



HAL
open science

Adaptive strategies based on artificial intelligence for radiotherapy of head-and-neck cancer treatment

Madalina-Liana Costea

► **To cite this version:**

Madalina-Liana Costea. Adaptive strategies based on artificial intelligence for radiotherapy of head-and-neck cancer treatment. Nuclear medicine. Université Claude Bernard - Lyon I, 2023. English. NNT : 2023LYO10017 . tel-04459738

HAL Id: tel-04459738

<https://theses.hal.science/tel-04459738>

Submitted on 15 Feb 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



THESE DE DOCTORAT DE L'UNIVERSITE CLAUDE BERNARD LYON1

Ecole Doctorale N° 160
Electronique, Electrotechnique, Automatique
Discipline : Physique Médicale

Soutenue publiquement le 22/02/2023, par :

Madalina-Liana COSTEA

Titre de la thèse

Adaptive strategies based on artificial intelligence for
radiotherapy of head-and-neck cancer treatment

Devant le jury composé de :

MAINGON Philippe	MD, Professeur, Université Sorbonne, Paris	Président de jury
CHIAVASSA Sophie	HDR, Physicienne médicale, Institut de Cancérologie de l'Ouest, Nantes	Rapporteuse
FRIBORG Jeppe	MD, PhD, Professeur Associé Université de Copenhague, Danemark	Rapporteur
ORKISZ Maciej	Professeur, Université Claude Bernard (UCBL), Lyon	Examineur
SARRUT David	Directeur de recherche CNRS, Lyon	Co-directeur de thèse
BISTON Marie-Claude	HDR, Physicienne médicale, Centre Léon Bérard, Lyon	Directrice de thèse
GREGOIRE Vincent	MD, PhD, Centre Léon Bérard, Lyon	Invite
LAFOND Caroline	HDR, Physicienne médicale, Centre Eugene Marquis, Rennes	Invitée

Table of contents

Abstract	4
Resumé.....	5
List of acronyms.....	6
Introduction.....	8
Chapter 1. Clinical context. Head-and-neck cancer	10
1.1. Cancer Epidemiology	10
1.2. Principle of external beam radiotherapy	15
1.3. Radiotherapy treatment workflow for HN cancer patients	18
1.4. Image guided radiation therapy	22
1.5. Adaptive radiation therapy	27
1.6. Artificial intelligence emergence in radiotherapy.....	35
1.7. Medical image automatic segmentation.....	38
1.8. Automated treatment planning	53
1.10. Synthetic CT image generation.....	58
1.11. Objectives of the study.....	66
Chapter 2. Evaluation of fully automated a priori MCO treatment planning in VMAT for head-and-neck cancer	67
Synthesis.....	76
Chapter 3. Comparison of atlas-based and deep learning methods for organs at risk delineation on head-and-neck CT images using an automated treatment planning system.....	77
3.1. Abstract	77
3.2. Introduction.....	77
3.3. Materials and methods	78
3.4. Results	82
3.5. Discussion	92
3.6. Conclusions.....	94
3.7. Synthesis.....	94
Chapter 4. Evaluation of different algorithms for automatic segmentation of head-and-neck lymph nodes on CT images	96
4.1. Abstract	96
4.2. Introduction.....	96
4.3. Materials and methods	97
4.4. Results	100
4.5. Discussion	108
4.6. Conclusion	109

4.7. Synthesis.....	109
Chapter 5. Evaluation of different methods for synthetic CT image generation from daily CBCT images	111
5.1. Introduction.....	111
5.2. Materials and methods	111
5.3. Results	115
5.4. Discussion	118
5.5. Conclusion	119
5.6. Synthesis.....	120
Chapter 6. Conclusions and perspectives of the study	121
Résumé étendu	123
Bibliography.....	133

Abstract

Intensity-modulated radiation therapy (IMRT) techniques with highly conformal dose distributions and steep dose gradients are the standard for radiotherapy (RT) of head-and-neck (HN) tumors. To be safely delivered, accurate patient positioning and accurate delineation of targets and organs-at-risk (OARs) is necessary. Manual contouring is performed on computed tomography (CT) images, and typically takes 2-3 hours for a skilled physician. Similarly, manual treatment planning is complex and the plan quality is highly dependent on the planner's experience. Furthermore, image-guided RT allows monitoring of the daily tumor and normal tissue changes, during the treatment course. To facilitate adaptive RT (ART), artificial intelligence (AI) solutions have been explored.

The goal of this PhD was to investigate different solutions for improving the treatment workflow of HN cancer patients and the facilitate implementation ART. We considered automatic segmentation (AS) of CT images, automated treatment planning (auto-planning), and synthetic-CT image (sCT) generation from cone-beam CT (CBCT) images. First, an a priori multicriteria optimization (MCO) algorithm for auto-planning based on the elaboration of a "wish-list" was evaluated and compared with manual planning. It was concluded that high-quality volumetric art therapy (VMAT) plans could be generated with a robust wish-list of dose objectives and constraints. Moreover, OARs sparing was superior compared with manual planning.

The second objective was to evaluate and compare six different methods for AS of OARs on CT images. Their performance was assessed with regard to resource demand, contour geometrical accuracy, computational time, and time needed to perform manual corrections. Additionally, auto-planning was used to evaluate the dosimetric impact of using the AS contours with and without manual correction. Four atlas-based and two deep learning (DL) solutions were investigated and compared for 14 OARs. One DL solution was trained with mono-centric data and the other one was a commercial solution trained with multi-centric data. Overall the results showed that the DL solutions had the best performances over the majority of the OARs. However, the contours generated by the mono-centric DL solution were the fastest to correct. Moreover, the dosimetric study demonstrated no significant impact on the radiation doses. Furthermore, the same six solutions were evaluated and compared for the AS of three lymph node (CTVn) levels on HN CT images. The same performance evaluation metrics were used. Results showed the superiority of the DL solutions and the preference of the experts for the contours obtained with the multi-centric DL solution, which were also the fastest to correct. Regarding the dosimetric consequences, a significant underdosage of the nodal targets was observed regardless of the AS method used.

The last objective was to investigate different methods for sCT generation based on CBCT images. One DL solution was investigated and compared against: adapted CBCT Hounsfield Units and electron densities (HU-ED) calibration curve, density assignment method (DAM) and deformable image registration (DIR). Image accuracy was evaluated by calculating the differences in HU numbers between the CT and the sCT. The accuracy of dose calculation was evaluated by comparing plans calculated on the CT with the plans calculated on the sCT. Results showed that DIR method was the most precise, followed by DL, DAM and the adapted HU-ED method.

In conclusion, for the delineation task, DL solutions were the most attractive. The wish-list based auto-planning algorithm produced high-quality plans, but the computational time was large. Lastly, DIR and DL solutions can offer fast sCT generation from CBCT images, that enable to perform dose calculations on the anatomy of the day. The human intervention is not yet to be overlooked, however the aspects investigated are key points for ART of HN patients.

Resumé

Les techniques de radiothérapie à modulation d'intensité avec des distributions de dose hautement conformes sont la norme pour la radiothérapie (RT) des tumeurs HN. Pour être délivré en toute sécurité, le positionnement précis du patient et le contournage correct des volumes cibles et des organes-à-risque (OAR) sont nécessaires. Le contournage manuel est réalisé à partir d'images tomodynamométrie (CT) et prend généralement 2-3 heures pour un médecin qualifié. De plus, la planification manuelle du traitement est complexe et la qualité du plan dépend de l'expérience du planificateur.

La radiothérapie guidée par l'image permet de surveiller les évolutions de la tumeur et des tissus normaux pendant le traitement. Pour faciliter la RT adaptative (ART), des solutions d'intelligence artificielle (IA) ont été explorées.

L'objectif de la thèse était d'étudier différentes solutions pour améliorer le flux de travail de traitement des patients HN et permettre l'ART. Nous avons considéré la segmentation automatique (AS) des images CT, la planification automatique (auto-planning), et la génération d'images CT synthétiques (sCT) à partir d'images de tomodynamométrie conique (CBCT) du jour.

Au premier, un algorithme d'optimisation multicritères (MCO) a priori pour l'auto-planning basée sur l'élaboration d'une "wish-list" a été évalué et comparé à la planification manuelle. Il a été conclu que des plans d'arthérapie volumétrique (VMAT) de haute qualité peuvent être générés avec une "wish-list" robuste d'objectifs et de contraintes de dose. De plus, l'épargne des OAR était supérieure à celle des plans optimisés manuellement.

Le deuxième objectif était d'évaluer et de comparer les performances de six méthodes différentes pour l'AS des OAR et niveaux ganglionnaires (CTVn) sur les images CT. De plus, l'auto-planning a été utilisée pour évaluer l'impact dosimétrique de l'utilisation des contours d'AS. Quatre méthodes basées sur une bibliothèque d'atlas (ABAS) et deux solutions d'apprentissage profond (DL) ont été étudiées et comparées pour 14 OAR et 3 niveaux CTVn. L'une des solutions DL a été entraînée avec les données monocentriques et l'autre était une solution commerciale entraînée avec des données multicentriques. Les résultats ont montré que les solutions DL avaient les meilleures performances sur la majorité des structures considérées. La solution DL monocentrique a été plus performante pour les OAR, par contre la DL multicentrique était meilleure pour les CTVn. De surcroît, l'étude dosimétrique a démontré un impact négatif seulement pour les plans générés avec des AS CTVn.

Le dernier objectif était d'étudier différentes méthodes pour la génération des images sCT à partir d'images CBCT. Une solution DL a été étudiée et comparée à d'autres méthodes: correspondance entre nombres Hounsfield de CBCT et densités électronique (HU-ED), affectation de densités (DAM) et recalage élastique entre l'image CT et CBCT (DIR). En rapport avec l'image CT, la précision des nombres HU a été analysée sur la base de l'erreur moyenne (ME) et de l'erreur absolue moyenne (MAE). La précision du calcul de la dose a été évaluée en comparant les distributions de dose calculées sur le CT aux doses calculées sur les sCT. Les résultats ont montré que la méthode DIR était la plus précise, suivie de DL, DAM et HU-ED.

En conclusion, pour la tâche de contournage, les solutions DL étaient les plus intéressantes. L'algorithme d'auto-planning basé sur la "wish-list" produit des plans de haute qualité, mais le temps de calcul est trop important pour être intégré dans l'ART. Enfin, les solutions DIR et DL peuvent offrir une génération rapide d'images sCT à partir d'images CBCT, ce qui permet d'effectuer des calculs de dose sur l'anatomie du jour. L'intervention humaine n'est pas encore à négliger, mais les aspects étudiés dans cette thèse sont des points clés pour l'ART des patients HN.

List of acronyms

3D-CRT	3-dimensional conformal radiotherapy
ADMIRE	Advanced Medical Imaging Registration Engine
AI	artificial intelligence
AJCC	American Joint Committee on Cancer
ANN	artificial neural network
ART	adaptive radiation therapy
AS	automatic segmentation
auto-planning	automated treatment planning
BMI	body mass index
CBCT	cone-beam computed tomography
CNN	convolutional neural network
CT	computed tomography
CTV	clinical tumor volume
DAM	density assignment method
DIR	deformable image registration
DNN	deep neural network
DVH	dose-volume histogram
ED	electron densities
FCNN	fully connected neural networks
FOV	field of view
GAN	generative adversarial network
GTV	gross tumor volume
HAS	hybrid automatic segmentation
HN	head-and-neck
HT	helical TomoTherapy
HU	Hounsfield Units
HU-ED	conversion between HU and ED
ICRU	Commission on Radiation Units and Measurements
IGRT	image-guided radiation therapy
IMRT	Intensity-modulated radiation therapy
IOV	inter-observer variation
LINAC	linear accelerator
LQ	linear-quadratic model

MCO	multicriteria optimization
ML	machine learning
MLC	multi-leaf collimator
MRI	magnetic resonance imaging
MV	majority voting
OAR	organs-at-risk
pCT	planning CT
PRV	planning organ-at-risk volume
PTV	planning target volume
QA	quality assurance
RT	radiation therapy
sCT	synthetic-CT
STAPLE	simultaneous truth and performance estimation
TNM	tumor-node-metastasis
TPS	treatment planning system
VMAT	Volumetric Modulated Arc Therapy

Introduction

Cancer is one of the biggest causes of death worldwide and can develop in several regions in the head-and-neck (HN): larynx, oropharynx, nasopharynx, hypopharynx, thyroid, salivary glands, oral cavity and lips. In 2020, among the 1 518 133 patients diagnosed with HN cancer, 34% did not survived [1]. External radiation therapy (RT) is one of the most efficient treatments for this type of cancerous tumors. Its principle is to deliver high-energy ionizing radiation (MV) with a linear accelerator (LINAC), in order to induce DNA damage in the cancer cells and block their ability to multiply. Intensity-modulated radiation therapy (IMRT) techniques with highly conformal dose distributions and steep dose gradients are the standard of care for RT of HN tumors by ensuring maximum target coverage and organs-at-risk (OARs) sparing. Commonly, a cumulative radiotherapy dose of 70Gy is delivered with curative intent over several weeks in daily fractions of 1.8 – 2.0Gy. Acquisition of the planning computed tomography (pCT) scan is a prerequisite in order to define the patient's reference positioning, have access to the patient's anatomy and to the electronic densities (ED) of the tissues. On the pCT image, the radiotherapist defines the planning target volume (PTV) and the OARs, so that dose calculations can be performed. Both contouring and planning require high accuracy. Nonetheless they are tedious and prone to intra and inter-observer variations (IOV). Moreover, anatomical changes (e.g. weight loss, tumor shrinkage, displacement of OARs) may occur during the RT treatment, that can cause differences between the planned and actual delivered doses.

Adaptive RT (ART) strategies have been developed to correct for intra-fractional anatomical variations. Ideally, 3 dimensional (3D) in-room imaging devices primarily used for patient positioning can be used to assess daily anatomical deformations and to perform dose calculation on the anatomy of the day. Low energy cone-beam CT (CBCT) systems integrated on the LINAC machine are widely spread and used for patient set-up verifications. However, their image quality is not suitable for dose calculations due to several limitations (e.g. image artifacts, inconsistency of the Hounsfield Unit (HU) numbers, and restricted field of view (FOV)). With the advent of artificial intelligence (AI) solutions, several applications have been proposed to facilitate the implementation of ART. Among them, automatic image segmentation (AS), automated treatment planning (auto-planning) and synthetic-CT (sCT) image generation from CBCT images are being discussed in this thesis manuscript.

The goal of this thesis was to investigate several automated solutions for different steps in the RT workflow of HN cancer patients. First, an auto-planning solution was evaluated. Then the performance of several AS solutions was compared for HN OARs and HN lymph node levels. Finally, several methods were investigated that enable dose calculations on daily CBCT images. The manuscript represents the work done in the last three years and is organized in six chapters.

The first chapter presents the context of the study. In the first part is described the standard process of a RT treatment, the modalities used in image-guided RT (IGRT), and the concept of ART with focus on HN cancer treatment. Furthermore, the emergence of AI solutions in RT is introduced with a focus on state-of-the-art AS methods, auto-planning solutions, and sCT generation methods. In the last part of the chapter are summarized the objectives of the thesis.

The second chapter describes the evaluation of the performance of an auto-planning solution against manually optimized treatment plans. This work represents a first contribution, as second author, to the article that has been published in the European Journal of Medical Physics in 2021 [2].

The third chapter presents the work of the second contribution, the evaluation of the performances of six AS methods for OARs segmentation on HN CT images. This work was published in the Radiotherapy and Oncology Journal [3]. The performance evaluation was based on resource demand, geometrical accuracy, computational time and the time needed for manual corrections. Additionally, the dosimetric impact on RT dose distributions was calculated using the auto-planning solution previously validated.

Similarly, in the fourth chapter, the performances of the same six AS methods were evaluated on HN lymph nodes levels (CTVn), which are typically irradiated as elective target volumes. The results of the study are presented as form of an article that will shortly be submitted also to the Radiotherapy and Oncology Journal.

In the fifth chapter are presented results from the evaluation of four methods for sCT generation based on daily CBCT images. One DL-based solution was investigated and compared to other methods proposed in the literature. With respect to the reference pCT, image uncertainty and dose calculation accuracy were measured.

Finally, chapter six summarizes the thesis conclusions and future perspectives.

The work done during these three years of thesis was financed by Elekta LTD, and has been conducted at the RT department of the Léon Bérard Cancer Center, in the TOMORADIO team of the CREATIS laboratories.

Chapter 1. Clinical context. Head-and-neck cancer

1.1. Cancer Epidemiology

Cancer is a disease characterized by an unwanted and uncontrolled growth of cells that have developed from normal body cells and which have structural and functional mutations. The danger of malignant growth is the ability to invade and infiltrate in the surrounding tissues, blood vessels, lymph vessels and other body cavities. Worldwide, over 18 million patients are diagnosed with cancer each year and only 45% survive [1,4]. After cardiovascular diseases, cancer represents the main cause of mortality. Cancer incidence and mortality can be visualized in Figure 1.1-1 as results of global cancer statistics from 2020 [1].

Cancer incidence and mortality global statistics 2020

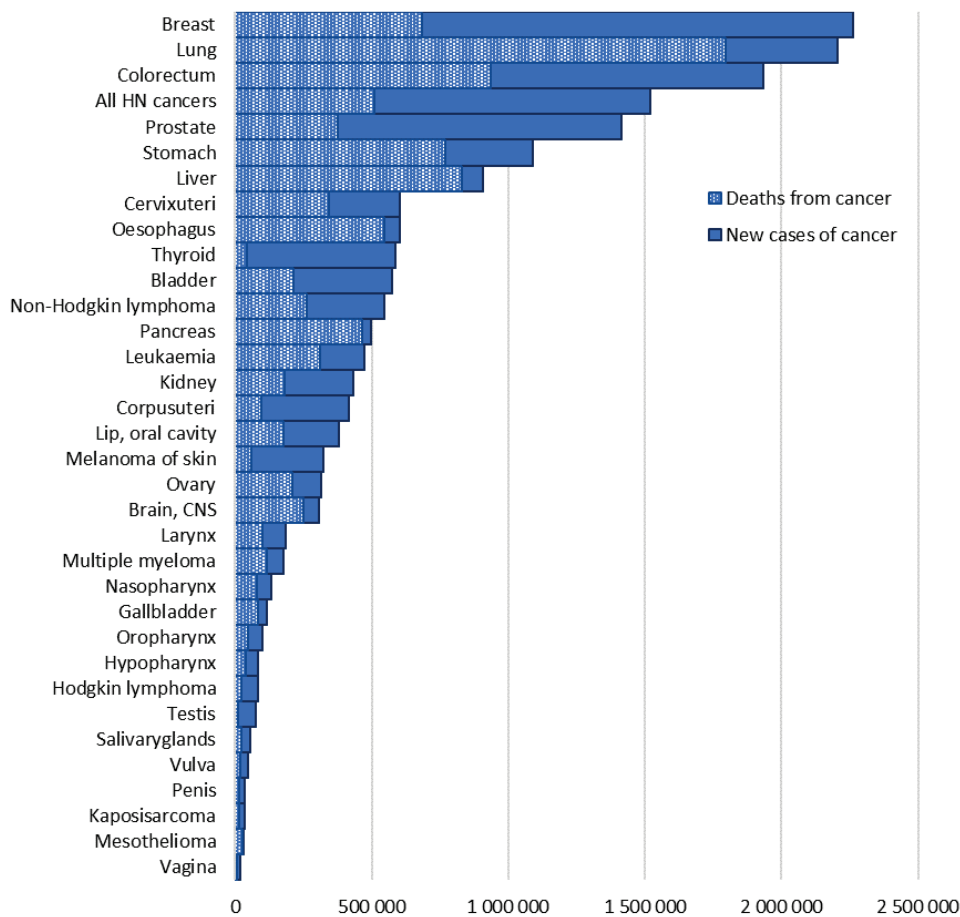


Figure 1.1-1 Cancer global statistics 2020 [1]; HN = head-and-neck; CNS= central nervous system

The most frequent reported cases are the breast and lung cancers, both representing approximately 12% from the total number of new cancers. Lung patients however have higher mortality (81% compared with 30% for breast cancer). Colorectum cancers are following with 10% incidence and 48% mortality. Head-and-neck (HN) tumors are the next, that in 2020 contributed with 8.2% and 5% to the global cancer incidence and mortality, respectively. More precisely, among the 1 518 133 patients diagnosed with HN

cancers, 34% did not survive. Malignant tumors in HN are called sarcomas because they originate from connective tissue such as skin, fat, muscle, cartilage, nerves and bones. The main categories of HN sarcomas are developed in the larynx, oropharynx, nasopharynx, hypopharynx, thyroid, salivary glands, oral cavity and lips. In Figure 1.1-2 is illustrated the HN anatomy with the primarily tumor regions and their associated percentage of incidence.

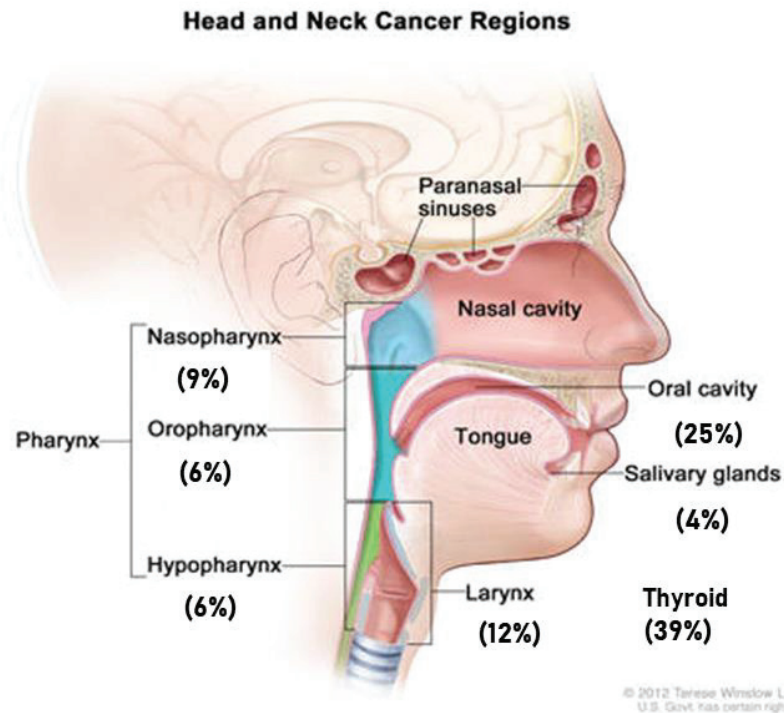


Figure 1.1-2 Primarily tumors regions in head-and-neck (Figure credit: © Terese Winslow)

Among the different localizations, thyroid cancer is the most common, representing 39% from the incidence of HN cancers, followed by oral cavity cancers that include also the lip cancers (25%) and larynx cancer (12%). At the same time, thyroid cancers have the highest survival (93%) while the oral cavity and lip cancers have a high mortality (57%), causing the highest proportion of deaths (35%) among the HN cancers. Nasopharynx cancers have however, the highest mortality (60%), but due to their low incidence (9%), they contribute with only 16% of the total mortality in HN cancers.

