat T, et al. CPAP (Continuous Positive Airway Pressure) is an effective and stable solution for heart sparing radiotherapy of left sided breast cancer. *Radiation Oncology*. 2020;15(1):59. doi:10.1186/s13014-020-01505-7

41. Kil WJ, Pham T, Kim K. Heart sparing breast cancer radiotherapy using continuous positive airway pressure (CPAP) and conventional supine tangential fields: an alternative method for patients with limited accessibility to advanced radiotherapy techniques. *Acta Oncologica*. 2019;58(1):105-109. doi:10.1080/0284186X.2018.1503711

42. Josipovic M, Persson GF, Dueck J, et al. Geometric uncertainties in voluntary deep inspiration breath hold radiotherapy for locally advanced lung cancer. *Radiother Oncol*. 2016;118(3):510-514. doi:10.1016/j.radonc.2015.11.004

43. Cheung PCF, Sixel KE, Tirona R, Ung YC. Reproducibility of lung tumor position and reduction of lung mass within the planning target volume using active breathing control (ABC). *Int J Radiat Oncol Biol Phys*. 2003;57(5):1437-1442. doi:10.1016/j.ijrobp.2003.08.006

44. Kimura T, Murakami Y, Kenjo M, et al. Interbreath-hold reproducibility of lung tumour position and reduction of the internal target volume using a voluntary breath-hold method with spirometer during stereotactic radiotherapy for lung tumours. *Br J Radiol*. 2007;80(953):355-361. doi:10.1259/bjr/31008031

45. Comsa D, Zhang B, Mosely D, Yeung I. Poster - Thur Eve - 26: Interfraction reproducibility of heart position during breast irradiation using Active Breathing Control. *Med Phys*.

2012;39(7Part3):4629. doi:10.1118/1.4740134

46. Koivumäki T, Tujunen J, Virén T, Heikkilä J, Seppälä J. Geometrical uncertainty of heart position in deep-inspiration breath-hold radiotherapy of left-sided breast cancer patients. *Acta Oncol*. 2017;56(6):879-883. doi:10.1080/0284186X.2017.1298836

47. Kil WJ, Pham T, Hossain S, Casaigne J, Jones K, Khalil M. The impact of continuous positive airway pressure on radiation dose to heart and lung during left-sided postmastectomy radiotherapy when deep inspiration breath hold technique is not applicable: a case report. *Radiat Oncol J*. 2018;36(1):79-84. doi:10.3857/roj.2018.00017

48. Appel S, Weizman N, Davidson T, et al. Reexpansion of atelectasis caused by use of continuous positive airway pressure (CPAP) before radiation therapy (RT). *Advances in Radiation Oncology*. 2016;1(2):136-140. doi:10.1016/j.adro.2016.03.002

49. Di Perri D, Colot A, Delor A, et al. Effect of continuous positive airway pressure administration during lung stereotactic ablative radiotherapy: a comparative planning study. *Strahlenther Onkol*. 2018;194(6):591-599. doi:10.1007/s00066-018-1278-2

50. Ko H, Chang JS, Moon JY, et al. Dosimetric Comparison of Radiation Techniques for Comprehensive Regional Nodal Radiation Therapy for Left-Sided Breast Cancer: A Treatment Planning Study. *Front Oncol*. 2021;11:645328. doi:10.3389/fonc.2021.645328

51. Mouawad M, Lailey O, Poulsen P, et al. Intrafraction motion monitoring to determine PTV margins in early stage breast cancer patients receiving neoadjuvant partial breast SABR. *Radiother Oncol*. 2021;158:276-284. doi:10.1016/j.radonc.2021.02.021

52. Rice L, Goldsmith C, Green MM, Cleator S, Price PM. An effective deep-inspiration breath-hold radiotherapy technique for left-breast cancer: impact of post-mastectomy treatment, nodal coverage, and dose schedule on organs at risk. *Breast Cancer (Dove Med Press)*. 2017;9:437-446. doi:10.2147/BCTT.S130090

53. Rodriguez GM, DePuy D, Aljehani M, et al. Trends in Epidemiology of Esophageal Cancer in the US, 1975-2018. *JAMA Network Open*. 2023;6(8):e2329497. doi:10.1001/jamanetworkopen.2023.29497

54. van Hagen P, Hulshof MCCM, van Lanschot JJB, et al. Preoperative chemoradiotherapy for esophageal or junctional cancer. *N Engl J Med*. 2012;366(22):2074-2084. doi:10.1056/NEJMoa1112088