The most significant risk factors for HN cancers are alcohol and tobacco consumption (including passive smoking). Moreover, when used together the risk of developing cancer rises [5]. Infection with human papillomavirus (HPV) is also an increasing risk factor particularly for oropharyngeal cancer that includes the tonsils and the base of the tongue [6,7].

Symptoms of HN cancer may include: lump in the neck, sore in the mouth or throat that does not heal, difficulty and/or pain in swallowing, troubles breathing and speaking as well as changes of the voice.

In finding the most appropriate treatment for cancer, staging of the tumors is one of the most important steps. Typically, it relies on results from several physical exams such as endoscopies, biopsies and imaging sets. American Joint Committee on Cancer (AJCC) developed the tumor-node-metastasis (TNM) system to evaluate the anatomic extent of the disease: where T refers to the size of the primary tumor, N describes involvement of the lymph nodes and M indicates whether the cancer has spread (metastasized) in other areas of the body. The 4 stages for HN cancer are presented in Figure 1.1-3, whereas classification might vary according to the specific anatomic subsite [8,9]. In general, primary

tumors are classified as T1 to T3 by increasing size, whereas T4 usually represents invasion of another structure such as bone, muscle, or root of the tongue. Lymph nodes are staged by size, number, and localization (ipsilateral or contralateral to the primary tumor location). Distant metastases (M1) are more common in patients with advanced lymph node stage.

From a therapeutic point of view, HN cancers are challenging to treat because many organs in this region are associated with physiological functions such as respiration, communication and nutrition. Severe consequences such as functional impairments and structural disfigurements can compromise considerably the quality of life and social integration of the patient. Therefore, the management of HN cancer patients imposes a multidisciplinary treatment approach involving surgery, external or internal radiation therapy and systemic treatments. In order to achieve the therapeutic goal, often the prescription includes a combination of these treatment options.

Surgery

Surgical treatment has the attempt of radical resection of the tumor and sometimes might include removal of structures with important functional functions. For instance, after a total laryngectomy (removal of the larynx), patients can have a decreased thyroid gland function that will require hormone medication. Depending on the type and location of the surgery, other common side effects include a temporary or permanent loss of voice, impaired speech, loss of hearing or swallowing difficulties. The removal of the lymph nodes can also cause stiffness in the shoulders. Complications causing breathing problems from swelling of the mouth and throat area, can be managed by tracheostomy, which creates a hole in the trachea to facilitate breathing. At the same time, some people that experience facial deformations may require reconstructive surgery.

Radiation therapy

Radiation therapy (RT), also called radiotherapy, is a local treatment aiming to kill the cancer cells by damaging their DNA through the use of ionizing radiation. It is given to most of the HN cancers patients either with curative (tumor annihilation) or palliative (to reduce the symptoms) intent. It is often combined with surgery and/or chemotherapy. Most commonly, radiation is delivered from a source outside of the patient body and it's called external beam radiotherapy. When a radioactive source is inserted inside the body, the treatment is called internal radiation therapy or brachytherapy. Superficial cancers such as lip cancers can benefit from this method of delivering radiation.

Acute side effects from radiation consist of redness or skin irritation, mucositis and dysphagia while long term complications include xerostomia (dry mouth), loss of taste, decreased mobility of the mouth, voice hoarseness, swallowing dysfunction, second malignancies, dysphagia and neck fibrosis [10–13]. Radiation can also cause dental problems such as tooth decay, loss of hearing or lymphedema if the lymph nodes are damaged. Moreover, every patient receiving radiation in HN should regularly check their thyroid function and get hormone medication in case of hypothyroidism.

Systemic treatments

Systemic treatments involve drug therapies that work throughout the whole body and can be given as an injection, infusion or oral medication. This category includes chemotherapy, targeted therapy and immunotherapy.

Chemotherapy aims to stop tumor growth by metabolic inhibition of the DNA synthesis and is typically reserved for patients whose cancers have metastasis. Common drugs used to treat HN cancers include: methotrexate, cisplatin and other platinum analogues. In general, chemotherapy may induce side effects such as nausea, vomiting, diarrhea, hair loss, poor appetite and weight loss.

Targeted therapies act on the cancer's specific genes and proteins responsible for the growth, division and spread of the tumoral cells. Most of the treatments target the epidermal growth factor receptor protein. Cetuximab is the only FDA approved targeted molecule for the treatment of HN cancers in addition to radiotherapy for locoregionally advanced tumors.

Immunotherapy has recently emerged from the growing understanding of the role of the immune system in tumor suppression. The two FDA approved immunotherapy drugs for the treatment of HN cancers are nivolumab and pembrolizumab. Common side effects can include skin reactions, sickness, diarrhea and weight changes.

Upon diagnosis and staging, HN cancers can be categorized in three main clinical groups: those with localized disease, those with locally or regionally advanced disease (lymph node positive) and those with recurrent and/or metastatic disease. Localized diseases, explicitly stage I and stage II lesions without detectable lymph node involvement or distant metastases, are usually treated with curative intent either by surgery or radiotherapy depending on the anatomical localization. To preserve voice function, radiotherapy might be preferred for the laryngeal cancer, and for small lesions in the oral cavity, surgery may be chosen in order to avoid long-term complications caused by radiation. For this category, the overall survival rate is high and most recurrences occur within the first 2 years and are typically local. Locally or regionally advanced diseases which involve large primary tumor with or without lymph node involvement, are present in the majority of the HN cancer patients. For these patients, a curative treatment intent will require a combined treatment modality. In most cases, and most efficiently, after surgery, radiotherapy is given with or without concomitant chemotherapy. Patients with recurrent and/or metastatic disease receive (with only few exceptions) a treatment with palliative intent. This can be RT for pain control but most often is chemotherapy having a median survival of 8-10 months after administration.







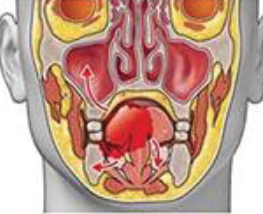

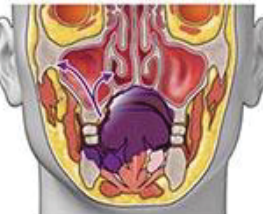
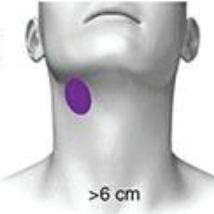
Definition of TNM				Stage groupings		
Stage I T1 	Tumor \leq 2 cm in greatest dimension without extraparenchymal extension	N0 	N0- No regional lymph node metastasis	T1	N0	M0
Stage II T2 	Tumor \geq 2 cm but not more than 4 cm in greatest dimension without extraparenchymal extension	N0 	N0- No regional lymph node metastasis	T2	N0	M0
Stage III T3 	Tumor \geq 4 cm and/or tumor having extraparenchymal extension	N1  \leq 3 cm	N1- Metastasis in a single ipsilateral lymph node, \leq 3 cm in greatest dimension	T3 T1 T2 T3	N0 N1 N1 N1	M0 M0 M0 M0
Stage IVA T4a 	Tumor invades skin, mandible, ear canal, and/or fascial nerve	N2  \leq 6 cm	N2a- Metastasis in a single ipsilateral lymph node, $>$ 3 cm but \leq 6 cm N2b- Metastasis in a multiple ipsilateral lymph node, none $>$ 6 cm N2c- Metastasis in a bilateral or contralateral lymph nodes, none $>$ 6 cm	T4a T4a T1 T2 T3 T4a	N0 N1 N2 N2 N2 N2	M0 M0 M0 M0 M0 M0
Stage IVB T4b 	Tumor invades skull base and/or pterygoid plates and/or encases carotid artery	N3  $>$ 6 cm	N3- Metastasis in a lymph node $>$ 6 cm in greatest dimension	T4b Any T	Any N N3	M0 M0
Stage IVC		M1		Any T	Any N	M1

Figure 1.1-3 Staging of head-and-neck cancer tumors based on TNM system [11]

1.2. Principle of external beam radiotherapy

The principle of external beam RT is inducing damage inside the cancer cells while limiting the effects to the normal healthy tissues. Due to an atypical cell cycle, with an accelerated division stage, cancer cells are more sensitive to radiation and therefore can be lethally affected by radiation, whereas the effect on normal cells is sublethal.

The unit of gray (Gy) is used to quantify the amount of energy absorbed per unit mass. To decrease acute and especially late toxicity to the surrounding normal tissue exposed to radiation, the desired doses are divided in smaller fragments over time, process called fractionation. The reasoning behind fractionation is based on the relative biologic effectiveness of radiation described by 5 radiobiological determinants, the so-called '5Rs' of radiotherapy: repair, repopulation, redistribution, reoxygenation, and radiosensitivity [14]. By delivering small doses of radiation, cells are allowed to repair the sublethal damage. Unlike normal tissue, malignant cells have often suppressed the cell repair pathways thus they cannot efficiently repair after radiation caused damage. Repopulation is the increase in cell division after radiation. Redistribution refers to the cells' cycles, which makes them more resistant or more sensitive to radiation damage. Reoxygenation is the phenomena by which hypoxic cells recover after radiation. It is not effective in the malignant tumors that are often characterized by acute or chronic hypoxia caused by their rapid proliferation that outgrows their surrounding vascularization. This condition allows a higher degree of tumors cell annihilation compared to normal tissues that have normal oxygen levels. Lastly, radiosensitivity is an intrinsic cell factor leading to variations in response to radiation among tissues, organs, or organisms.

The linear-quadratic model (LQ) is used to describe the cell survival probability as a function of radiation dose (S), where α/β ratio characterizes the sensitivity of a certain tissue to fractionation [15] (Figure 1.2-1). Following the LQ model, curative treatments are commonly delivered in doses of 2Gy per fraction over several weeks and hypo-fractionated regimen consist in delivering doses higher than 2Gy in one treatment session.

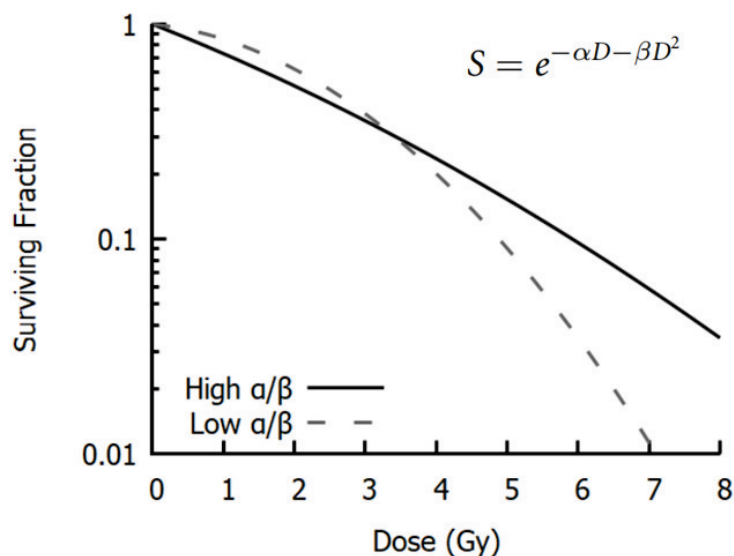


Figure 1.2-1 Illustration of linear quadratic model. High α/β cell lines have nearly-constant rates of cell killing with increasing dose, while low α/β lines show a pronounced curvature, with greater killing per unit dose at higher doses [15]

Several radiation qualities can be used for treatment, namely electrons, photons or charged particles such as protons and carbon ions. Each of them is presenting particular penetration capabilities caused by their interaction with matter. The linear accelerator (LINAC) is the most widely used equipment for radiation delivery in RT. It is able to produce high intensity X-rays, by accelerating electrons towards a tungsten target. Most of the electron's kinetic energy is transformed into heat and a small fraction of it is emitted in the form of X-ray photons. If the target is removed, the electron beam can also be used for treatment. Protons or carbon ions are also being effectively used for radiation treatments but they necessitate stricter conditions and bigger facilities (such as cyclotrons, where the heavy particles can be accelerated), which implies higher costs. Their advantage over photons and electrons consists in the peak of maximum dose deposition, which can reach deeper depths in the patient with less harm to the traversed healthy tissues.

Once a radiation beam is produced in a LINAC, it can be modeled with the help of several elements inside of the treatment head that include filters, blocks and collimators. With the introduction of the multi-leaf collimator (MLC) in the LINAC design, more precise shaping of the radiation field to the outline of the tumoral targets was made possible (see Figure 1.2-2). This allowed the introduction of IMRT.

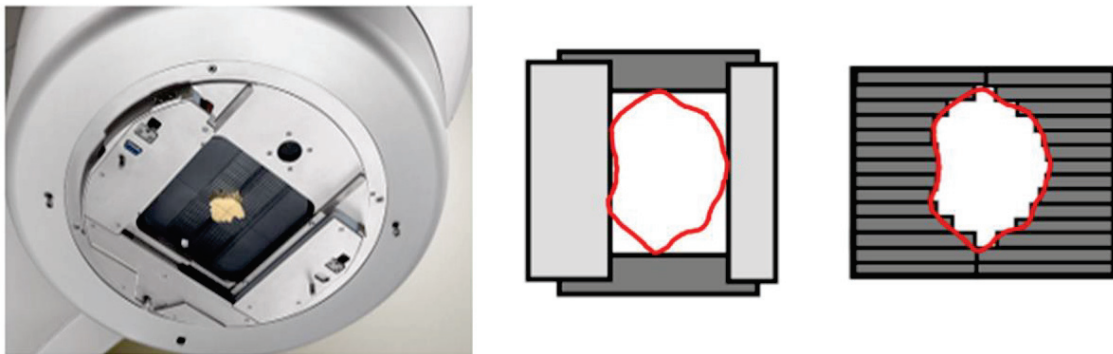


Figure 1.2-2 LINAC treatment head with rectangular blocks and multi leaf collimators

Based on the available equipment, different treatment planning strategies are possible. Among them, 3-dimensional conformal radiotherapy (3D-CRT) involves stationary radiation beams aimed at the tumor from several directions. This approach is called direct treatment planning because the planner chooses the fixed parameters (field intensity, angle, collimator angle, jaws, wedges) and the MLC configuration (following the shape of the target) for each radiation field. The superposition of multiple radiation fields will yield the desired dose to the target. IMRT is a more sophisticated approach that uses inverse optimization starting from dose-volume objectives introduced by the planner [16]. Each beam direction is divided into multiple segments in order to achieve highly conformal dose distributions (see Figure 1.2-3). It can be used in step-and-shoot or sliding-window mode, which defines the way the MLC will move. Similarly, intensity modulation can be delivered in a continuous manner, using volumetric modulated arc therapy (VMAT) technique [17], where the treatment head moves around the patient while the radiation beam is on. Compared to IMRT, VMAT allows faster delivery of the planned treatments [18].

IMRT became the standard modality for RT treatments of complex cases such as HN cancers where the radiation fluence has to be optimized in function of multiple targets and OARs. Moreover, it was shown that IMRT can improve the long-term quality of life, by reducing xerostomia and dysphagia in HN patients [19].

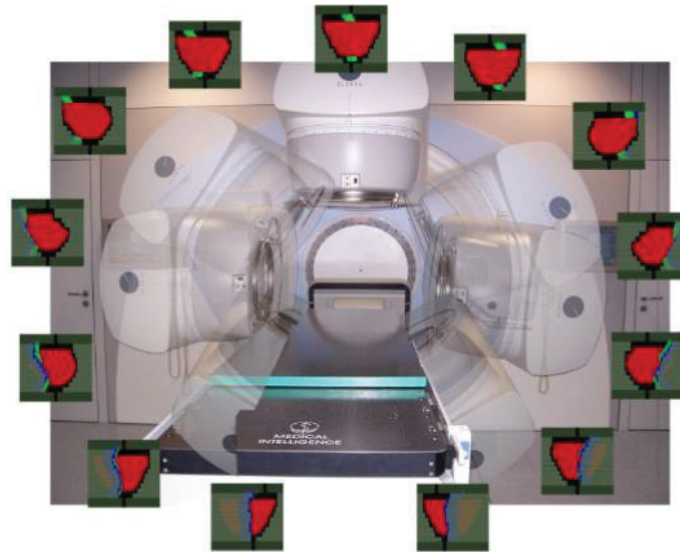


Figure 1.2-3 Illustration of radiation field modulation

From a dosimetric point of view, VMAT and IMRT plans for HN patients present comparable results and the target volume coverage depends mostly on the planner's expertise and the treatment planning algorithm employed. VMAT plans however demonstrated improved sparing of the normal tissues, reduction of the monitor units (MU, the measure of output of a LINAC) with subsequent reduction of the treatment delivery times [20–22]. The high risk of developing secondary cancer due to high number of MU is however a constant challenge for both intensity modulation techniques. Therefore, to be safely delivered, high accuracy in volume definition as well as high precision in patient positioning during the entire course of the treatment is crucial.

1.3. Radiotherapy treatment workflow for HN cancer patients

RT remains one of the most cost-effective cancer treatment modalities. Commonly, a cumulative dose of 70Gy is delivered with curative intent over several weeks in daily fractions of 1.8–2Gy. Altered fractionation regimes are also considered within this group of patients [23]. Often the HN cancer treatment requires more than one target volume to be irradiated in order to achieve the therapeutic goal. A typical dose prescription for HN patients at Léon Bérard Cancer Center consists in delivering 70Gy to the PTV associated to the primary tumor, and 54.25Gy to the PTV associated to the prophylactic nodal target, in 35 fractions of 2Gy. Other treatment strategies may separate the planned target volume into 3 distinct dose levels intervals: low dose [54 – 56Gy], intermediate dose [60 – 63Gy] and high dose [66 – 70Gy]. Typically, the external beam RT treatment can be decomposed in four major steps as illustrated in Figure 1.3-1 and detailed later.

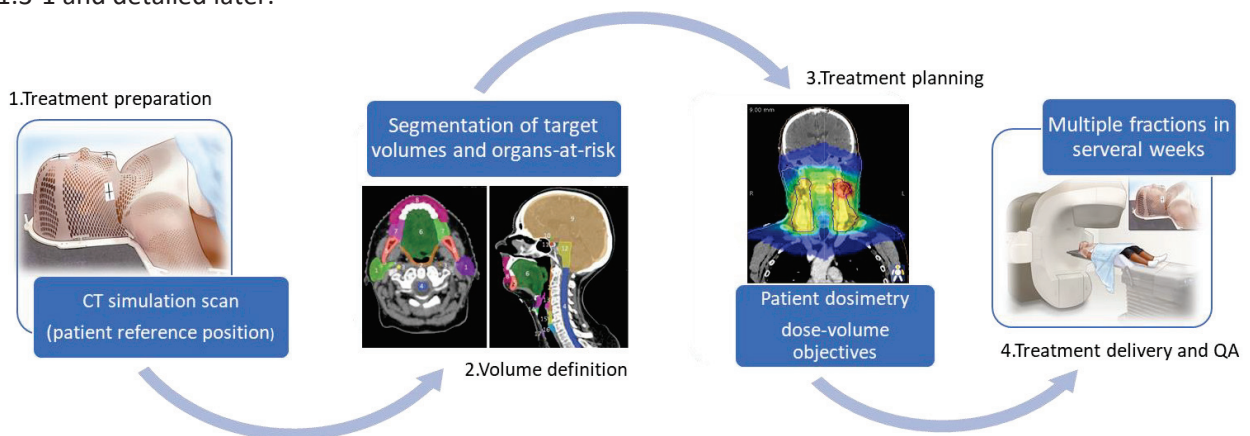


Figure 1.3-1 The radiotherapy treatment chain decomposed in 4 major steps

1. Treatment preparation

Imaging of the patient lies at the beginning of the treatment chain and is responsible for a correct target detection leading to an appropriate dose prescription. Moreover, it defines the reference positioning of the patient during the whole treatment. To ensure a reproducible positioning of the patient, different immobilization devices, markers or skin tattoos are being used. Nonetheless, patient must be fairly comfortable to maintain a stable position during each treatment session. For the HN patients, a personalized thermoplastic mask with five-clip immobilization is recommended [24].

Generally, patient imaging consists of a CT scan either alone, or in combination with other imaging modalities for better tissue visualization such as magnetic resonance imaging (MRI), ultrasound or positron emission tomography (PET). The CT scan is a prerequisite because it contains the relationship between the image intensity information, expressed in Hounsfield Units (HU) and the correspondent electron density (ED) information. This relative HU-ED relationship is defined in the commissioning of the CT imaging system and it is essential for the dose calculation algorithm.

A contrast agent injection is recommended before the CT-scan acquisition for a good visualization of the HN anatomy [25]. Usually a dose of 90ml iodine solution is administered in 2 steps with 45 seconds pause in between.

2. Volume definition

The radiation oncology physician, dosimetrist or planner are responsible for (often manually) contouring on the CT images, slice-by-slice, all the necessary volumes including the tumoral targets and

the OARs. This step in which the image is divided into anatomical groups is also called image segmentation, annotation or labeling. To enable consistency between contouring practices, international delineation guidelines and recommendations have been established for both targets and OARs [26–30]. Additionally, the International Commission on Radiation Units and Measurements (ICRU), provides instructions and various concepts for defining the target volumes and OARs [31–33]. For the target volumes, the volume expansion starts from the gross tumor volume (GTV) that is the visible and/or palpable macroscopic part of the tumor (Figure 1.3-2). The clinical tumor volume (CTV) incorporates the GTV and an additional margin to account for the tissue microscopic infiltration of the tumors.

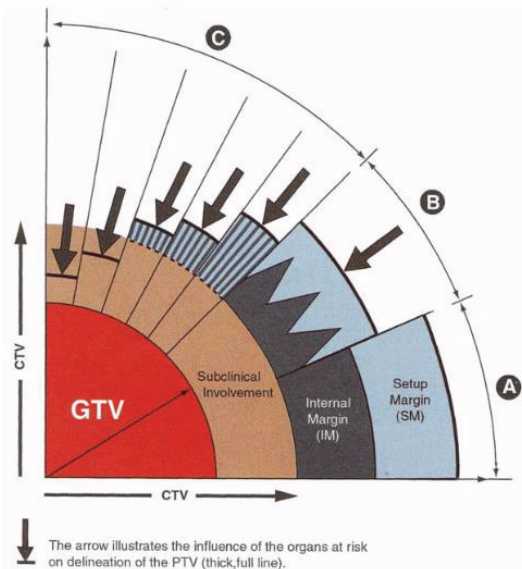


Figure 1.3-2 Margin concepts proposed by ICRU reports 50 and 62 [31]

Like for the GTV, several CTVs can be defined including the primary tumor volume (CTVt) and the HN lymph nodes (CTVn). Several recommendations can be found for the delineations and the selection of these primary tumor areas and lymph nodes areas at risk (Figure 1.3-3) [27–30]. Lastly, with the goal of ensuring adequate CTV dose coverage, ICRU defines the concept of PTV, which is obtained by applying additional margin that accounts for the patient positional uncertainties linked to internal movements or patient set-up [33]. Depending on the tumor localization, the choice of treatment immobilization and the imaging system used for repositioning, isotropic margins between 2 and 5mm are recommended [34].

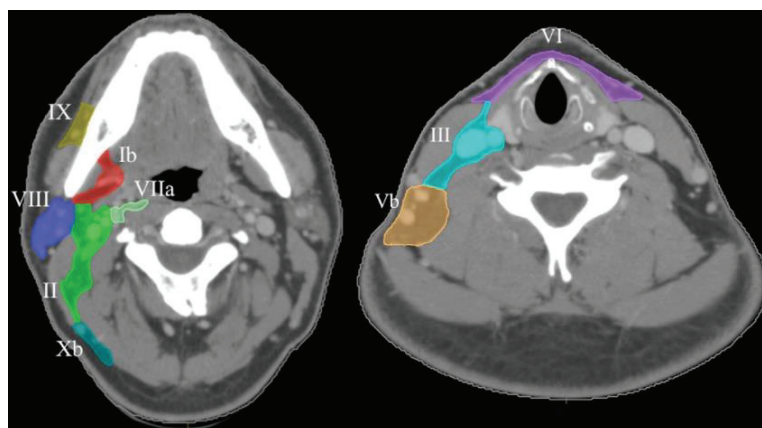


Figure 1.3-3 Various node areas delineated on axial slices of contrast enhanced CT image [25]

Similar as for the PTV, margins are applied to the OARs to account for the same patient positioning uncertainties and this concept is called planning organ-at-risk volume (PRV). Margins are typically recommended for the “serial-like” organs that include the spinal cord, brainstem, optical nerves, chiasma or brachial plexus. For the “parallel organs” such as the parotid glands or the constrictor muscle, this margin can be zero. For both PTV and PRV, 4mm margin strategy is used in the radiotherapy department at Léon Bérard Cancer Center. A visualization of OARs and lymph nodes (CTVn) volumes typically contoured on HN CT images is presented in Figure 1.3-4.

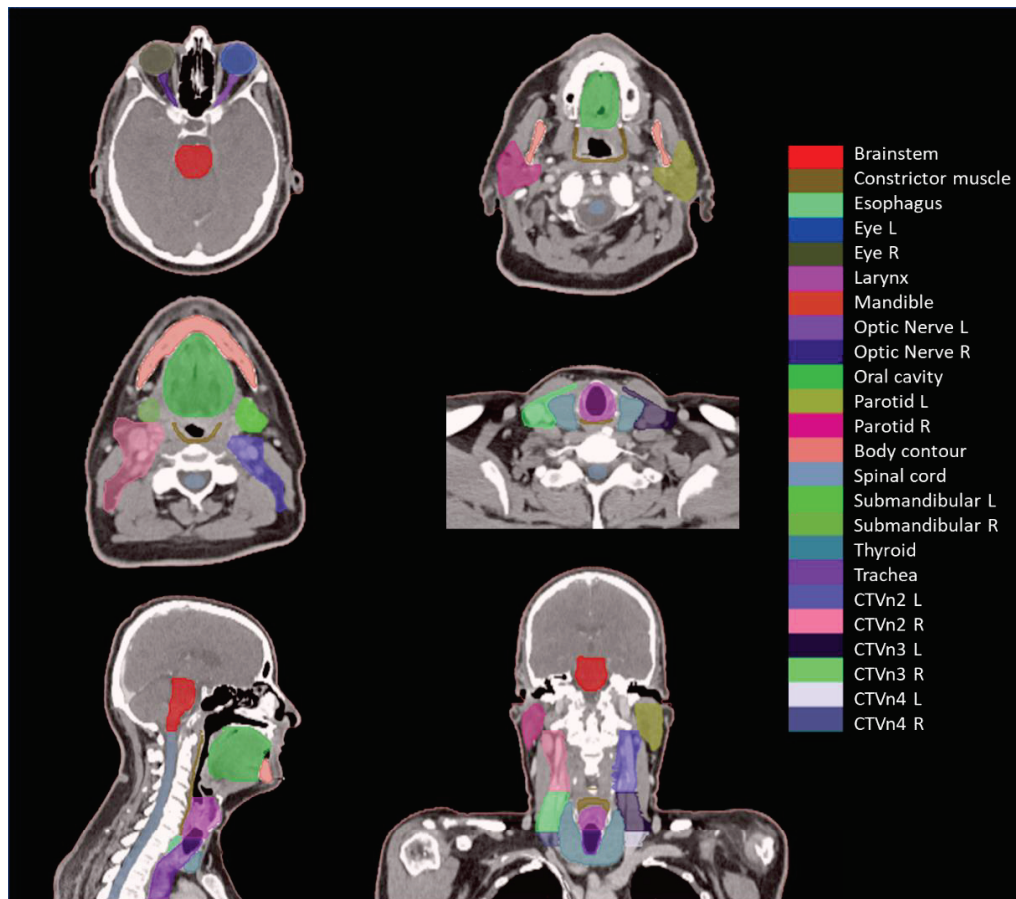


Figure 1.3-4 Organs-at-risk and lymph node level volumes typically contoured in head-and-neck cancer patients

3. Treatment planning

Treatment plan or patient dosimetry is performed by a medical physicist or a dosimetrist, on a treatment planning system (TPS) console using the patient CT image and the volumes defined in the previous step. Optimal treatment plan geometry and radiation field characteristics must be set, to deliver the prescribed dose to the PTV and minimal doses to the normal tissues. For the optimization of the HN treatment plans, dose-volume objectives and constraints (see Table 1.3-1) are imposed for each OARs and targets upon prescription, clinical protocols and recommendations [25,35].

Table 1.3-1 Risk of late toxicity for different organs-at-risk in HN and their correspondent dose-volume objectives and constraints used for treatment plan optimization (adapted from [25])

Region of interest	Toxicity	Dose-volume objectives and constraints
Planning target volumes (PTV)	\	$D_{50\%} = D_p$ (Gy) $D_{95\%} \geq 95\% D_p$ (Gy) $D_{98\%} \geq 90\% D_p$ (Gy) $D_{2\%} \leq 107\% D_p$ (Gy)
Brain	Symptomatic necrosis	$D_{2\%} < 60$ Gy
Brainstem	Necrosis or neuropathy	$D_{2\%} < 50$ Gy
Spinal Cord	Myelopathy	$D_{2\%} < 50$ Gy
Optic nerve/Chiasma	Optic neuropathy	$D_{2\%} < 55$ Gy
Cochlea	Hearing loss	$D_{mean} < 45$ Gy
Brachial plexus	Clinical neuropathy	$V_{70Gy} < 10\%$
Parotid glands	Xerostomia > grade 2	$D_{mean} < 25$ Gy
Submandibular glands	Xerostomia grade 4	$D_{mean} < 30$ Gy
Pharyngeal constrictor muscle	Dysphagia	$D_{mean} < 45$ Gy
Larynx	Hoarseness	$D_{mean} < 30$ Gy
Mandible	Osteoradionecrosis	$D_{5\%} < 70$ Gy

D_p : dose prescribed, D_x : dose that covers X% of the volume, V_x : volume receiving X% of the dose

4. Treatment delivery and QA

During a standard RT treatment, one fraction of the prescribed dose is delivered per day to the patient that lies on the treatment couch in the reference position defined during CT simulation. Prior to the treatment, a pre-treatment 2D or 3D image is acquired and registered to the planning CT image, to ensure the correct positioning of the patient. This procedure is called image-guided RT (IGRT) and is highly recommended for the HN patients repositioning to reduce systematic and random errors [24]. Most of the time, simple translation and rotation shifts are performed between the simulation CT image and the new (kV or MV) image of the day. Prior to each treatment session, a quality assurance (QA) of the TPS calculated dose distribution is conducted through phantom irradiations that enable the dose evaluation with regards to the actual output of the treatment machine.

In particular for HN cancer patients, accurate radiation delivery remains yet challenged by the patient anatomy alteration between the planning CT scan (on which the treatment plan is being created) and the scheduled days for the treatment fractions. A review of IGRT strategies for HN cancer treatments can be found in literature [36] that aims to provide clinics with best practice recommendations.

1.4. Image guided radiation therapy

IGRT is the process of regular imaging, during a course of RT, used to guide the position of the patient, by comparing the simulation CT images to the pre-treatment images, acquired in the treatment room prior to dose delivery. The main goal of IGRT is to reduce errors in patient set-up and positioning by correcting the alignment of different images of the same patient.

Image alignment, typically referred to as image registration, is the process of estimating a spatial transformation between two similar images according to different choices available for similarity measures (e.g. neighborhood correlation, mutual information, mean squares), optimizers (e.g. conjugate gradient line search, gradient descent) and transformation functions (e.g. rigid, affine, B-spline, dense deformation field). For the purpose of patient positioning, most commonly two images of a patient are aligned by simple rotation and translation shifts, process called rigid registration.

Imaging modalities are mainly split in two branches based on their dimensionality: 2D or planar imaging and 3D or volumetric imaging.

Planar imaging

Electronic portal imaging device (EPID) or portal imaging is the first 2D imaging modality integrated on the linear accelerators. It uses the megavoltage (MV) source and a detector underneath the treatment couch to capture the signal of the attenuated radiation field. It is often used for QA of the 3D-CRT radiation fields' shape and equally for the QA of the treatment machine. It can be also used for checking the patient positioning but its image quality in terms of tissue contrast is less attractive due to the high intensity radiation and comes additionally at expense of increased patient imaging doses.

Kilovoltage (kV) imaging allows better distinction of the bony anatomy due to the dominant contribution of the photoelectric effect at lower doses, and thus it is well suited for supporting daily patient positioning. Compared to the MV systems, delivering average dose per image as high as 20-70 mGy, kV systems are reducing the patient imaging dose to 0.1-0.3 mGy per image [37,38]. They imply however an additional kV source integrated in the treatment room, either as a separate system (e.g. ExacTrac X-ray system; Brainlab AG, Munich, Germany) or integrated on the LINAC (Figure 1.4-1). 2D-kV systems on the treatment machine are providing digitally reconstructed radiographs (DRRs), that acquired at 90° distance between each other (at typically 90° and 270°) are used to assess and correct patient translational shifts. Rotational errors can however be resolved only if the patient couch allows rotational movements.

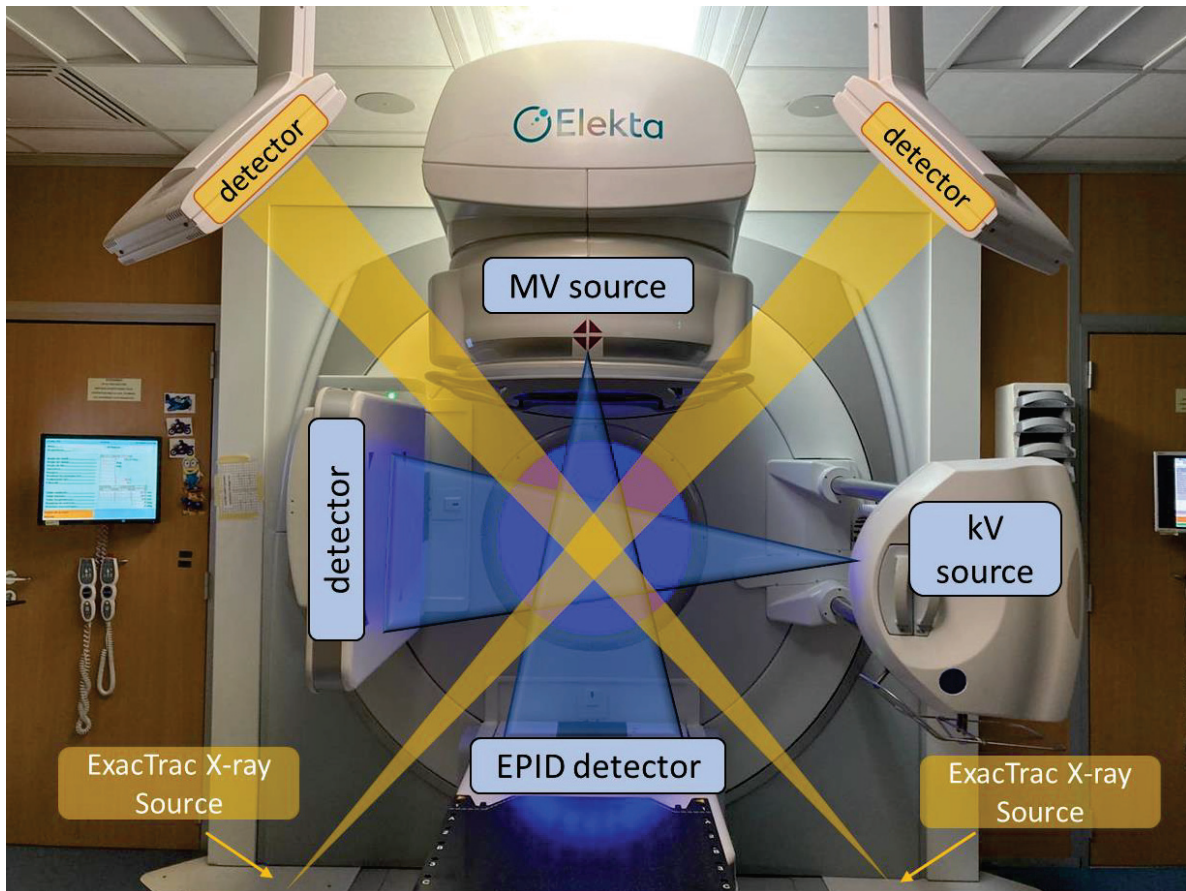


Figure 1.4-1 Planar imaging systems in a treatment room with an Elekta linear accelerator; MV and kV sources on the treatment head and additionally integrated ExacTrac X-ray system from Brainlab.

The ExacTrac X-ray system (Figure 1.4-1) has two separate radiation sources implanted in the floor and two detectors suspended in the ceiling that allows two paired-images to be acquired prior and during the treatment.

Volumetric imaging

The 3D imaging systems that can be found in a treatment room are: 3D-kV CBCT (kV-CBCT), 2D-kV imaging, high energy CT (2D or 3D MV-CT), CT on rail and MRI.

The 3D kV-CBCT system is the most used and became the standard imaging practice prior to RT treatment delivery (Figure 1.4-2). It is fixed to the treatment machine (at 90° of the treatment head), and uses a separate X-ray source with a correspondent flat panel amorphous silicon detector positioned perpendicular to the radiation treatment field. A single rotation of the tube delivering a conic shape beam is sufficient to achieve a reconstruction of the 3D image of the patient with adequate bone anatomy contrast to serve for patient positioning verifications, and subsequent translational and rotational corrections. Repeated scans may serve in monitoring intra-fractional motions and treatment response, and after adequate image processing, they may be used for dose recalculation and treatment plan adaptation. Moreover, the imaging dose is on average only 10-30mGy per kV-CBCT image [38,39], which is low compared to the treatment dose, scatter and leakage. Nevertheless, it is well justified because ensured an increased accuracy of patient positioning. Although it is very convenient due to its integration on most LINAC machines, the kV-CBCT system has several disadvantages: limited field of view (FOV),

limited soft tissue contrast resolution, presence of image artifacts and scatter in case of a bulky patient. An illustration of the CBCT system on an Elekta linear accelerator can be visualized in Figure 1.4-2 and a of a HN cancer patient position correction using kV-CBCT imaging is presented in Figure 1.4-3.

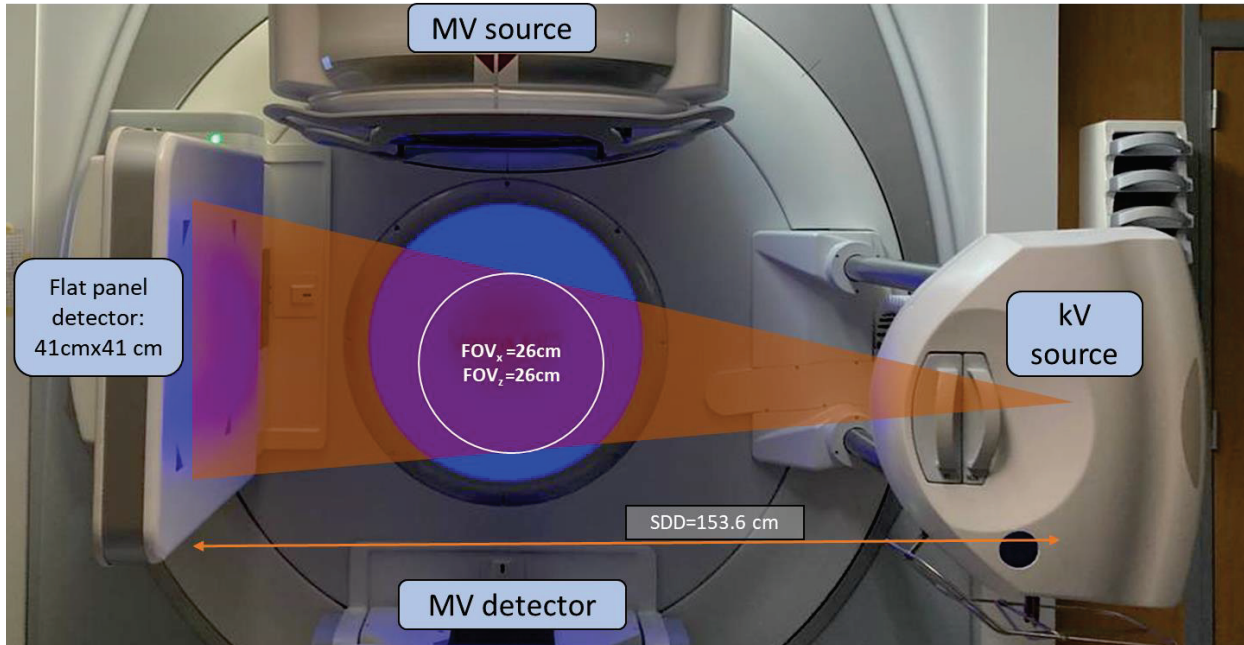


Figure 1.4-2 Elekta linear accelerator with kV and MV- CBCT imaging system; SDD = source to detector distance; FOV=Field of view, MV=megavoltage, kV=kilovoltage

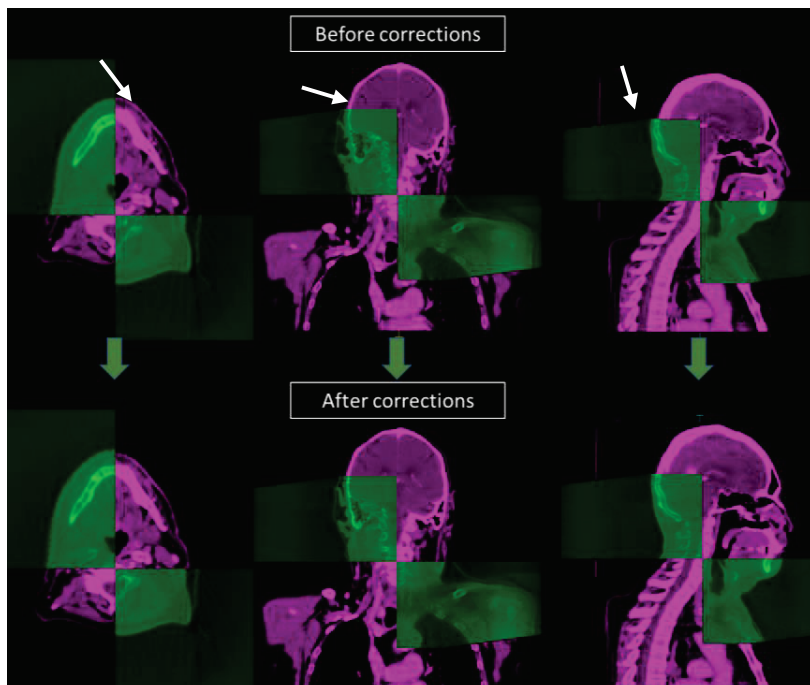


Figure 1.4-3 CT and CBCT rigid registration; With green is represented the kV-CBCT image of the day, and pink is the reference CT image.

High energy MV-CT imaging are currently integrated in TomoTherapy systems (Accuray, Sunnyvale, CA) where the same X-ray source is used to generate both the treatment beam (6 MV) and the imaging beam (3.5 MV). Essentially, it is a hybrid between a LINAC and a helical CT scan, where the CT component allows targeted regions to be visualized prior to, during, and immediately after each treatment [40](Figure 1.4-4). Imaging data is collected on a xenon detector located on the gantry, opposite to the radiation source. Varying with the pitch setting, the average imaging patient dose is typically 10-30mGy per scan [38,41].