55. Aupérin A, Le Péchoux C, Rolland E, et al. Meta-analysis of concomitant versus sequential radiochemotherapy in locally advanced non-small-cell lung cancer. *J Clin Oncol*. 2010;28(13):2181-2190. doi:10.1200/JCO.2009.26.2543

56. Huang EX, Robinson CG, Molotievschi A, Bradley JD, Deasy JO, Oh JH. Independent test of a model to predict severe acute esophagitis. *Advances in Radiation Oncology*. 2017;2(1):37-43.

doi:10.1016/j.adro.2016.11.003

57. Chapet O, Kong FM, Lee JS, Hayman JA, Ten Haken RK. Normal tissue complication probability modeling for acute esophagitis in patients treated with conformal radiation therapy for non-small cell lung cancer. *Radiotherapy and Oncology*. 2005;77(2):176-181. doi:10.1016/j.radonc.2005.10.001

58. Chen M, Wang Z, Jiang S, et al. Predictive performance of different NTCP techniques for radiation-induced esophagitis in NSCLC patients receiving proton radiotherapy. *Sci Rep*. 2022;12(1):9178. doi:10.1038/s41598-022-12898-8

59. Wang D, Lee SH, Yegya-Raman N, et al. Interpretable Machine Learning Models for Severe Esophagitis Prediction in LA-NSCLC Patients Treated with Chemoradiation Therapy. *International Journal of Radiation Oncology, Biology, Physics*. 2023;117(2):e490. doi:10.1016/j.ijrobp.2023.06.1720

60. Luna JM, Valdes G, Berman AT, et al. Novel Use of Machine Learning for Predicting Radiation Esophagitis in Locally Advanced Stage II-III Non–small Cell Lung Cancer. *International Journal of Radiation Oncology, Biology, Physics*. 2017;99(2):E476-E477. doi:10.1016/j.ijrobp.2017.06.1743

61. Luna JM, Chao HH, Shinohara RT, et al. Machine learning highlights the deficiency of conventional dosimetric constraints for prevention of high-grade radiation esophagitis in non-small cell lung cancer treated with chemoradiation. *Clinical and Translational Radiation Oncology*. 2020;22:69-75. doi:10.1016/j.ctro.2020.03.007

62. Zheng X, Guo W, Wang Y, et al. Multi-omics to predict acute radiation esophagitis in patients with lung cancer treated with intensity-modulated radiation therapy. *Eur J Med Res*. 2023;28(1):126. doi:10.1186/s40001-023-01041-6

63. Ma Z, Liang B, Wei R, et al. Enhanced prediction of postoperative radiotherapy-induced esophagitis in non-small cell lung cancer: Dosiomic model development in a real-world cohort and validation in the PORT-C randomized controlled trial. *Thoracic Cancer*. 2023;14(28):2839-2845. doi:10.1111/1759-7714.15068

64. Xie C, Yu X, Tan N, et al. Combined deep learning and radiomics in pretreatment radiation esophagitis prediction for patients with esophageal cancer underwent volumetric modulated arc therapy. *Radiotherapy and Oncology*. 2024;199:110438. doi:10.1016/j.radonc.2024.110438

65. Trotti A, Colevas AD, Setser A, et al. CTCAE v3.0: Development of a comprehensive grading system for the adverse effects of cancer treatment. *Seminars in radiation oncology*. 2003;13(3):176-181. doi:10.1016/S1053-4296(03)00031-6

66. Maaten L van der, Hinton G. Visualizing Data using t-SNE. *Journal of Machine Learning Research*. 2008;9(86):2579-2605.

67. Liu Z, Lin Y, Cao Y, et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. Published online August 17, 2021. doi:10.48550/arXiv.2103.14030

68. Jiang J, Veeraraghavan H. Self-distilled Masked Attention guided masked image modeling with noise Regularized Teacher (SMART) for medical image analysis. *arXiv e-prints*. Published online October 1, 2023. doi:10.48550/arXiv.2310.01209

69. Grattafiori A, Dubey A, Jauhri A, et al. The Llama 3 Herd of Models. Published online November 23, 2024. doi:10.48550/arXiv.2407.21783

70. BehnamGhader P, Adlakha V, Mosbach M, Bahdanau D, Chapados N, Reddy S. LLM2Vec: Large Language Models Are Secretly Powerful Text Encoders. Published online August 21, 2024. doi:10.48550/arXiv.2404.05961

71. Kirillov A, Mintun E, Ravi N, et al. Segment Anything. Published online April 5, 2023. doi:10.48550/arXiv.2304.02643

72. Zheng S, Guo J, Langendijk JA, et al. Survival prediction for stage I-IIIA non-small cell lung cancer using deep learning. *Radiotherapy and Oncology*. 2023;180. doi:10.1016/j.radonc.2023.109483