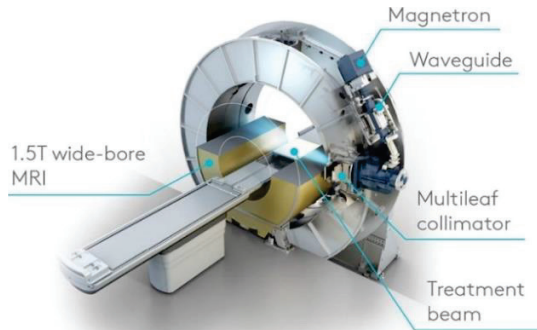


Figure 1.4-4 TomoHD (Accuray, Sunnyvale, CA)-TomoTherapy system

This treatment strategy is particularly employed for treatments that require high degree of radiation intensity modulation, patients having large lesions and those having prosthesis or dental artefacts where the MV-CT image quality is superior compared with kV-CBCT images. In addition, TomoTherapy is used for patients with recurring cancer, secondary cancers or metastases. HN cancers are also candidates for TomoTherapy treatments particularly in the event of a difficult localization of the tumor (next to a vital organ) that require extremely accurate patient positioning, or when maximum normal tissue dose tolerances have been reached after a previous administration radiation treatment. Compared with other IMRT treatments for HN patients, helical TomoTherapy (HT) treatments were shown to provide the most homogeneous target coverage with better sparing of the spinal cord, brainstem, the parotids and the swallowing apparatus [42,43].

Furthermore, non-invasive MRI can be used for image-guiding purposes, and it is known under the name of magnetic resonance-guided radiation therapy (MRgRT). This approach is found in the MR-LINAC systems that are exploiting the enhanced soft tissue contrast resolution provided by the non-ionizing radiation, without additional dose exposure to the patient. This ensure better target localization compared to standard kV or MV imaging and ultimately enables delivering of lower doses to the healthy tissues. These systems are however not often available as the equipment is rather expensive. MRgRT is particularly suitable for tumors in the brain or in the pelvic area where the soft tissue differentiation is highly important [44–47]. For the HN localization, MRI guidance is still at its infancy. Increased interest is however evolving in exploiting the new hybrid MR-LINAC platforms for this indication [48–51]. Two MR-LINAC systems are currently commercially available: MRIdian™ (ViewRay, Mountain View, CA, USA) with a low intensity magnetic field of 0.35 T and Unity™ (Elekta AB, Stockholm, Sweden) with a high-intensity magnetic field of 1.5 T [52,53] (Figure 1.4-5). The concept has the potential to become the next generation of RT standard by providing real-time visualization of the patient anatomy and real-time dose optimization [54].

A.



B.

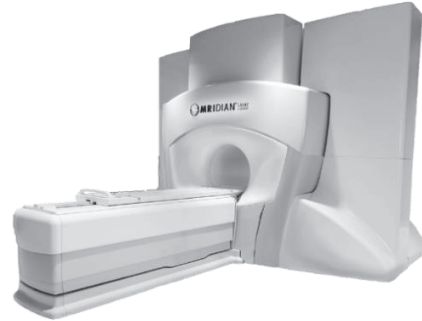


Figure 1.4-5 Commercially available MR-Linac platforms; A. Elekta-Unity™ and B. ViewRay - MRIdian™ system

Finally, by integrating sophisticated high-resolution real-time imaging equipment, IGRT is the foundation of high precision in radiation oncology and remains a major driving force for innovation enabling the shift towards a personalized treatment care. A summary of the IGRT techniques is presented in Table 1.4-1.

Table 1.4-1 Summary of in-room IGRT methods

Dimensionality	Imaging device	Advantages	Disadvantages
Planar imaging (2D)	MV (EPID)	<ul style="list-style-type: none"> • Same source as used for treatment (practical) • ↑Tissue contrast 	<ul style="list-style-type: none"> • ↑Imaging dose • ↓Soft-tissue contrast
	kV (DRR)	<ul style="list-style-type: none"> • ↑Bony anatomy contrast • ↓Imaging dose 	<ul style="list-style-type: none"> • Separate source
Volumetric imaging (3D)	kV-CBCT	<ul style="list-style-type: none"> • Translation and rotation • ↑Bony anatomy contrast • ↓Imaging dose • System accessibility 	<ul style="list-style-type: none"> • Separate kV source and detector • Artefacts for dense materials • Limited field of view • ↓soft-tissue contrast
	MV-CT (TomoTherapy)	<ul style="list-style-type: none"> • Same source as used for treatment • ↑Tissue contrast • Tumor localization • ↑OAR sparing 	<ul style="list-style-type: none"> • ↑Imaging dose • System accessibility • Image acquisition time
	MRI (MR-LINAC)	<ul style="list-style-type: none"> • No imaging dose • Translation and rotation • ↑Soft-tissue contrast • ↑Tumor localization 	<ul style="list-style-type: none"> • System accessibility • Image acquisition time • Costs

1.5. Adaptive radiation therapy

ART aims to undertake corrective measures, when necessary, based on daily tumor and normal tissue changes monitored by the in-room imaging techniques. By evaluating patient anatomy in daily images prior to each treatment fraction, a decision can be made whether the initial treatment plan requires adaptation. A major limitation of ART is the exhaustive labor and time needed to perform the plan adaptation. Therefore, currently, ART can be performed in two ways: online and offline (Figure 1.5-1), both requiring highly efficient workflows.

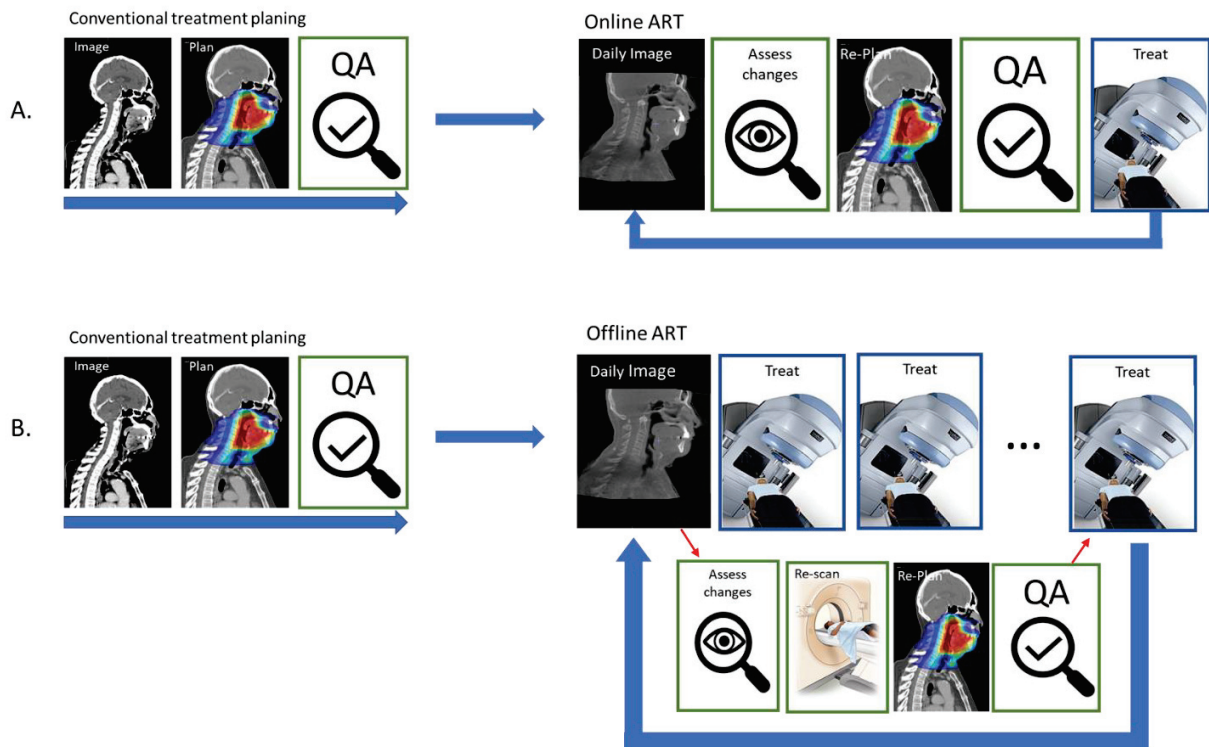


Figure 1.5-1 Adaptive radiation therapy strategies. (A) Workflow online-ART, (B) Workflow offline-ART

Online-ART uses IGRT to immediately adapt each fraction of the treatment to the daily changes in patient position and/or anatomy [55]. Although the concept is clear, the clinical implementation is hampered by several technical limitations of the elements involved: in-room imaging, uncertainty of the deformable image registration (DIR) algorithms, re-contouring time, plan re-optimization and plan QA. Moreover, during all the process of online adaptation, the patient must remain in the treatment position, in the treatment room. Thus, all the steps must be integrated seamlessly. Additionally, the extra time needed to perform online ART reduces the number of patients that can be treated in a day. Therefore, due to its complex and interdisciplinary teamwork demand, only few centers are practicing online ART.

Offline-ART uses the daily IGRT images to assess anatomical changes after each treatment fraction. When considered necessary, a new treatment plan will be performed based on a new simulation CT. The adapted plan will be then administered in the future fractions.

There is a particular interest in ART strategies for HN cancer patients due to the complex anatomy and tumor volumes in the proximity of OARs related to important physiological functions [56,57]. Moreover, there is a high potential for anatomical changes during the course of treatment in both tumor volume and surrounding tissue, that can have serious consequences in terms of underdosage of targets and

overdosage of the important OARs [58]. An average weight loss of 6-10% has been reported for HN patients throughout the course of a RT treatment causing inhomogeneous doses in the target volumes [59,60]. Reduction of target volumes as well as involved nodes is highly heterogeneous among the HN patients, and can cause important alterations of the normal anatomy [61,62]. Among studies, 3-16% reduction was found in the mean primary tumor volume after the first 10 fractions of the treatment, 7-48% after the first 20 fractions and 6-66% by the end of the treatment [56]. Parotid glands shrinkage was also reported in several studies indicating up to 48% volume reduction by the end of the treatment [56]. Additionally, their migration towards high dose gradient regions was shown to cause severe overdoses, up to 19Gy [63] which can have substantial impact in increasing the risk of xerostomia [64,65]. Figure 1.5-2 illustrates a patient case where volume changes can be observed from the baseline and the 24th fraction with respect to the primary tumor, the involved neck nodes and the parotid glands.

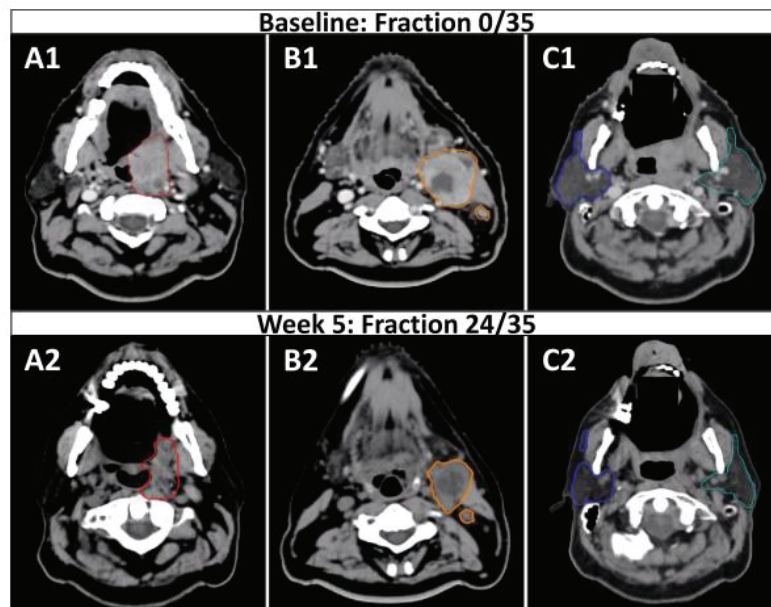


Figure 1.5-2 Primary tumor, nodal, and parotid volumes decrease over the course of radiation treatment of a 54-year-old patient with p16-positive T4 N1 M0 squamous cell carcinoma of the left tonsil. Patient received adaptive radiotherapy due to significant tumor response and weight loss observed through daily CBCT imaging. From baseline (A1) to week 5 (A2), the primary tumor decreased by 25%. The involved nodes decreased by 48.6% from baseline (B1) to week 5 (B2). The left parotid decreased by 37.2% (cyan) and the right parotid (blue) decreased by 41.9% from baseline (C1) to week 5 (C2) [50]

Spinal cord and brainstem are also important OARs that may be overdosed due to anatomical changes during the course of treatment. Although the dose variations are modest (2-4Gy) for most of the patients, significant dose escalation up to 15 and 10Gy was reported in few studies with regard to maximal doses in the spinal cord and brainstem, respectively [63,66,67].

Currently, there is no consensus on the timing regimen for ART administration, either for identifying the appropriate patient to benefit from it. Given the wide range of anatomical variations in both targets and OARs, a single ART regimen might not be applicable to all the patients. Several attempts have been made to identify baseline or dosimetric factors that can trigger the need of plan adaptation, among which one was able to assess a threshold for deviations larger than 3Gy in the mean doses to parotid glands [68]. A pilot study prospectively evaluated the need of ART based on weekly CT dose recalculations in 22 patients with oropharyngeal cancer. Poor target coverage or inadequate OARs sparing were considered triggers for ART administration, resulting in all 22 patients receiving at least 1 adapted re-plan and 8 of

them receiving 2 adapted re-plans. Similarly, in another study based on weekly CT re-planning, 8/10 patients were selected for ART when the PTV dose coverage was <96% or the spinal cord max dose >45Gy. Moreover, 41% of the adapted plans were triggered in the first 2 weeks of the treatment. More recent efforts are focused on the use of daily CBCT images to calculate accumulated daily received doses and identify the patients that can benefit from ART [69–71].

With respect to Heukelon et al. [72], Figure 1.5-3 illustrated a selection of possible methods of ART implementation for HN cancer patients with increasing frequency of re-planning. ART can be implemented at fixed-interval points (Figure 1.5-3 A) or triggered based on qualitatively and quantitatively assessment of daily imaging (Figure 1.5-3 B). A “serial” or “sequential” adaptation approach (Figure 1.5-3 C) involves more than weekly image acquisition where adaptation is performed using DIR between the planning CT and the image of the day. Ideally, a cascade ART scenario (Figure 1.5-3 D) is preferred where daily deformations are incorporated between the deformed images subsequent to all fractions.

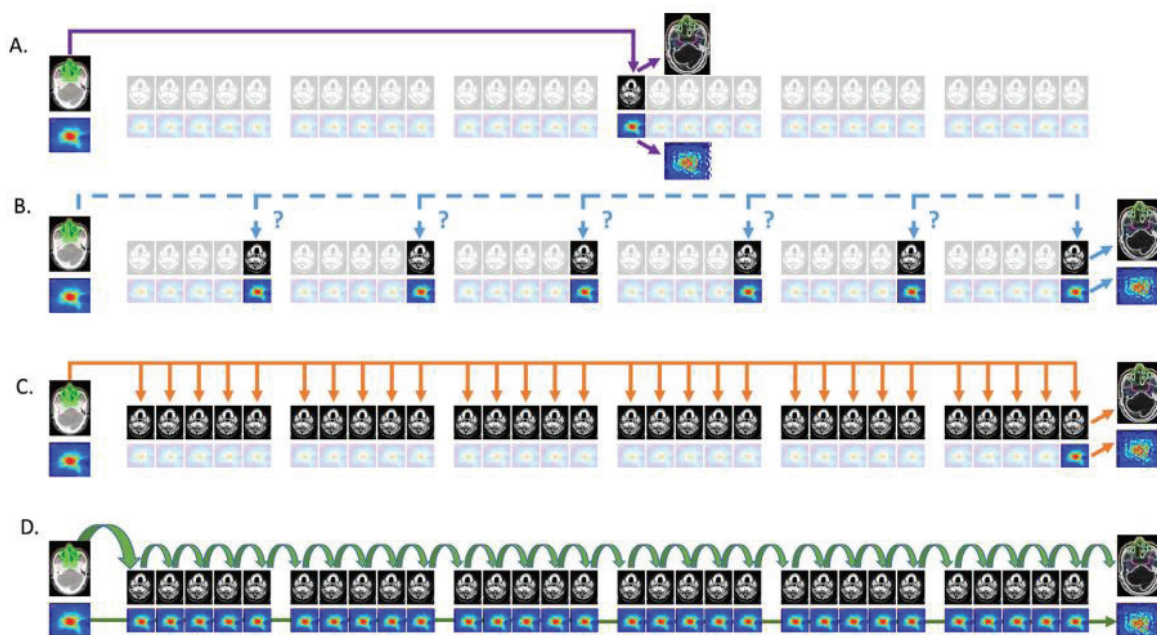


Figure 1.5-3 Possible typologies of ART implementation. A: fixed interval approach; B: ‘triggered’ ART; C: serial ART; D: cascade ART; ART= adaptive radiation therapy [68].

Generally, adaptive strategies may rely on repetitive CT, CBCT or MRI scans. An online plan adaptation based on daily CT proposed for HN cancers by the researchers at the Medical College of Wisconsin consisted of 2 step correction-scheme: segment aperture morphing (SAM) and segment weight optimization (SWO). During SAM the spatial relationship between the MLC and the contours of the targets and OARs are adjusted to the anatomy of the day using DIR and a new dose distribution is generated for each new aperture. Subsequently, optimal weights of the new segments are efficiently determined using a newly developed SWO package. After the daily CT scan, the procedure required 5-8min for the studied HN cases and was able to achieve the planned objectives by efficiently correcting the inter-fractional patient set up and anatomical changes [73]. Upon a retrospective study and clinical benefit assessment, the scheme was also adopted for the online adaptation of prostate and breast cancer patients at the clinic of Wisconsin [74,75].

Other interesting offline-ART methods are using the daily images from the first few days of the treatment to create composite target volumes and generate a library of adapted plans in function of tumor

volume and location. The adapted plan which fits the most the patient anatomy of the day, can be chosen for delivery in the future treatment sessions. This strategy is successfully used for bladder cancer patients to account for altered anatomy caused by the bladder filling. In one study, the CTV to PTV margins were safely reduced from 15 to 10mm and average small bowel volume spared was $31\pm 23\text{cm}^2$ when using the CBCT-assisted plan selection based on PTV of the day [76]. In another study, the CBCTs from the first 5 days of the treatment were used to create 3 adaptive plans which were used in 98.5% of the next treatment sessions [77].

Because most of the LINACs are equipped with CBCT imaging system for pretreatment positioning of the patients, it makes them very attractive for performing online-ART. However, the CBCT-based dose calculation is not trivial due to the “poor” image quality, limited size of the FOV and inconsistency of the HU numbers. There is an active area of research concerning solutions that enable to calculate doses directly on daily CBCT images. The proposed methods in the literature include: the use of CBCT adapted HU-ED calibration curve [78], density assignment methods [79,80], and DIR between the CT and the CBCT [81]. Ultimately, AI-based methods are expected to overcome this issue by automatic generation of synthetic-CT images that can provide accurate dose calculation based on the anatomy of the day [82]. Evaluation of four CBCT-based dose calculation methods for HN cancer patients is part of the last contribution of this manuscript. A commercially available system for performing online-ART based on daily CBCT images is Ethos™ (Varian Medical Systems, Palo Alto, CA, USA) which enables on-couch adaptive workflow within typically 15minutes per treatment session, which may vary among patients and tumor localizations [83]. Promising results on contouring accuracy, treatment plan quality and treatment time were demonstrated on prostate cancer patients [84]. Using the integrated automatic contouring algorithm, 11% of the contours required no changes, and 81% required only minor adjustments prior to dosimetry planning. The adapted treatment plans were chosen in 95% of the fractions, and demonstrated improved dosimetric outcomes compared to the initial plan, in 78% of the cases. Overall, the adaptation process was carried out in 19minutes on average. A clinical trial, called DARTBOARD has been recently initiated for evaluating the system’s clinical benefits of treating HN cancers with margin-less daily adapted radiotherapy (a 1 mm margin is kept for intrafraction motion) [85].

MRI-guided ART is also an evolving area of research thanks to improved soft-tissue contrast [86,87]. However, similar with CBCT images, dose calculations on MRI images require appropriate ED information that is usually obtained by DIR between the simulated CT and the daily MR images. The accuracy of DIR results is however uncertain, and sometimes manual density overrides and contour corrections may be needed. Methods for density assignment are also being used [88]. More advanced solutions are sought after for an accurate generation of synthetic-CT from MR images [89]. Commercially available solutions for online-ART are integrated in the Elekta-Unity™ and ViewRay-MRIdian™ systems [52,53]. Their dosimetric benefits have been recently demonstrated through a retrospective analysis on several anatomical sites (liver, lung, multiple abdominal lymph nodes, pancreas, and prostate) [90]. All groups showed a clear improvement of the PTV coverage with the adapted plans, and the largest reduction (-87%) in median dose to OARs was found on the pancreas patients [90]. However their clinical implementation involves complex workflows that can prolong the total treatment time up to 45minutes on average depending on the anatomical site [91].

In Table 1.5-1 is presented a summary of the current methods used to perform ART based on different imaging techniques with their associated strengths and limitations. Additionally, in Table 1.5-2 are summarized the current and upcoming trials on ART for HN cancers. Their aim is to assess the clinical benefit of ART when compared to traditional treatment regimens and at the same time, help identify which patients will benefit the most from this novel approach.

Table 1.5-1 Advantages and disadvantages of offline-ART and online-ART approaches

Method	Imaging Technique	Advantages	Disadvantages
Offline-ART	CT-based	<ul style="list-style-type: none"> • Good resolution • HU-ED relationship • Accurate dose calculations 	<ul style="list-style-type: none"> • High imaging dose • Limited soft-tissue differentiation compared to MRI, but better than CBCT images • Extra scanning time
	CBCT-based	<ul style="list-style-type: none"> • Availability of the system • Fast CBCT image acquisition • Fast registration CT/CBCT • Good contrast for bony anatomy 	<ul style="list-style-type: none"> • Limited soft-tissue contrast (↓ delineation precision) • Artefacts • Limited FOV of the CBCT • No HU-ED correlation • Imaging dose
	MRI-based	<ul style="list-style-type: none"> • Good soft-tissue contrast (↑ delineation precision) • Variety of sequences • Non-ionizing radiation • No imaging dose 	<ul style="list-style-type: none"> • image distortions • no HU-ED correlation • Functional imaging • Patients with contra-indications • Limited availability of the MRI scans • The costs
Online ART	CBCT-based (Ethos™)	<ul style="list-style-type: none"> • Automatic CT/CBCT registration • Automated segmentation of organs • Access to CT/PET or MR images on the console • Human decision making at each step 	<ul style="list-style-type: none"> • Limited soft-tissue contrast (↓ delineation precision) • Artefacts • Limited FOV of the CBCT • No HU-ED correlation • Imaging dose • Lower costs compared to MR-LINAC
	MR-LINAC (MRIdian™, Unity™)	<ul style="list-style-type: none"> • Good soft-tissue contrast (↑ delineation precision) • Variety of sequences • Non-ionizing radiation • Real-time tumor tracking • Daily adaptation 	<ul style="list-style-type: none"> • Image distortions • No HU-ED correlation • Patients with contra-indications • Increased treatment time • Limited availability of the systems • The costs

Table 1.5-2 Current clinical trials in adaptive radiotherapy of head and neck squamous cell carcinoma

Clinical trial	Location	Description	Eligibility criteria	Targeted nr. of patients (status)
A Prospective Non-Inferiority Trial of the Use of Adaptive Radiotherapy for Head and Neck Cancer Undergoing RT [92]	Memorial Sloan Kettering Cancer Center	Comparing LRSF (at 2 years) from patients receiving ART vs patients receiving IMRT without ART with the intent of assessing non-inferiority	SCC Receiving exclusive RT	64 (completed)
MRI - Guided Adaptive Radiotherapy for Reducing Xerostomia in Head and Neck Cancer (MARTHA-trial) [93]	University Hospital Zurich	MRI-guided IGRT adaptation protocol with the objective of evaluating xerostomia by salivary flow measurements at 6, 12 and 24 months.	Stages II-IVB receiving definitive or postoperative RT	44 (active, recruiting)
A Prospective Study of Daily Adaptive Radiotherapy to Better Organ-at-Risk Doses in Head and Neck Cancer (DARTBOARD) [85]	University of Texas Southwestern Medical Center, Dallas United States	Margin less daily adapted RT using Varian Ethos adaptive therapy software vs conventional RT margins strategy. Primary outcome: 1-year xerostomia	Stage I-IVB oropharyngeal, laryngeal, or hypopharyngeal SCC (exception T1-2 glottic carcinoma) receiving RT with/without chemotherapy	50 (active, recruiting)
PET-based Adaptive Radiotherapy Clinical Trial (PEARL) [94]	Velindre Cancer Center, Wales, United Kingdom	Prospective phase II feasibility study of biological dose adaptation using PET/CT at baseline and at 2 weeks. Primary outcome: LFS at 2 years	P-16 positive Oropharyngeal Cancer T1-3 N0 M0 treated with definitive CRT and non-smoker for >2 years	50 (Active, not recruiting)
Comparison of Adaptive Dose Painting by Numbers With Standard Radiotherapy for Head and Neck Cancer (C-ART-2) [95]	University Hospital Ghent, Belgium	Phase II randomized study comparing Adaptive Dose Escalation versus Standard Radiotherapy Primary outcome: LC at 1 year	T1-4, N0 oral cavity, larynx, oropharynx or hypopharynx SCC and T(any) N1-3 if glottic cancer, receiving definitive RT or CRT	100 (Active, not recruiting)
Adaptive, Image-guided, Intensity-modulated Radiotherapy for Head and Neck Cancer in the Reduced Volumes of Elective Neck [96]	University Hospital Ghent, Belgium	Phase II randomized study for patients receiving IMRT vs ART (2 re-scans) with the objective to reduce elective neck volume based on tumor response. Primary outcome: reduction of acute and late treatment-induced dysphagia	SCC oral cavity, oropharynx, hypopharynx and larynx. Stage T1-4, N0-3; and T any N1-3 if glottis	100 (completed, not yet published)
Adaptive Radiation Treatment for Head and Neck Cancer (ARTFORCE) [97][98]	The Netherlands Cancer Institute	Phase III trial randomizing participants to cisplatin or cetuximab and standard RT (70GY/35Fx) or ART (70-84GY/35Fx) with re-scans at week 2. Primary outcomes: 2-year grade 3+ toxicity, 2-year LRFs	stage T3-4, Nx M0 oropharynx, oral cavity or hypopharynx	268 (Active, not recruiting)
Trial of MRI-Guided Radiotherapy Dose Adaptation in Human Papilloma Virus Positive Oropharyngeal Cancer (MR-ADAPTOR) [99]	MD Anderson Cancer Center	Phase II trial using MRI imaging to reduce primary PTV volume. Stage 2 will randomize patient to standard IMRT od MRI-guided RT Primary outcome: LC at 6 months	P16 positive T1-2 N0-2b, lymph node < 3 cm receiving definitive RT	75 (active, recruiting)

Abbreviations: SCC = squamous cell carcinoma, LC = local control, LRSF = local-regional failure free survival;

ART may bring significant improvements by accounting for tumor and OARs changes during the course of treatment. However, as the initial contours based on the planning image set change, the initial planned dose may not accurately represent the actual delivered dose. At the moment no perfect solution exists for the issue of dose accumulation from the daily adapted treatment fractions. A voxel-by-voxel dose accumulation from each treatment session can be performed by deforming the dose based on the deformable vector fields (DVF) obtained from DIR, with the dose warped back to the initial planning CT image, to accumulate the doses from the delivered fractions [100]. An alternative approach is to deform the initial planning CT onto the daily patient image and calculate the “dose of the day”. Nonetheless, both methods for estimating the cumulative dose rely on the choice of the DIR algorithm and the quality of the underlying image of the patient. Moreover, the accuracy of the DVF may also be limited by internal target changes and OAR displacement. At the same time, the interpolation of doses constitutes another set of uncertainties depending on the method used (linear interpolation or energy/mass mapping). These daily deformable dose accumulation uncertainties have been subject of intensive research for the ART of HN cancers [101–108]. From [101] an illustration of the accumulated dose scheme can be seen in Figure 1.5-4 where weekly CT scans were used for plan adaptation. They quantified the performance of multiple DIR algorithms on 15 HN patients and found differences in the cumulated mean doses to the parotid gland from 1 – 8.9Gy. In conclusion, they emphasized that the choice of an image processing metric was at least as important as the choice of the registration algorithm.

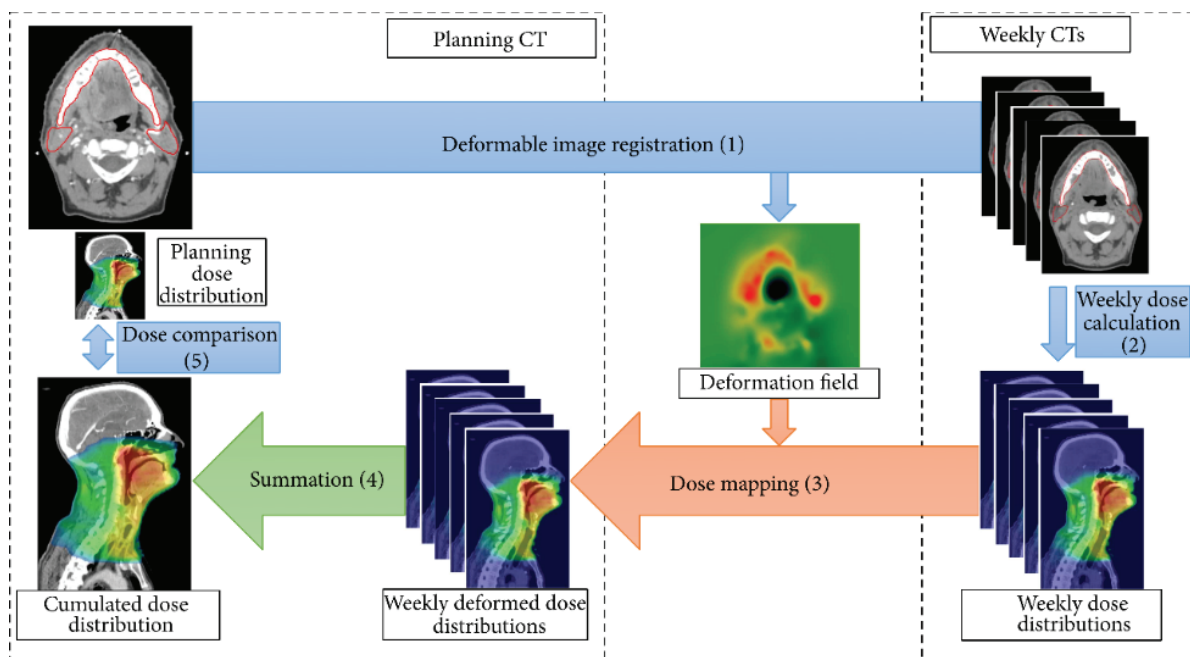


Figure 1.5-4 Dose accumulation workflow (from [101])

In the context of CBCT-based IGRT workflows, one study on 5 HN patients investigated the uncertainties associated with the choice of the DIR algorithm with respect to the accumulated doses [109]. Their maximum difference in estimated doses in OAR reached 2.8% of the prescribed doses and it was concluded that the choice of the DIR algorithm had a higher impact in the region of high dose gradient and/or when the CBCT image quality was poor. Another study demonstrated the accuracy of Varian’s SmartAdapt DIR algorithm on 12 HN patients with respect to target and OAR accumulated doses and presented a methodology that can be used for other DIR implementations [105]. As proposed in their study, a DIR algorithm can be evaluated by calculating the net voxel displacement after successive application of the forward and backward DVF. Similarly, the benchmark of a DIR algorithm can be performed with the help

of a virtual phantom, containing CT image set from the start and from the end of treatment, and the ground-truth DVF that links them together [110]. Furthermore, another commercial solution SureCalc from MIM Maestro software package based on Monte Carlo dose recalculations directly on CBCT images was compared with a dose mapping approach for 19 patients with HN cancers [108]. They showed that the dose discrepancies increased (>5%) when the calculated dose distributions were deformed back to the planning CT which resulted from workflow-related issues.

Taking all these uncertainties into account is part of the challenges when performing ART. This problematic of dose accumulation has not been part of this thesis objectives, however it constitutes future perspectives of the work.

Finally, current limitations to implementation of ART in routine clinical practice include:

- Optimal selection criteria of patients
- Optimal timing for ART
- Thresholds for OAR of increased toxicity probability
- Thresholds for tumor underdosage/overdosage
- Image quality of in-room imaging devices
- Accuracy of deformable image registration
- Time consuming task of re-contouring and re-planning
- Dose accumulation of adapted treatment fractions

To facilitate implementation of ART, AI solutions have emerged in RT for several applications (Figure 1.5-5) such as:

- automatic image segmentation (AS) or automatic contouring
- automated treatment planning (auto-planning)
- synthetic-CT image generation from CBCT and MRI images
- automated plan QA.

In the next chapters of this thesis manuscript, I will further discuss on auto-planning (Chapter 2), automatic contouring (Chapter 3 & 4) and synthetic-CT generation (Chapter 5) methods, with respect to the treatment of the HN cancer patients (Figure 1.5-5). Together with generalities on AI methods, the state-of-the-art for these different tasks will be introduced in the next sections of the clinical context.

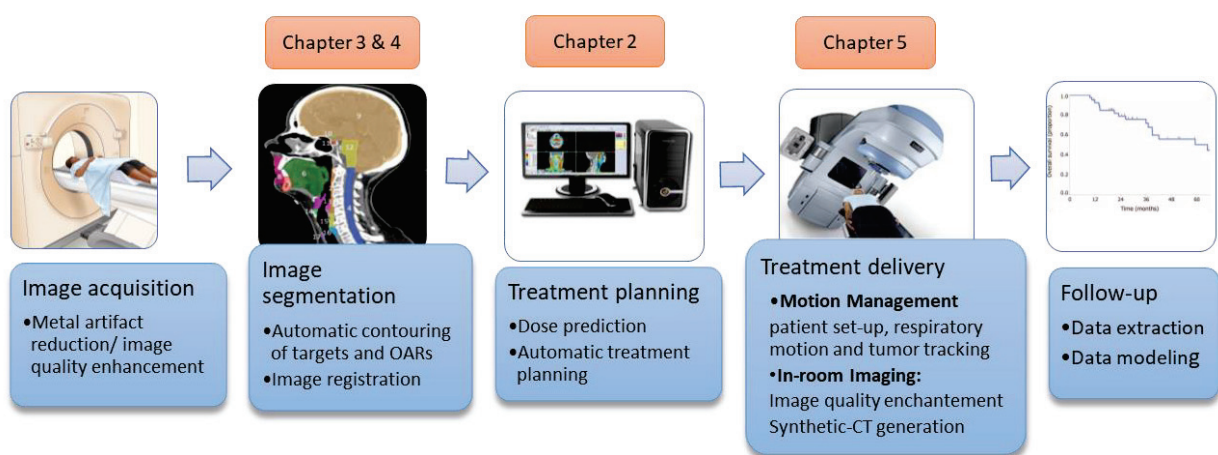


Figure 1.5-5 Potential applications of artificial intelligence methods in radiotherapy workflow

1.6. Artificial intelligence emergence in radiotherapy

AI is a universal term for all that implies modeling of intelligent human behavior by a computer [111]. Described as the science and production of intelligent technologies, AI was officially born in 1956 at a conference at Dartmouth College organized by Marvin Minsky, John McCarthy, Claude Shannon, and Nathaniel Rochester, who became the founding fathers of AI. Its physical branch deals with the manufacturing of robotics whereas the virtual branch includes informatics approaches. The former is called machine learning and represents a mathematical algorithm that learns from offered scenarios. We can envision that humans' process of learning is based on knowledge and experience, unfortunately both limited by the time factor. In a considerably shorter amount of time, a computer can handle a significantly larger amount of data and gain experience by using appropriate algorithms. The advantage today is that a massive amount of data is available and can be used for training algorithms on modern computational hardware. The AI subcategories are illustrated in Figure 1.6-1 and detailed later.

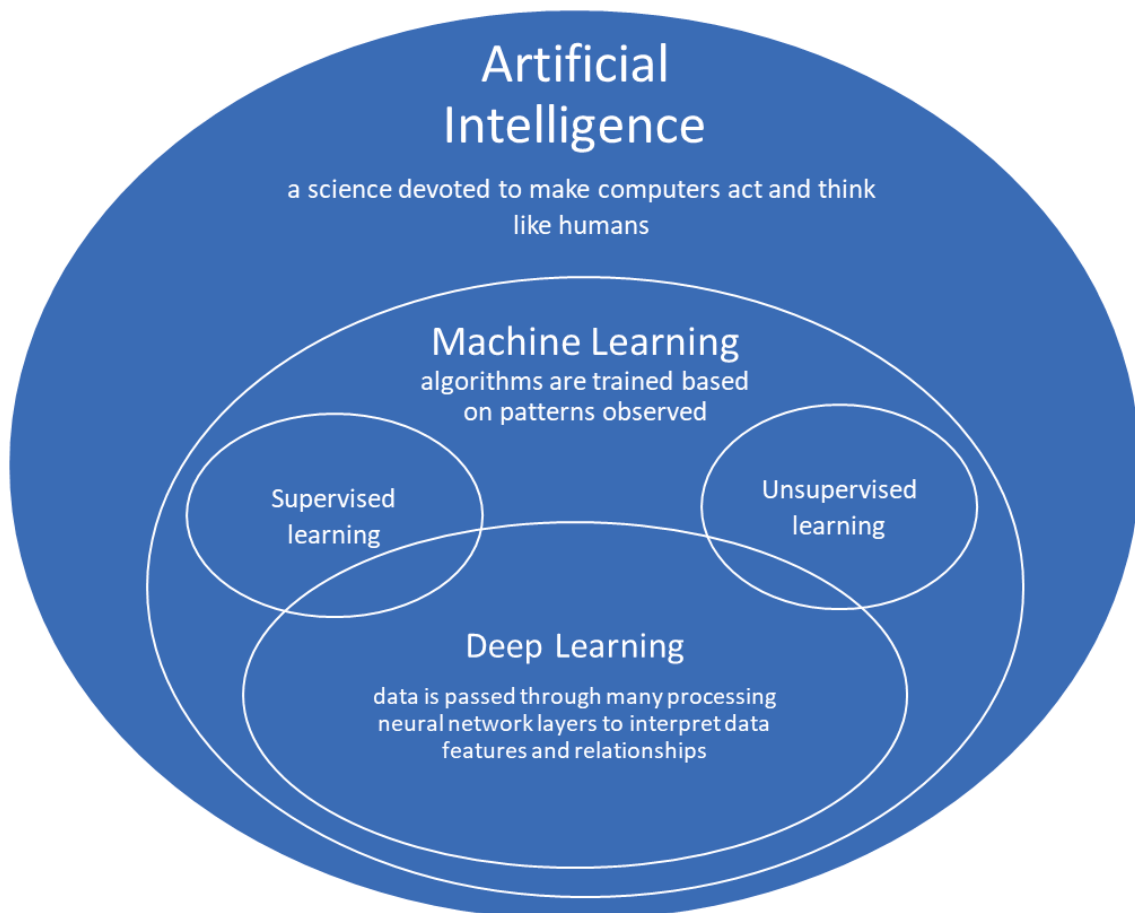


Figure 1.6-1 Subcategories of Artificial Intelligence (adapted from [112])

Machine learning (ML) is part of AI science and is the umbrella of all the other subcategories of virtual learning. In a broad sense, ML is the idea of a computer learning to perform tasks by studying examples grouped in a training set. Generally, it is divided into 2 main categories namely: supervised learning and unsupervised learning.

In supervised learning, the training database contains both data and its corresponding correct output, what it's called "labeled" or "annotated data". Output of new data is predicted by using a functional

relationship between input observations (cause) and output observations (effect). Various algorithms are developed to characterize this relationship, the most important being: regression algorithm, classification and reinforcement learning [113].

In unsupervised learning, the training database does not contain the correspondent output so that the computer must find alone the unknown relationship between cause and effect in order to predict outcome for future data. Based on this approach, numerous algorithms have been derived such as: dimensionality reduction algorithms, clustering, blind source separation, or density estimation [114]. Semi-supervised learning combines the two categories of learning by using a training set that contains both labeled data and unlabeled data (without the solutions) [115]. This allows reduction of labeled data that is not always available or sometimes expensive.

Deep learning (DL) is a subset of ML where training is based on artificial neural networks (ANNs), inspired by the structure of the neurons in the biological brain [116,117]. An ANN is composed of neurons organized into layers. We can distinguish 3 elements: input layer, hidden layers and output layer. The hidden layer neurons are the processing element which gathers all the inputs or signals from the previous layer of neurons with each input being multiplied by its associated weights on the connection. This result is then passed through an activation transfer function to give the final output signal to the next layer. See in Figure 1.6-2 a representation of an ANN and analog a biological neuron.

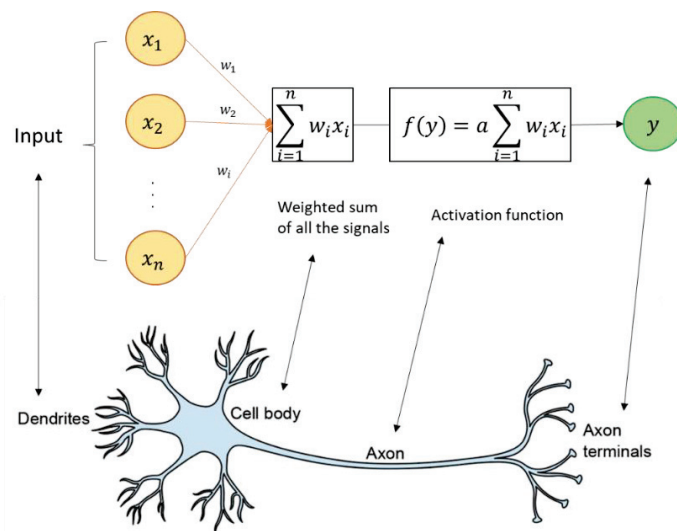


Figure 1.6-2 Representation of an artificial neural network vs a biological neuron (adapted from [118])

The advantage of DL is that during the training phase, the system adjusts its internal parameters based on an error computed between the observed and desired output. Along with the computed weights to reduce the error, the system calculates a gradient vector for each weight that indicates the error deviation triggered by the weight adjustment. A general method to find an effective set of weights is called stochastic gradient descent (SGD) method. The system takes the input vector of a few examples and computes its output, error and average gradient and the whole process is repeated multiple times on small sets of training data until the average of the object function stops decreasing [117].

Deep neural networks (DNNs) are combinations of multi-layered ANNs where the output of the first layer becomes the input for the next layer and the process repeats until the last layer that gives the output of the whole system (Figure 1.6-3). Moreover, because each neuron from one layer is connected to all neurons from the next layer, DNNs are in general (but not necessarily) fully connected neural networks (FCNN). They are dependent on the input data and learn its hierarchical representation without requiring additional feature extractors. Moreover, just like basic ML algorithms, they can be either supervised or

unsupervised. The main disadvantage is that they need large quantities of input data in order to be effective.