73. Ma B, Guo J, Chu H, et al. Comparison of computed tomography image features extracted by radiomics, self-supervised learning and end-to-end deep learning for outcome prediction of oropharyngeal cancer. *Physics and Imaging in Radiation Oncology*. 2023;28:100502. doi:10.1016/j.phro.2023.100502

Abstract in Korean

방사선 치료 독성 예측을 위한 대규모 언어 및 영상 기반 통합 모델

방사선 치료(RT)는 항암화학요법 및 수술과 더불어 흉부암 환자에게 중요한 치료 방식이다. RT 는 종양을 정밀하게 표적하는 것을 목표로 하지만, 인접한 정상 조직이 상당한 방사선량을 받을 수 있으며, 그 결과 식도염, 심장 독성, 폐렴 등 방사선 유발 독성이 발생할 수 있다. 이러한 독성은 환자의 삶의 질에 악영향을 줄 수 있으며, 장기 생존 환자의 수가 증가함에 따라 치료 관련 독성을 줄이는 것이 방사선 치료 계획의 핵심 과제로 부각되고 있다.

이러한 독성 관리는 방사선 치료 계획의 여러 단계에서 고려될 수 있다. 시뮬레이션 단계에서는 치료 계획의 기반이 되는 CT 영상이 획득된다. 그러나 호흡으로 인한 움직임은 영상에 아티팩트를 유발하고, 계획된 선량과 실제 전달된 선량 간의 차이를 초래할 수 있다. 이를 줄이기 위해 숨참기 기법이 활용되지만, 기존의 표준 방식은 환자가 오랜 시간 숨을 참아야 하므로 일부 환자에게 부담이 된다. 치료 계획 단계에서는 위험 장기(OAR) 및 종양의 정확한 윤곽선 구획이 필수적이나, 여전히 주요 도전 과제로 남아 있다. 수동으로 수행되는 윤곽선 구획은 많은 시간이 소요되고 관찰자 간 편차가 커 임상 워크플로우의 병목으로 작용할 수 있다. 마지막으로, 치료 계획과 실제 치료 사이의 기간에는 환자 맞춤형 독성 예측 모델을 도입함으로써 임상의가 잠재적인 부작용을 보다 잘 예측하고 대응할 수 있는 의사결정 지원이 가능하다. 그러나 방사선 치료에 대한 환자의 반응이 개별적으로 다르기 때문에 예측 모델을 구축하는 데 어려움이 존재한다.

따라서 본 논문의 목적은 방사선 치료로 유발되는 독성 관리와 관련된 주요 문제를 해결하기 위한 새로운 방법을 개발하는 데 있다. 본 논문은 세 개의 장으로 구성되며, 각 장은 방사선 치료 과정 내 독성 관리를 향상시키기 위한 독립적인 연구 성과를 제시한다. 첫 번째 장에서는 지속적 양압 호흡법(CPAP)이라는 새로운 숨참기 기법의 임상 적용을 다룬다. 해당 기법은 방사선 치료를 받은 유방암 환자를 대상으로 적용되었으며, 기존의 자유호흡 및 깊은 흡기 숨참기 방식과 기하학적, 선량학적으로 비교 분석되었다.

두 번째 장에서는 딥러닝 기반의 자동 분할 알고리즘이 방사선 치료 계획을 효율화하는 도구로서의 역할을 탐색한다. 본 알고리즘은 유방암 환자의 후향적 데이터를 기반으로 적용되었으며, 생성된 분할 결과의 기하학적 정확도를 평가하였다. 또한 기존의 아틀라스 기반 분할 방식과 비교하여 정확성과 효율성의 향상 정도를 분석하였다.

마지막 장에서는 식도암 환자에서 방사선 치료로 유발되는 식도염을 예측하는 다중 모달 예측 모델의 개발을 다룬다. 본 장에서는 영상 정보와 임상 정보를 통합하는 새로운 접근 방식을 도입하였다. 의료 영상으로부터 특징을 추출하기 위해 사전학습된 이미지 인코더를 사용하였으며, 임상 정보는 대형 언어 모델을 통해 반영하였다. 이는 기존의 영상 기반 예측 모델에서 벗어나 다중 모달 기반의 통합 예측 모델을 제안하는 것으로, 식도염 예측의 정확도와 임상적 활용 가능성을 높이고자 한다. 해당 프레임워크는 환자 맞춤 치료를 위한 보다 포괄적인 도구로 기능할 수 있다.

핵심되는 말 : 딥러닝, 대규모 언어 모델, 멀티모달, 예측모델