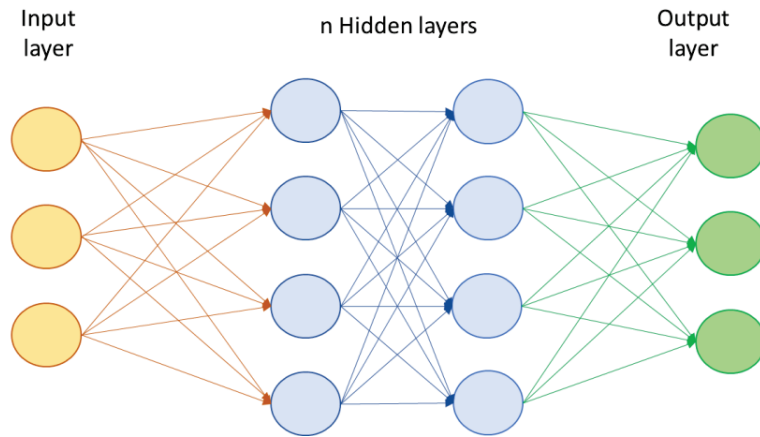


Figure 1.6-3 Deep neural network

Convolutional neural networks (CNNs) hold a different architectural scheme that takes advantage of spatially structured information, making them suitable for processing (medical) images. A typical architecture consists of convolution layers, pooling layers and fully connected layers, whose purpose is to find first simple representative features of the input data and progressively search for more elaborate features as the layers succeed each other. From an image that is considered a matrix of pixels, the neurons of the convolutional layer are dividing the image into small blocks and extract patterns/features and form, so-called, feature maps. A feature map is a collection of multiple neurons each holding the location of a particular feature in an image. The pooling layer performs a down sampling operation that aims to reduce the dimensionality of the feature map. The pooled feature maps are eventually converted into a single long continuous linear vector that will be the input layer of a fully connected layer used to end the classification task. See in Figure 1.6-4 an example of a CNN solving an image classification task.

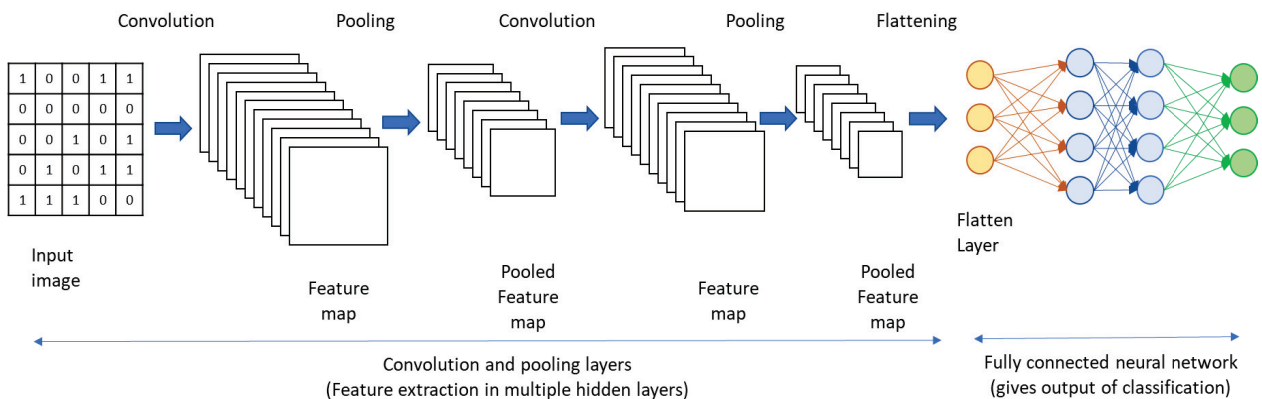


Figure 1.6-4 Convolutional neural network in solving an image classification task (adapted from [119])

An evolutionary history of the CNNs architectural developments as well as performance comparison of recent architectures can be found in the survey from Artificial Intelligence Review Journal done by Khan et al. [120]. In 2012, a remarkable performance was brought by AlexNet [121] increasing the depth of the

network to 8 connected layers and several optimization parameters, reaching unprecedented performances on classical image categorization problems. Later, in 2015 the concept of skipped connections and residual learning gained popularity with the introduction of ResNet which proposed a CNN of 152 deep layers [122]. The performance of CNNs has significantly improved over the years by exploiting depths and structural modifications. Nowadays the ambition lies in the development of new effective architectural designs that include blocks as auxiliary learners.

Three branches where AI is empowered in RT will be further elaborated in this thesis manuscript (Figure 1.5-5) that can help implementation of ART for HN cancer patients. The next 3 sections will introduce the state-of-the-art with regard to each of the thesis contributions.

1.7. Medical image automatic segmentation

By definition, medical image segmentation is the process of dividing an image into multiple areas according to different landscapes in order to label anatomical structures. It is a difficult and time-consuming task but it is a prerequisite for a successful treatment plan. Complex cases like HN patients are the focus of this thesis manuscript and they are particularly challenging and tedious due to the localization of tumors in close proximity of multiple OARs. Routinely contouring is performed manually by a physician or a dosimetrist and can take up to 1-2h or even 3h per patient depending on the case and the experience of the delineator [123–125]. Although guidelines exist [26,27] important variations between contouring practices are still observed [126–130]. Moreover, the high intra- and inter-observer variation exhibited by manual delineation proved to have important dosimetric impact for the patients [61,131]. In the new area of ART, one of the barriers to overcome is the time spent on delineating the volumes on which dose constraints are imposed. Therefore, at the basis of high-quality treatment plans stands the accurate volume definition. In this section, we propose an extensive literature review on automated image segmentation methods of organs principally focused on HN localization.

The goal of automatic segmentation (AS) or automatic contouring is to harmonize contouring practices by improving consistency in contour definition and to increase efficiency by reducing the manual delineation time. Ultimately, the goal is to facilitate implementation of ART.

The continuous technical expansion shows promising results in the development of different automated methods [132–134]. Each organ however exhibits particular limitations (such as appropriate image contrast and well defined anatomical boundaries) and thus finding a AS solution to perform well over a big range of anatomical structures is a continuous challenge. At the same time, quantifying the quality of a segmentation method has also been another persistent challenge, and up to now there is no consensus on how to assess performance of AS solutions. In general, results are reported in terms of geometric accuracy, assessment of manual time reduction, clinical acceptability and dosimetric impact. Recommendations on what metrics can be used exist, but no standardized protocol yet to account for the strengths and shortcomings of each [135,136].

For the geometrical accuracy evaluation, a measure that characterizes the volume overlap between two considered contours is recommended in combination with a measurement of positional displacement which provides complementary information. One of the most commonly used coefficients is dice similarity coefficient (DICE). DICE measures the volumetric overlap between the ground truth contour (A) and the predicted segmentation (B), leading to a value between 0 (no overlap) and 1 (perfect overlap). It is defined as:

$$DICE = \frac{2 \times |A \cap B|}{|A| + |B|}$$

With the intention of setting a benchmark for AS, DICE values were investigated and it was concluded that in the validation process of AS method, a value of “0.80” for the DICE can be interpreted as a good performance and looking at the value of 1 as performance benchmark is not realistic [133]. At the same time, what seems to be a good volumetric overlap can still conclude in significant organ overdose [137].

While DICE is limited to the volume intersection without considering the shape differences, a second most used metric that accounts for the magnitude of contour displacement is the Hausdorff distance (HD). HD is a boundary-based metric that measures the surface distances between the two contours. Due to its sensitivity to outliers, often the 95-percentile HD ($HD_{95\%}$) is chosen in contour evaluation:

$$HD_{max} = \max[d(A, B), d(B, A)]$$

$$HD_{95\%} = \max_{k_{95\%}}[d(A, B), d(B, A)]$$

$$d(A, B) = \max_{a \in A} \min_{b \in B} \|a - b\|$$

where $d(A, B)$ is the directed HD and A and B are the set of non-zero pixels in the images. Similarly, HD metric has its own limitation that it does not focus on the object itself, therefore does not punish a prediction with a large hole inside or with a spotted pattern within the contour [136].

Another important aspect to consider when evaluating an AS solution is the manual ground truth segmentation that has a considerable impact on the interpretation of the results. The choice of a gold standard or reference contour varies in the literature from a mathematical average contour, a radiologist-defined contour, an experienced oncologist defined contour or a consensus contour that is decided upon by a panel of experts [138]. The quality of the manual segmentation is hampered on one hand by the image quality and structure visibility on the medical images and on the other hand by the IOV as cause of individual eye sensitivity of the physician, eye-hand motoric fatigue, experience as well as the contouring tools available. Although physical and digital phantoms can be constructed for which ground truth is known or easily estimated, such phantoms cannot fully reproduce the whole range of imaging characteristics and patient anatomical variability observed in clinical data. Poor overlap results do not necessarily mean that the AS solution is worse than the manual reference, but that it differs in terms of spatial localization and geometric similarity. For this reason, a blind evaluation of the contours from several clinicians can help make the decision upon the clinical acceptability of an automatically generated contour. Alternatively, publicly available data sets such as The Cancer Imaging Archive (TCIA) [139] or Public Domain Database for Computational Anatomy (PDDCA) dataset released as part of the 2015 segmentation challenge [140], can be used to evaluate and prove generalizability of a segmentation method.

Furthermore, a highly recommended aspect to address when evaluating an AS solution concerns the dosimetric consequences of using AS contours. This implies reference dose distributions to be compared with those generated from AS contours. The task involves additionally exhaustive labor and is not systematically conducted in studies, first because of the time requested to be completed, and secondly because of the intra-planner factor, which could introduce bias in the observations [141]. To address these difficulties, research groups have adopted different strategies on performing the dosimetry. Among other methods, some authors proposed to superpose the original clinical plan onto the automatically delineated contours [142,143]. Others have employed the use of auto-planning such as knowledge-based planning [131,144] or conserved the original beam configuration parameters [145]. Finally, most of the dosimetric studies results concluded non-significant differences in the delivered doses and inconsistent correlations between geometric indices and dosimetric endpoints. One study investigated regions of higher sensitivity around the PTV where contour accuracy is particularly important [144]. Most of the times, organs located in short vicinity of the PTV required manual corrections, however the correlation between organ distance to targets and dosimetric impact was not always confirmed. Additionally, it is difficult to isolate the effect from each individual contour error (inaccurate contour borders) as the dosimetric effect is rather a cumulative one. For this, ideally would be to create separate plans per investigated contour.

Eventually, an inaccurate delineation provided by a AS method will result in additional time for manual adjustment, therefore reporting of manual post editing time is an important aspect to help conclude about an AS solution. Remaining conscious that a machine cannot fully replace human judgement, the challenge

of AS is to get as close to clinical acceptability as possible and to require minimal time for manual corrections. A summary of evaluation methods for AS solutions with their associated strengths and limitations can be followed in Table 1.7-1.

Table 1.7-1 Methods for evaluating automatically segmented contours

Method	Evaluation	Advantage	Disadvantage
Overlap metrics	Volume overlap Surface overlap (volume DICE, Surface DICE)	Easy to calculate Practical during training/validation	Volume dependent No information on contour shape and localization Not correlated to subjective clinical acceptance
Distance metrics	Distance between surface point of the true and predicted volume (e.g. HD, MSD, APL)	Focus on the boundary of the contours	Not dependent on absolute volume Difficult to interpret for small contours Outliers
Volume	Comparison of absolute volume	Easy to calculate and interpret (Systematic over/under estimation of a contour)	No correlation to contour location
Dose calculations	Dosimetric impact of delineation uncertainty	Clinical impact of differences between delineations	Labor intensive Dependent on the planning strategy Subjective to planner input
Clinical assessment	Turning test Blinded evaluation	Rating from multiple observers	Labor intensive Subjective to observer's experience
Manual corrections	Time record IOV assessment	Assessment of benefit in the clinical workflow	Labor intensive Subjective to observer's experience

Abbreviations: HD= Hausdorff Distance, MSD= Mean surface distance, APL= Added path length, IOV=inter-observer variation

Finally, all the resources involved in developing an AS solution as well as the computational time of generating the desired contours are important factors to consider when evaluating AS solutions.

Traditional segmentation methods are still at the foundation of the newly complex methods and their ideas are worth credit [146–148]. The origins of computer-aided segmentation were set by the successful implementation of digital image processing and mathematical techniques such as threshold-based method [149], region-based [150] and edge detection [151]. The logic is simple and calculations are fast but they lack precision in terms of details. At present, methods based on more complex concepts and empowered by AI research have made remarkable achievements and the segmentation accuracy has surpassed the traditional methods. The state-of-the-art of medical images AS methods fall into two main categories: atlas-based (ABAS) and deep learning (DL) solutions (Figure 1.7-1). Hybrid methods are also proposed under the atlas-based methodology.

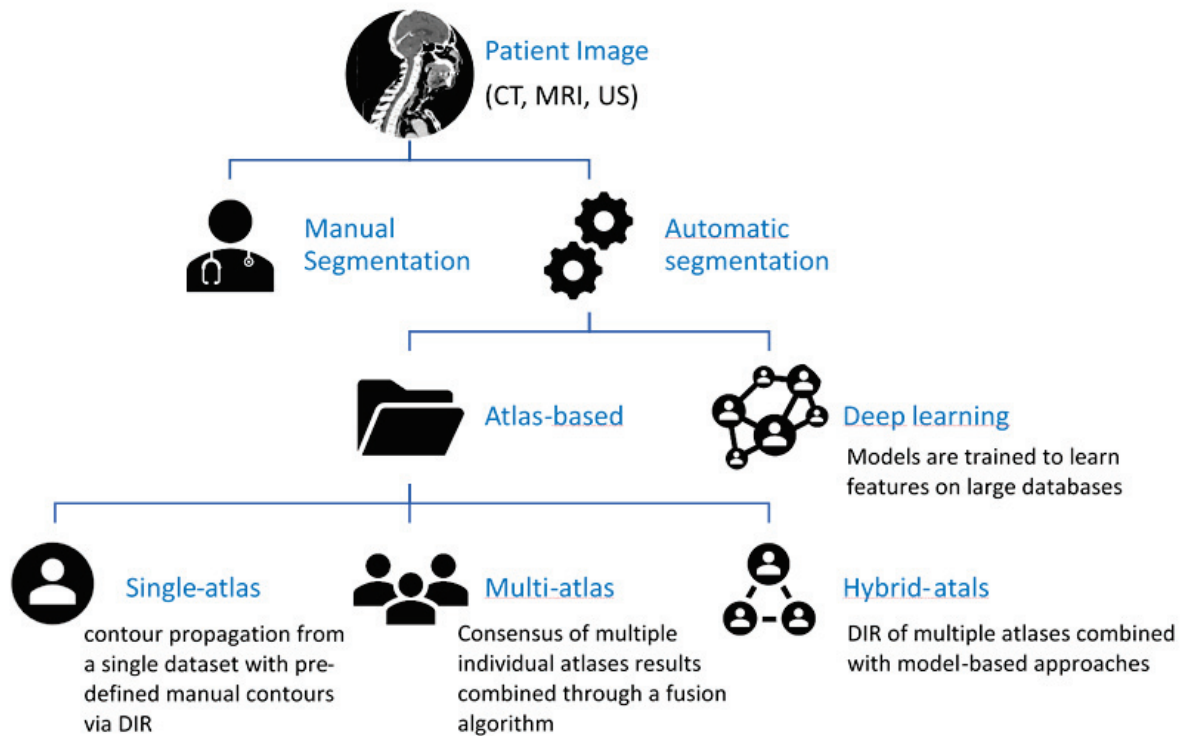


Figure 1.7-1 Medical image segmentation methods; CT=computed tomography, MRI=magnetic resonance imaging, US=ultrasound

Atlas-based methods

ABAS is a segmentation technique that uses previously annotated data (i.e atlases) to propagate organ contours onto a new patient image via deformable image registration. As atlas-based method considers only contour information from a defined atlas collection or library, the quality of the segmentation is highly dependent on the similarity of the atlas and the underlying patient. If only one atlas is used, it has to be thoughtfully chosen in order to embody an average patient anatomy. Secondly, the accuracy of atlas-based solutions depends on the accuracy of the DIR algorithm that can be hampered by large anatomical deformations (particularly in the HN region).

Single atlas-based techniques make use of a single pre-defined dataset of gold standard volumes, whereas multi-ABAS use combined information from multiple atlases to reduce the variability in anatomy between the atlas and the new patient. The multiple individual segmentations generated from each atlas can be combined to form a population-based average atlas, through a process called fusion. Those methods are widely spread due to their convenient implementation that require minimum of resources. However, they do have several drawbacks: atlas selection strategy (single vs multi-atlas)[152]; performance plateau reached after 10-20 atlases [153]; poor performance for small and low contrast soft tissue structures [154]; increased computational time with each added atlas [155]. An illustration of single-atlas based and multiple-atlases based segmentation can be followed in Figure 1.7-2.

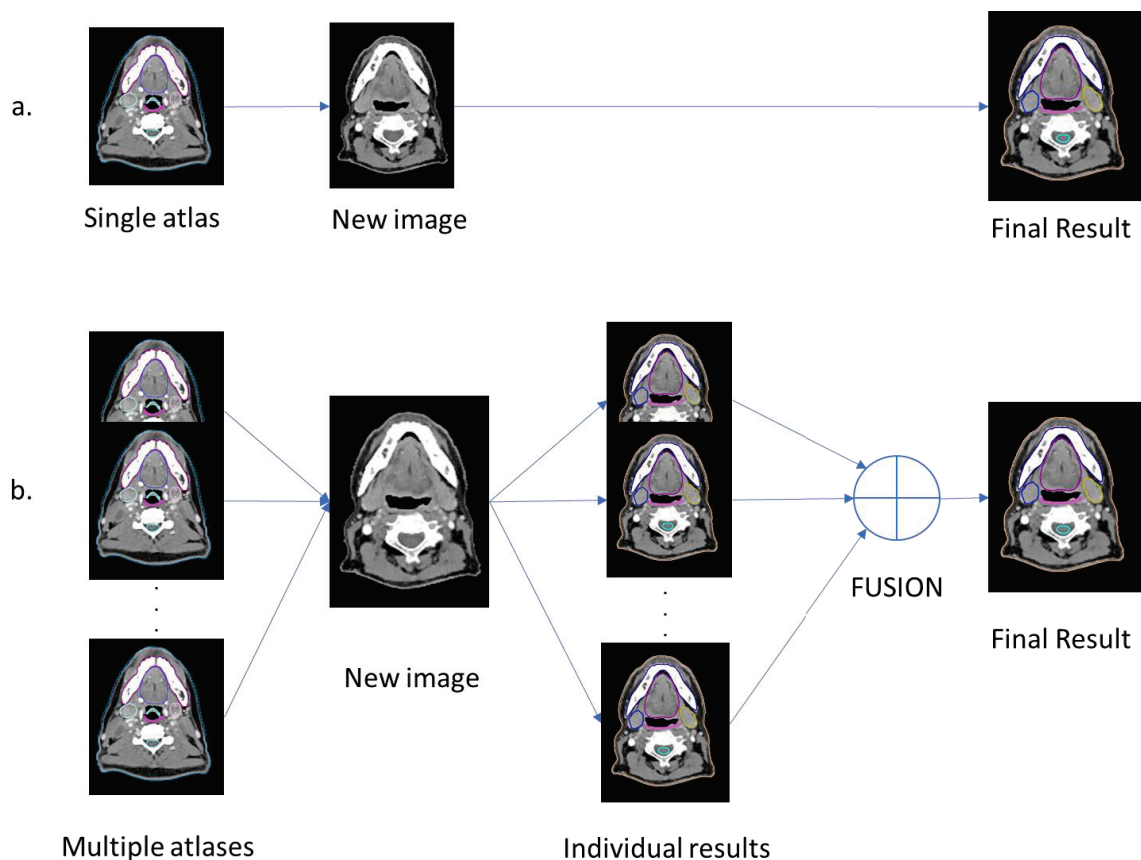


Figure 1.7-2 Atlas-based segmentation a. Single atlas; b. multiple atlases segmentation

Among the atlas fusion algorithms, one of the firsts, introduced in 2004 by Warfield et al. is the simultaneous truth and performance estimation (STAPLE) [156]. STAPLE algorithm consists in estimating the optimal combination of the segmentations by weighting each segmentation upon the estimated performance level based on expectation maximization algorithm. Several studies have used the STAPLE algorithm to compare single and multi-ABAS techniques for delineation of OARs and lymph nodes levels, and overall multi-ABAS performed better than single ABAS methods [157–159]. At the same time, STAPLE algorithm has been widely used to define the reference ground truth segmentation from multiple manual delineation to overcome the IOV. Stapleford et al. followed this approach for generating a consensus between five physicians’ manual contours and used it as a benchmark in evaluating bilateral nodal CTV segmentation in HN cases [160]. They also measured a reduction of IOV when STAPLE was used.

Another algorithm is similarity and truth estimation for propagated segmentation (STEPS) [161], that in addition to STAPLE includes a spatially variant image discriminator that discards the atlases that have the least anatomical similarity to the underlying patient. A comparison between STAPLE and STEPS was performed by Hoang et al. for the segmentation of OARs in HN patients [161]. They showed that STEPS algorithm outperformed STAPLE for a number of structures such as parotids, brainstem and spinal cord but not for smaller OARs such as optic chiasm and ocular globes.

Majority voting (MV) fusion is another algorithm used for atlas fusion [162,163] where for each voxel in the new image, each atlas attributes a vote denoting what structure or background the voxel belongs to. The final label of the voxel is then decided by the one label that has the most votes. Lee et al. [162] evaluated a commercial ABAS software (with STAPLE label fusion) vs MV fusion, and stated that their performances were similar and further research was need to investigate their differences. Different atlas-

based methods were compared by varying the atlas selection criteria and the fusion algorithm by Chen et al. [164]. It was demonstrated that a weighted combination of atlas individual segmentation results based on correlation coefficient (as measure of similarity between the atlases and the new patient), outperformed standard combination methods such as MV and STAPLE.

Another method to combine results from multiple atlases is patch-based segmentation [165]. Unlike STAPLE, patch fusion (PF) algorithm takes advantage of the image intensity information when weighting the individual atlases results. PF algorithm has been compared against STAPLE, within Advanced Medical Imaging Registration Engine (ADMIRE, Elekta, AB Stockholm, Sweden) [166] for the segmentation of 7 HN OARs [167]. Using 10 atlases and leave-one-out strategy for atlas selection, PF results were superior to STAPLE.

Similarly, using ADMIRE software, Liu et al. performed an evaluation of commercial ABAS (Elekta AB, Stockholm, Sweden, with STAPLE fusion) and intra-patient deformable contour propagation for offline-ART in HN patients. Three separate image datasets were used for each patient: pretreatment planning CT, in-treatment planning CT, and a CBCT, the last two acquired in the same day. For the 7 OARs, initial contours generated by STAPLE had good conformity to manual reference (DICE>0.8 and mean surface distance <2mm). Moreover, similar accuracy was obtained between CT-to-CBCT and CT-to-CT deformable registration therefore demonstrating the feasibility of the method for adaptive workflow [168].

The performance of three commercially available solutions (ABAS 2.0, Elekta AB, Stockholm, Sweden; MIM 5.1.1, MIM Vista Corp, Cleveland, Ohio; VelocityAI 2.6.2, Velocity Medical Systems, Atlanta, Georgia) was evaluated by La Macchia et al. for the accuracy of deformable registration and contour propagation when using single atlas strategy. The results revealed modest absolute differences between systems and significant time savings compared to manual contouring regardless of the solution [154]. Although manual corrections of the deformed contours were needed, the use of the atlas-based methods reduced the delineation time with 1h for HN patients, 40min for prostate and 20min for mesothelioma patients.

Furthermore, in addition to the image registration and fusion algorithms, to be noted that the quality of the atlases themselves impacts the resulting quality of the AS contours. For this, strict implementation of the international guidelines is the key, as well as the involvement of multiple experts. Although additional manual review and corrections are needed, simple atlas-based solutions can reduce the manual delineation time with minimum resources required [125].

Hybrid atlas-based methods

In addition to the multi-ABAS strategy that relies only on structures shape variation from the contours in the atlas library, the hybrid segmentation (HAS) approaches are aiming to compensate for the lack of reliable image information by imposing prior shape constraints in the segmentation process. These constraints are constructed by the image features learned from the contours' variation within the atlas library. Several implementations of HAS methodologies are present in the literature.

A method investigated by Qazi et al. [169] uses a probabilistic mask to guide the DIR by a boundary refinement approach. The segmentation starts with a single-ABAS registration and then it is refined down to voxel level classification. This combination of local low-level features and global high-level prior shape information has greater potential in achieving more reliable and robust AS contours. The technique could potentially be improved by integrating atlas selection or combination of multiple atlases using a fusion algorithm in the first step of atlas registration. Walker et al. studied a smart probabilistic image contouring engine (SPICE) algorithm, which performs an initial registration, followed by a deformable registration, and finally a probabilistic (model-based) refinement [170]. For most contours in HN, statistically significant differences compared to the reference were observed but when the contours were manually corrected, no significant differences remained. In conclusion, they affirmed that the human oversight remains critical. At the same time, when Thomson et al. evaluated the SPICE algorithm, the results after manual corrections

of SPICE contours did not improve significantly and no time-saving evidence was confirmed [171]. Fritscher et al. examined a different HAS approach, by combining multi-ABAS with geodesic active contours (GAC) and statistical appearance models (SAM) for segmentation of the parotid glands and brainstem [172]. This way, they tried to combine the strengths of each single method to overcome the other one's limitations. In their approach, starting with the robust atlas-based method, boundary features were learned by the GAC and prior information about anatomically plausible appearances of structure was added by the SAM. They observed statistically significant improvement for the model-based approach when compared with multi-ABAS technique alone. Fortunati et al. proposed a similar framework of combined shape prior from atlas-based registration with intensity modeling [173]. They also found improvements in accuracy of most of the investigated tissues when compared to a typical atlas-based segmentation based on MV fusion.

Furthermore, training a voxel classifier in parallel with atlas registration has also been investigated by Han et al. [174]. They used Random Forest (RF) algorithm which is a supervised learning algorithm designed for voxel-wise classification using both local and contextual image features. Particularly structure border regions benefit from training of RF classifiers where atlas-based segmentation errors typically occur. A framework of an atlas-based segmentation algorithm using a prior shape model or a voxel classifier alongside a traditional atlas-based approach can be followed in Figure 1.7-3. More explicit, classifiers trained using the atlas library as training data, were applied to re-estimate a structure contour at the level of "ambiguous" voxels where decision from multiple atlas segmentations did not fully agree. Finally, the classification result combined with the traditional atlas fusion result was demonstrated to achieve improved accuracy compared to the baseline method in the experimental results on rib cage segmentation and HN organs [174]. Finally, the HAS techniques show potential in bringing improvements in regions with distinguishable image intensity features, but no clear evidence for time saving has been reported.

The Chapter 3 and Chapter 4 of this thesis manuscript contains an evaluation of traditional atlas-based fusion methods (STAPLE, PF, and MV) and one hybrid algorithm (RF) for OARs and CTVn, respectively on HN CT images.

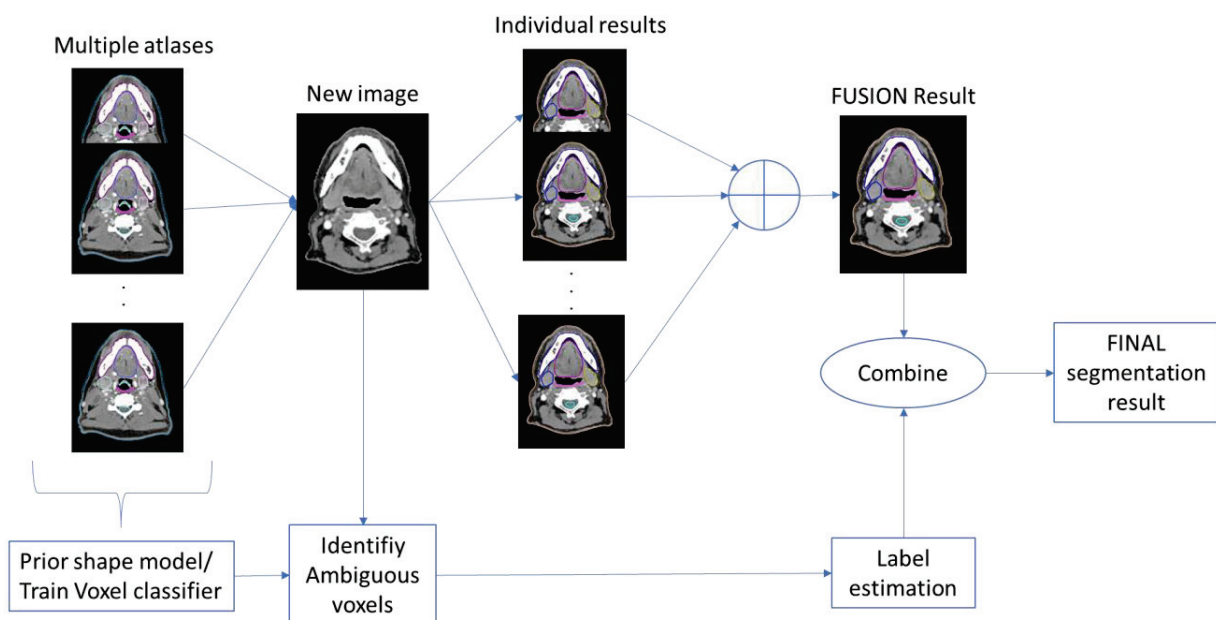


Figure 1.7-3 Hybrid atlas-based segmentation framework

Deep Learning methods

Enormous amount of work has been recently directed into developing image segmentation approaches based on DL models [175,176]. One of the milestones in DL-based segmentation was the introduction of CNN composed of multiple hidden convolutional layers and a fully connected layer. Some of the most well-known CNN architectures include: AlexNet [121], VGGNet [177], ResNet [122], GoogLeNet [178], MobileNet [179], and DenseNet [180].

Long et al. [181] was the first to propose a fully convolutional network (FCN) configuration. By replacing all fully-connected layers with fully-convolutional layers the model was able to handle arbitrary size-input and produce correspondingly-sized output, in an end-to-end segmentation fashion. Furthermore, the authors introduced skip connections such that the information from deep coarse layers is combined with the information from shallow, fine layers in order to produce more detailed and accurate segmentations. By demonstrating that models can be trained on variable sized images, this work was considered a milestone in image segmentation research. However, the limitation when using FCN and CNN is that after many layers of pooling, the final resolution is potentially low or suboptimal. To address this issue, authors of U-net network proposed to add a symmetrical part to the first convolutional part, where the pooling layers were replaced by interpolation layers and a large number of feature channels were added in this up-sampling part to support the propagation of the contextual information to higher resolution layers [182]. Because the contracting part and the expansive part are fairly equal in size, the model resembles a U-shape architecture (Figure 1.7-4). Since this breakthrough, U-net network architectures have been explored in many medical image applications [183], and its architectural design has been extended to the 3D U-Net [184], V-Net [185] and AnatomyNet [186]. Moreover, Isensee et al. [187] reported state-of-the-art segmentations with a self-configuring U-net network, called nnU-Net, that automatically adapts to any new data set without manual intervention. Besides computational efficiency nnU-Net is also able to cope with limited training data. The network code is implemented in Python using PyTorch framework and is available for users on GitHub (<https://github.com/MIC-DKFZ/nnUNet>).

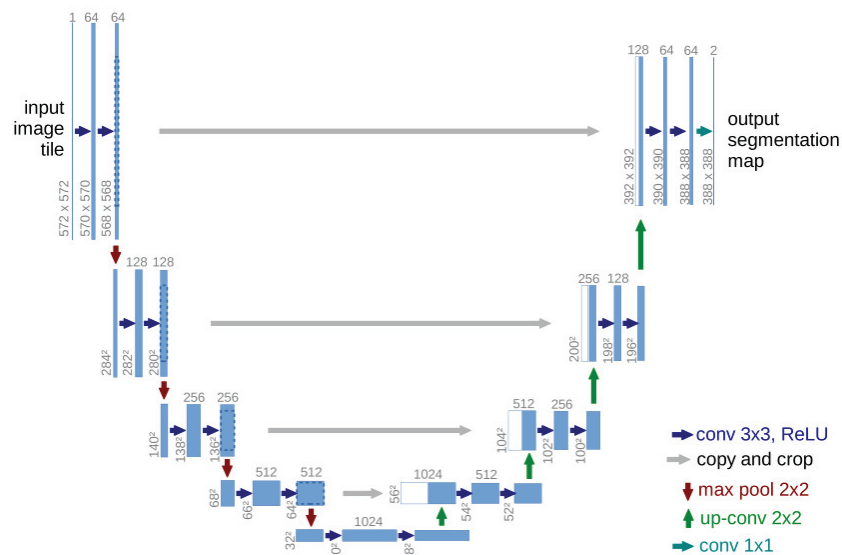


Figure 1.7-4 U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations [182]

Generative Adversarial Network (GAN) is another category of deep learning that had gained interest for segmentation applications due to its data generation capability that can alleviate the problem of training data shortage [188]. A typical GAN integrates two networks, namely a generator network and a discriminator network into one framework. The generator is trained to generate artificial data and the discriminator is trained to differentiate the synthetic samples from real samples (Figure 1.7-5). The training stage is time-consuming since the two components need to be trained sequentially and iteratively in a competing manner to enhance the performance of the other. The final goal is to generate artificial data that cannot be differentiated from real data. Once trained however, only the generator will be used to perform segmentations. A U-net-GAN framework was proposed by Dong et al. [189] and demonstrated superior segmentation accuracy compared to U-Net network configuration alone for OARs in the thoracic region. Moreover, particularly improved results were observed on esophagus when compared to the results from 2017 AAPM Thoracic Auto-segmentation Challenge [190].

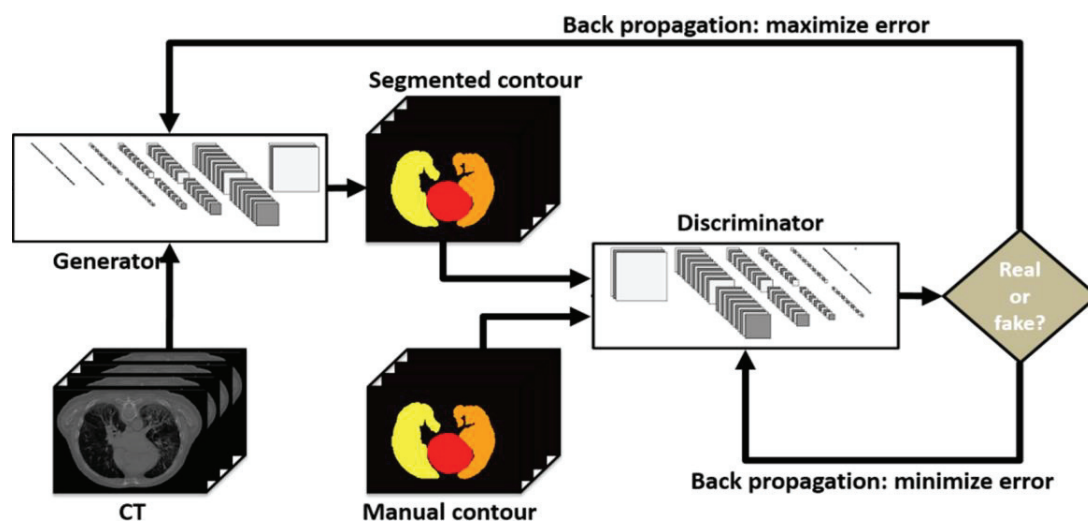


Figure 1.7-5 An example illustrating the process of generative adversarial network; back-propagation is applied in both networks so that the generator produces more realistic segmentation, while the discriminator becomes more skilled at flagging segmented contours against manual contours [189]

Among the recently published studies, Chen et al. [191] propose a whole-body CT segmentation model that combines three OARs segmentation models with one anatomic site detection model. In total the model is able to automatically segment 50 OARs across three localizations. All models are ramifications of U-net architecture, specifically Ua-Net [192] was used for HN localization, 2.5 U-net [182] for the thorax model and 3D U-Net [184] for abdomen and pelvis. The final WBNet model was trained and evaluated on the same data against two similar DL networks (nn-Unet [187] and AnatomyNet [186]) and one commercial atlas-based solution (ABAS [152]). For the majority of OARs, WBNet superiority was demonstrated in terms of geometrical accuracy by means of DICE and $HD_{95\%}$ distance. Moreover, compared to manual delineation it was reported that WBNet was able to significantly reduce the delineation time by 51%, 73%, 86% and 71% for HN, thorax, abdomen and pelvis, respectively. Furthermore, they conducted a dosimetric study on 20 HN patients and observed that the dose differences were clinically acceptable for most cases. However, particular attention was suggested for small volume organs or OARs with relatively lower DICE such as the brachial plexus, hypophysis, optic nerve, optic chiasm, and sublingual gland [191]. The study has a particular strength since it performs an evaluation of four systems using the same data for training and for testing, thus providing pertinent comparison between different DL solutions.

Another recent study, investigated general (diversified multi-institutional datasets for training) and custom (single-institution data sets for training) CT-based AS models for 42 OARs in 3 major tumor sites:

HN, male pelvis and abdomen [124]. Additionally, they evaluated an adaptive spatial resolution (ASR) approach for the narrow or small volume OARs, which proved to improve the accuracy for the eye lens, optical nerves, inner year and bowels. Overall, even with smaller datasets for training, the custom models performed slightly better than the multi-centric models. With regards to the contouring times, after counting for the necessary manual editing time, the whole process of delineation was reduced by 80%, 88% and 65% for HN, pelvis and abdomen, respectively [124].

From most recent studies, state-of-the-art DICE results for CT-based OARs segmentation can be followed for HN, thorax, abdomen and pelvis region in Table 1.7-2, , Table 1.7-4, and Table 1.7-5, respectively. Compared with other networks from literature, WEBNet showed superior DICE results for several OARs such as larynx, lenses, oral cavity, lungs, gallbladder, kidneys, pancreas, and femoral heads. Competitive results were from the custom-made models, having the highest overlap results for the optical nerves, the heart, duodenum, liver, stomach, small and large bowel, bladder and rectum.

Table 1.7-2 Review of DICE results for organs-at-risk in Head and Neck region.

	Chen et al. [191] (trained and tested on same data)				Ibragimov et al. [193]	Willems et al. [194]	Nikolov et al. [195]	Amjad et al. [124]
	WBNet	ABAS	Anatomy-Net	nnUnet	CNN	CNN (DeepVoxNet)	CNN	DCNN (ResUnet3D)
Brainstem	0.87	0.78	0.86	0.90	n.a.	0.92	0.88	0.90
Eye L	0.92	0.88	0.90	0.84	0.88	n.a.	0.95	0.91
Eye R	0.93	0.89	0.91	0.86	0.88	n.a.	0.95	0.91
Larynx	0.90	0.81	0.88	0.79	0.86	0.71	n.a.	0.85
Lens L	0.83	0.62	0.75	0.77	n.a.	n.a.	0.81	0.74
Lens R	0.84	0.56	0.76	0.79	n.a.	n.a.	0.80	0.80
Mandible	0.94	0.88	0.92	0.94	0.99	0.96	0.96	0.88
Optic Chiasm	0.64	0.52	0.61	0.69	0.37	n.a.	n.a.	n.a.
Optic nerve L	0.76	0.59	0.70	0.72	0.64	n.a.	0.76	0.78
Optic nerve R	0.75	0.60	0.71	0.75	0.64	n.a.	0.77	0.79
Oral Cavity	0.91	0.86	0.88	0.91	n.a.	0.84	n.a.	n.a.
Parotid L	0.85	0.68	0.81	0.80	0.76	0.86	0.85	0.82
Parotid R	0.85	0.71	0.81	0.78	0.78	0.90	0.85	0.82
SMG L	0.82	0.61	0.75	0.79	0.69	0.79	0.85	0.77
SMG R	0.82	0.55	0.75	0.80	0.73	0.88	0.85	0.80
Spinal Cord	0.86	0.86	0.86	0.91	0.87	0.96	0.88	0.88

Abbreviations: L= left, R=right, SMG=Submandibular gland; n.a.= not available; bold numbers highlight the best results

Table 1.7-3 Review of DICE results for organs-at-risk in thorax region.

	Chen et al. [191] (trained and tested on same data)				Feng et al. [196]	Yang et al. [190]
	WBNet	ABAS	AnatomyNet	nnUnet	DCNN	DCNN
Heart	0.91	0.85	0.88	0.93	0.93	0.93
Lung R	0.98	0.95	0.96	0.98	0.98	0.97
Lung L	0.98	0.96	0.96	0.98	0.97	0.97
Spinal Cord	0.90	0.86	0.86	0.91	0.84	0.88
Esophagus	0.76	0.54	0.71	0.81	0.61	0.72

Abbreviations: L= left, R=right; bold numbers highlight the best results.

Table 1.7-4 Review of DICE results for organs-at-risk in abdomen.

	Chen et al. [191] (trained and tested on same data)				Zhou et al. [197]	Gibson et al. [198]	Kim et al. [199]	Amjad et al. [124]
	WBNet	ABAS	AnatomyNet	nnUnet	FCN	DenseVNet	CNN	DCNN
Duodenum	0.77	0.40	0.72	0.74	0.76	0.63	0.81	0.82
Gallbladder	0.87	0.53	0.81	0.87	0.65	0.73	0.59	n.a.
Kidney L	0.96	0.83	0.91	0.88	0.91	0.93	0.90	0.96
Kidney R	0.96	0.81	0.93	0.89	0.92	n.a.	0.91	0.94
Liver	0.94	0.91	0.95	0.96	0.95	0.95	0.96	0.97
Pancreas	0.84	0.52	0.75	0.83	0.62	0.75	n.a.	0.76
Spleen	0.96	0.80	0.93	0.95	0.92	0.95	n.a.	0.96
Stomach	0.90	0.58	0.83	0.91	0.76	0.87	n.a.	0.95

Abbreviations: L= left, R=right; n.a.= not available; bold numbers highlight the best results.

Table 1.7-5 Review of DICE results for organs-at-risk in pelvic region

	Chen et al. [191] (trained and tested on same data)				Men et al [200]	Liu et al. [201]	Amjad et al [124]
	WBNet	ABAS	AnatomyNet	nnUnet	DDCNN	CNN	DCNN
Large Bowel	0.80	0.37	0.77	0.82	0.62	n.a.	0.87
Small Bowel	0.82	0.41	0.78	0.81	0.65	0.83	0.84
Femoral R	0.94	0.90	0.93	0.86	0.92	0.90	n.a.
Femoral L	0.93	0.91	0.90	0.85	0.92	0.90	n.a.
Bladder	0.93	0.62	0.89	0.92	0.93	0.92	0.96
Rectum	0.80	0.61	0.76	0.76	n.a.	0.79	0.89

Abbreviations: L= left, R=right; n n.a.= not available; bold numbers highlight the best results

Important to mention are results from auto-segmentation challenges, where teams had access to the same datasets and tested different network configurations [140,190]. Moreover, such competitions allowed other studies as well to benchmark their results on the same datasets. For thorax and HN region, DL-based methods' results using international challenge data sets can be followed in Table 1.7-6 and Table

1.7-7, respectively. The 2017 AAPM thoracic AS challenge provided a benchmark dataset of 60 thoracic CT images, separated in 36, 12 and 12 for training, offline testing and online testing, respectively [190]. Out of the 5 OARs, the esophagus had the lowest DICE scores and the largest variations among the solutions. This can be explained by its difficult differentiation on CT images.

Similarly, the 2015 MICCAI HN AS challenge provided a benchmark dataset of 40 CT images, divided into 25, 10 and 5 for training, off-site testing and on-site testing, respectively [140]. Among the 9 OARs contours, the lowest contour agreement was observed for the chiasma followed by the optical nerves. The MRI-aided method provided improved result particularly for these structures and additionally for submandibular glands. For the rest of the OARs (parotids, mandible and brainstem) the CT image contrast provided good organ differentiation with $DICE > 0.80$ and $HD_{95\%} < 5\text{mm}$.

Table 1.7-6 DL-based methods using the 2017 AAPM Thoracic Auto-segmentation Challenge datasets; * participating teams at the auto-segmentation thorax challenge [176,190]

Metric	Organ	DCNNN Team Elekta*	3D U-Net [196]	Multi-class CNN Team Mirada*	2D ResNet Team Beaumont*	3D and 2D U-Net Team WUSTL*	U-Net GAN [189]
DICE	Esophagus	0.72 ± 0.10	0.72 ± 0.10	0.71 ± 0.12	0.61 ± 0.11	0.55 ± 0.20	0.75 ± 0.08
	Heart	0.93 ± 0.02	0.93 ± 0.02	0.91 ± 0.02	0.92 ± 0.02	0.85 ± 0.04	0.87 ± 0.05
	Lung L	0.97 ± 0.02	0.97 ± 0.02	0.98 ± 0.02	0.96 ± 0.03	0.95 ± 0.03	0.97 ± 0.01
	Lung R	0.97 ± 0.02	0.97 ± 0.02	0.97 ± 0.02	0.95 ± 0.05	0.96 ± 0.02	0.97 ± 0.01
	Spinal Cord	0.88 ± 0.04	0.89 ± 0.04	0.87 ± 0.11	0.85 ± 0.04	0.83 ± 0.08	0.90 ± 0.04
HD95%	Esophagus	7.3 ± 10.31	8.71 ± 10.59	7.8 ± 8.17	8.0 ± 3.80	37.0 ± 26.88	4.52 ± 3.81
	Heart	5.8 ± 1.98	6.57 ± 1.50	9.0 ± 4.29	8.8 ± 5.31	13.8 ± 5.49	4.58 ± 3.67
	Lung L	2.9 ± 1.32	2.10 ± 0.94	2.3 ± 1.30	7.8 ± 19.13	4.4 ± 3.41	2.07 ± 1.93
	Lung R	4.7 ± 2.50	3.96 ± 2.85	3.7 ± 2.08	14.5 ± 34.4	4.1 ± 2.11	2.50 ± 3.34
	Spinal Cord	2.0 ± 0.37	1.89 ± 0.63	2.0 ± 1.15	2.3 ± 0.50	8.10 ± 10.72	1.19 ± 0.46

Abbreviations: L= left, R=right; bold numbers highlight the best results

Table 1.7-7 DL-based methods using the 2015 MICCAI Head-and-Neck AS Challenge datasets [140,176]

Metric	Organ	Shape model constrained FCN [140]	2-stage U- Net [202]	AnatomyNet [186]	DL-based [192]	Synthetic MRI-aided [203]	3D U-Net [184]	3D-CNN [204]
DICE	Brainstem	0.87 ± 0.03	0.88 ± 0.02	0.87 ± 0.02	0.87 ± 0.03	0.91 ± 0.02	0.80 ± 0.08	n.a.
	Chiasm	0.58 ± 0.1	0.45 ± 0.17	0.53 ± 0.15	0.62 ± 0.01	0.73 ± 0.11	n.a.	0.58 ± 0.17
	Mandible	0.87 ± 0.03	0.93 ± 0.02	0.93 ± 0.02	0.95 ± 0.01	0.96 ± 0.01	0.94 ± 0.02	n.a.
	Optic Nerve L	0.65 ± 0.05	0.74 ± 0.15	0.72 ± 0.06	0.75 ± 0.07	0.78 ± 0.09	0.72 ± 0.06	0.72 ± 0.08
	Optic Nerve R	0.69 ± 0.05	0.74 ± 0.09	0.71 ± 0.1	0.725 ± 0.06	0.78 ± 0.11	0.70 ± 0.07	0.70 ± 0.09
	Parotid L	0.84 ± 0.02	0.86 ± 0.02	0.88 ± 0.02	0.89 ± 0.02	0.88 ± 0.04	0.87 ± 0.03	n.a.
	Parotid R	0.83 ± 0.02	0.85 ± 0.07	0.87 ± 0.04	0.88 ± 0.05	0.88 ± 0.06	0.85 ± 0.07	n.a.
	SMG L	0.76 ± 0.06	0.76 ± 0.15	0.81 ± 0.04	0.82 ± 0.05	0.86 ± 0.08	0.76 ± 0.09	n.a.
	SMG R	0.81 ± 0.06	0.73 ± 0.01	0.81 ± 0.04	0.82 ± 0.05	0.85 ± 0.01	0.78 ± 0.07	n.a.
HD _{95%} (mm)	Brainstem	4.01 ± 0.93	2.01 ± 0.33	n.a.	n.a.	n.a.	n.a.	n.a.
	Chiasm	2.17 ± 1.04	2.83 ± 1.42	n.a.	n.a.	n.a.	n.a.	2.81 ± 1.56
	Mandible	1.50 ± 0.32	1.26 ± 0.50	n.a.	n.a.	n.a.	n.a.	n.a.
	Optic Nerve L	2.52 ± 1.04	2.53 ± 2.34	n.a.	n.a.	n.a.	n.a.	2.33 ± 0.84
	Optic Nerve R	2.90 ± 1.88	2.13 ± 2.45	n.a.	n.a.	n.a.	n.a.	2.13 ± 0.96
	Parotid L	3.97 ± 2.15	2.41 ± 0.54	n.a.	n.a.	n.a.	n.a.	n.a.
	Parotid R	4.20 ± 1.27	2.93 ± 1.48	n.a.	n.a.	n.a.	n.a.	n.a.
	SMG L	5.59 ± 3.93	2.86 ± 1.60	n.a.	n.a.	n.a.	n.a.	n.a.
	SMG R	4.84 ± 1.67	3.44 ± 1.55	n.a.	n.a.	n.a.	n.a.	n.a.

Abbreviations: L= left, R=right, SMG= submandibular gland; bold numbers highlight the best results

Lastly, a review of commercially available AS solutions is summarized in Table 1.7-8, together with studies attesting their clinical validation. Majority of the study show of the superiority of DL-based solutions over the atlas-based algorithm, in both improved contour consistency through geometrical accuracy, and time savings when considering the manual corrections. For instance, Zabel et al. [205] quantified >50% compared to 20% time reduction when using a DL-based solution compared to an atlas-based solution for the contouring of the rectum and bladder. Similarly, for a set of 7 OARs in HN van Dijk et al. [143] has demonstrated significant time savings when using a DL solution compared to an atlas-based solution, especially when corrections were performed by a young observer. They also showed a potential reduction of >50% compared to the standard manual delineation. Another study presented the experience from clinical integration of three commercial AS solutions (e.g. ART-plan Annotate (Therapanacea, Paris, France), DLCExpert (Mirada Medical, Oxford, UK) and RayStation v9B (RaySearch Laboratories, Stockholm, Sweden)) in three centers in France [206]. Additionally, the paper presented state-of-the-art of DL solutions for medical image segmentations as well as recent recommendations on implementation, commissioning and QA of AI methods, proposed by the European Society of Radiation and Oncology (ESTRO) [207]. The observed benefit of integrating DL contours in clinical practice, was reflected through objective measures such as DICE or Jaccard coefficient and subjective assessment upon their clinical acceptability performed by doctors. ART-plan Annotate solution used >600 data for model training depending on localization and provided OARs and target contours for male pelvis, thorax, breast, HN, digestive and the brain. From the experience at Léon Bérard Cancer Center (Lyon), most of the OARs contours in pelvis and breast region were 100% validated by doctors. However, for thorax and HN, only 67% and 85% respectively were accepted by the experts. Regarding targets, the DL contours were 100%

accepted on the male pelvis and the breast but none were accepted on HN localization. The deployment of DLCExpert (>200 data for model training) at the European Hospital Georges Pompidou (Paris) resulted in 100% acceptance of the DL contours for OARs and targets on breast localization. At the Pasteur-ONCORAD Clinic (Toulouse), the commissioning of RayStation v98 solution, trained on about 100 patient data, resulted in 100% validated OAR and target contours for the male pelvis region. Another positive outcome from implementation of these solutions was the harmonization of organs' nomenclatures which may support creation of prospective databases, that are extremely valuable in the growing context of machine learning. They also highlighted the lack of shared experiences among centers and the necessity of understanding the AI basis in order to have a critical and constructive communication with the industry.

To be noted that one of the most important considerations when developing a DL-based solution is the collection of the data, that requires proper curation and pre-processing (e.g. data augmentation, image resizing, image cropping, image normalization) prior to training. Moreover, in supervised learning-based methods, a bias introduced by the physicians' ground truth manual contours is to be expected. The computing power may also represent a limitation when training a model, therefore depending on the GPU power availability and the network design, some methods use the whole image as input, whereas others use 2D image-based approaches to reduce the computational costs. 3D feature information has also been exploited with the use of 2D kernels on multidirectional 2D images or ultimately with 3D patches as network input. A post-processing step is also often applied for the refinement (smoothing) of contour boundaries using morphological operations such as conditional random fields. GAN-based methods are expected to be used more in the future for penalizing implausible structures. Finally, the multi-organ class imbalance remains a continuous problem challenging the choice of an appropriate loss function to overcome the segmentation accuracy over a wide range of structures [208]. To increase the community awareness, challenges of data-limitation are well described in a survey on DL methods for medical image segmentation with few associated solutions in order to further inspire the efforts in this impactful area of research [209].

The two contributions presented in [Chapter 3](#) and [Chapter 4](#) of this thesis manuscript, present the results from evaluating the commercial DL model (ART-plan Annotate, Therapanacea, France) trained with multi-centric data and another non-commercial DL model trained exclusively with data from a single center, for the segmentation of OARs and CTVn levels on HN CT images.

Table 1.7-8 Commercial solutions for automatic segmentation

Supplier	Product Name	Method	Segmentation site	TPS integrated	Reference
Accuray	multiPlan 5.0	AB	HN, brain, pelvis, thorax	Yes	Chaney et al. (2004) [210]
BrainLab	iPlan	AB	HN, rain, pelvis, spine, thorax	Yes	Grosu et al. (2003) [211]
DOSisoft	IMago	AB	HN, brain	Yes	Commowick et al. (2008) [212]
Elekta	ABAS 2.0	AB	HN, pelvis	No	Han et al. (2008) [159] Liu et al. (2016) [168]
	ADMIRE	Hybrid AB	HN	No	McDonald et al. (2021) [213]
	ADMIRE-DL	DL-based	HN, thorax	No	Yang et al. (2018) [190] McDonald et al. (2021) [213]
MIM Software	MIM Maestro 6+	AB	HN	No	Hu et al. (2008) [125]
	ProtégéAI	DL-based	HN, prostate	No	Urago et al. (2021) [163]
Mirada	RTX	AB	HN, pelvis, thorax	No	Gooding et al. (2013) [214]
	DLCExpert	DL-based	HN	No	van Dijk et al. (2020) [143]
Philips	SPICE 9.8	AB model-based	HN, abdomen, pelvis, thorax	Yes	Qazi et al. (2011) [169]
RaySearch Laboratorie	RayStation 4.0	AB model-based	HN, abdomen, pelvis, thorax	Yes	Stewart et al. (2010) [215]
	DLS (Raystation 11B)	DL-based	HN, breast, abdomen, pelvis, thorax	Yes	-
Varian	Smart Segmentation	AB	HN, thorax, pelvis	Yes	Haas et al. (2008) [216] Zabel et al. (2021) [205]
Velocity Therapanacea	Velocity AI 3.0.1	AB model- based	HN, brain, pelvis	No	Stapleford et al. (2010) [160]
	ARTplan-Annotate	DL-based	HN, thorax, abdomen pelvis	No	Ung et al. (2020) [217]
Limbus AI	Limbus Contour	DL-based	HN, breast, thorax, abdomen pelvis	No	Wong et al. (2020) [218] Zabel et al. (2021) [205]
Siemens	Syngo.Via RT	DL-based	HN, thorax, abdomen pelvis	No	-
RAD formation	AutoContour	DL-based	HN, thorax, abdomen, pelvis, brain	No	-
MVISION	MVision AI	DL-based	HN, breast, thorax, abdomen, pelvis	No	-

Abbreviations: AB = atlas-based, DL = deep learning, HN = head-and-neck

1.8. Automated treatment planning

The goal in the RT treatment planning step is to determine the optimal irradiation parameters (number of beams or arcs, beam angles, shape of the collimators etc.) to achieve a desired dose distribution. Normally, the task is carried out by a human user on a dedicated TPS. Manual treatment planning is a labor-intensive task that implies interactive optimization, based on trial and error, to reach the best trade-off between all dose-volume objectives. Moreover, it is strongly dependent on the planner's skills and experience. The difficulty of a plan increases with more OARs accounted for in the optimization. This increases further the variation in plan quality results obtained between planners. Auto-planning techniques have the potential to overcome these issues and reduce operator inconsistencies while shortening the time a planner would spend on manual optimization. The ultimate goal of auto-planning is efficient generation of high-quality plans.

From the literature [219,220] and from RT solutions vendors, we can identify several categories of auto-planning solutions: knowledge-based planning (KBP), protocol-based automatic interactive optimization (PB-AIO), multicriteria optimization (MCO) also known as multi-objective optimization, and DL methods.

Knowledge-based planning (KBP)

KBP relies on a database of prior treatment plans to predict the best achievable plan based on the anatomical information of the new patient. It can be atlas-based or model-based. In the atlas-based approach the knowledge from atlases is used to select the closest matching patient to give a better starting point of the inverse optimization [221–224]. An interesting work of Chanyavanich et al. presents a method of predicting starting parameters of the treatment machine using a data-base of prior fixed-field IMRT plans [224]. Further, the model-based approaches are able to build predictive models by using prior information about the relationship between anatomical and geometrical features from clinically acceptable plans. Dose-volume histograms (DVHs) can be used to create predictive site-specific models based on similar contours and quality of prior treatment plans [225–247]. The limitation is that the DVHs are only predicted for the regions of interest (ROIs) that are delineated in the database. This means that other volumes which a human planner may also optimize to reduce dose, may not be considered. An alternative is voxel-based dose prediction where dose to individual voxels can be predicted from the prior plans [248–252]. For this approach the drawback is that the quality of the plans generated in the past directly impacts the plan quality for the new patients.

Published studies on clinical implementation of KBP are via the use of the RapidPlan commercial solution from Varian Medical Systems. The results show equally or slightly better plan quality compared to manual planning with improved consistency and efficiency. Using KBP, Foy et al. [234] reported a time reduction by a factor of 6 in the planning of the VMAT for stereotactic body RT of the spine (from 1–1.5 h down to 10–15 min). Most studies are focused on single institution experience and results. However, Fogliata et al. [239] performed a study on esophageal cancer involving three centers with different dose protocols and demonstrated dosimetric improvements when compared to the reference data especially for the center that did not participate with data for training the KBP model. Similarly, other studies' results highlight the potential benefit of a heterogeneous dataset, as well of outliers, in contributing to the model strength [224,253].

Protocol-based automatic interactive optimization

The PB-AIO approach mimics the iterative optimization steps of a human operator starting from a user-defined template of parameters including beam configuration and planning goals for targets and OARs [254–269]. The automatic solution will iteratively adjust the planning parameters to meet the required planning goals. One of the drawbacks of the method is that it is dependent on the experience of the

planner to define the set of settings and adequate protocol. Secondly, it is hard to judge if the resulting plans could potentially be further improved.

Studies have reported results for the clinical implementation of the commercial PB-AIO solution AutoPlanning in Pinnacle³ TPS (Philips Radiation Oncology Systems, Fitchburg, WI) with either equivalent or superior quality compared to both IMRT and VMAT manual planning. Apart from evidence of time reduction [266], studies have also reported IOV reduction when PB-AIO auto-planning strategy was used [257]. Contrarily, certain studies were arguing the feasibility of fully automated PB-AIO, indicating situations where further manual optimization was still deemed necessary [265].

Multicriteria optimization

The core of the MCO approach is the concept of “pareto optimal solution” which represents a plan that can only be further improved with the cost of degrading at least one of the dose-volume objectives. There are two directions of this approach: *posteriori* and *a priori*.

In the *posteriori*-MCO also called the “pareto surface based technique”, the system creates a database of plans (also called “pareto surface”) that satisfy different planning goals and where any change in dose to one organ results in a trade-off in another organ [91]–[104]. This, allows the physician to explore the compromise between different planning goals and choose the preferred plan for the case. One of the method’s limitation is the computing resource power since there is an infinite number of pareto plans that can be generated. Moreover, the plan selection is totally operator-dependent, and it can become challenging especially when there are a large number of clinical objectives under consideration. Another limitation of *posteriori*-MCO method is that the pareto plans do not consider directly the machine parameter optimization, thus the chosen plan has to be converted into a deliverable plan using direct aperture optimization. This can translate into significant dosimetric differences that may require manual adjustments [275]. A commercial solution using *posteriori*-MCO was implemented clinically in RayStation TPS. All studies reported comparable or better plan quality compared to manual planning with benefit in time reduction up to 88min [273]. Chen et al. [280] used the DVH information from IMRT automatically generated plans in RayStation TPS, to optimize VMAT plans with shorter delivery times and less monitor units (MUs). They showed that most of the times the plan quality was conserved and particularly better results were observed for HN cases and hypo-fractionated prostate cases, whereas standard fractionated prostate cancer cases required further modification of the objectives’ or constraints’ weights [280]. In another study on lung cancer [273], evaluation of the plan quality between manual and MCO solution was conducted via a double blinded examination and concluded that clinicians preferred the automated plans in 8/10 cases, whereas the 2 situations chosen in the favor of the manual planning were due to better skin and spinal cord sparing but at the expense of higher esophagus dose.

In the *a priori*-MCO approach, a single pareto-optimal plan is generated based on a wish-list with predefined clinical dose objectives and constraints [2,284–294]. Each objective and constraint have a priority order, and the automatically generated treatment plan contains clinically favorable trade-off between all the treatment goals. The idea of this method is to have a single wish-list per clinical protocol that provides consistent results over all the patients. Its limitation is that it is highly dependent on the experience of the person defining the wish-list.

Erasmus-iCycle is an in-house developed *a priori*-MCO algorithm that was validated as beam fluence pre-optimizer for the Monaco TPS (Elekta AB, Stockholm, Sweden) after demonstrating superiority over manual planning in several clinical sites [287,295]. In a prospective study on HN cancer patients, the Erasmus-iCycle plans were preferred in 97% of the cases when compared to manually optimized plans [287]. An *a priori*-MCO solution fully integrated in Monaco TPS is not yet clinically available. Based on the Erasmus-iCycle algorithm, Elekta AB vendor developed the mCycle solution, and studies have been initiated to validate its clinical implementation. Two studies evaluated and confirmed the feasibility of

using the mCycle solution for plan adaptation on prostate and rectal cancer [289,290]. Although both studies reported a slight increase in the MUs in the automated plans, this had no impact on deliverability and clinical acceptability of the plans. The robustness of the wish-list has been demonstrated for VMAT planning of HN cancer patients, where a blind evaluation of 2 physicians confirmed the preference of mCycle plans over the manual ones [2]. This work represents the first contribution of this thesis study and will be detailed in the [Chapter 2](#).

Deep learning solutions

DL-based solutions have also been investigated for auto-planning, where two approaches can be distinguished among studies: direct generation of fully 3D dose predictions or generation of fluence maps that can be converted into deliverable treatment plans by a commercial TPS. With respect to studies on HN cancer patients, dose distribution predictions were obtained either by using GAN [296] or U-net network [297]. Both approaches obtained realistic plans that better satisfied clinical criteria. Similarly, in other studies, in only few seconds, fluence maps were generated using GAN [298] or U-net network [299] with acceptable plan quality. This evidence holds great potential for clinical applications and real-time planning. A commercial DL solution was proposed by Raysearch Laboratories and its clinical feasibility was demonstrated on prostate [300] and breast [301] cancer localization.

A summary table of commercially available solutions integrated in TPS with details on studies investigating their clinical implementation can be followed in Table 1.8-1.

Table 1.8-1 State-of-the-art of automated treatment planning solutions and their clinical validation studies per tumor localization and in chronological order

Autoplanning Solution	Technique/ method	TPS	Tumor sites	References	Year
RapidPlan™ (commercial solution)	KBP (DVH-guidance)	Eclipse TPS (Varian Medical Systems, Palo Alto, USA)	HN	Krayenbuehl et al.[247]	2015
				Tol et al.[227]	2015
				Fogliata et al.[246]	2017
			Prostate	Krayenbuehl et al.[256]	2018
				Fogliata et al.[231]	2014
				Yang et al.[228]	2015
				Hussein et al.[230]	2016
				Schubert et al.[229]	2017
				Powis et al.[232]	2017
			Breast	Wang et al.[233]	2017
				Spinal metastasis	Foy et al.[234]
			Lung	Fogliata et al.[231]	2014
				Chin Snyder et al.[236]	2016
				Delaney et al.[235]	2017
			Upper GI	Fogliata et al.[240]	2014
Fogliata et al.[239]	2015				
Habraken et al.[238]	2017				
AutoPlanning (commercial solution)	PB-AIO	Pinnacle ³ TPS (Philips Radiation Oncology Systems, Fitchburg, WI)	HN	Hassen et al.[266]	2016
				Hazzel et al.[265]	2016
				Gintz et al.[269]	2016
			Prostate	Speer et al.[267]	2017
				Krusters et al.[268]	2017
				Krayenbuehl et al.[256]	2018
				Kanabu et al.[257]	2017
			Esophagus	Xiadong et al.[258]	2017
				Hansen et al.[259]	2017
			Brain	Krayenbuehl et al.[255]	2017
Wang et al.[260]	2017				
Rectal	Song et al.[261]	2016			
RayStation (commercial solution)	<i>posteriori</i> -MCO	RayStation TPS, Eclipse TPS	HN	Chen et al.[280]	2014
				Kierkels et al.[281]	2015
				Krayenbuehl et al.[256]	2018
			Prostate	Wala et al.[283]	2013
				McGarry et al.[272]	2014
				Chen et al.[280]	2014
				Ghandour et al.[271]	2015
				Müller et al.[282]	2017
			Brain	Craft et al.[279]	2010
				Müller et al.[282]	2017
			Pancreatic	Craft et al.[279]	2010
			Lung	Kamran et al.[273]	2016
			Lower GI	Rønne et al.[274]	2017

	<i>DL methods</i>	Raystation TPS	Breast prostate	Bakx et al.[301] Nilsson et al.[300]	2021 2021
Erasmus-iCycle	<i>a priori</i> -MCO	Combined with Monaco TPS (Elekta AB, Stockholm, Sweden)	HN	Voet et al. [287]	2013
			Prostate	Voet et al.[292]	2014
				Buschmann et al.[291]	2018
			Gastric cancer	Sharfo et al.[294]	2018
			Lung	Della Gala et al.[285]	2017
			Spinal metastases	Buergy et al.[293]	2017
		Cervical	Sharfo et al.[286]	2016	
mCycle (not yet commercially available)	<i>a priori</i> -MCO	Monaco TPS (Elekta AB, Stockholm, Sweden)	HN	Biston et al. [2]	2021
			Prostate	Naccarato et al.[289]	2022
			Rectal	Jagt et al. [290]	2022

Abbreviations: HN=head-and-neck, GI=gastro-intestinal

1.10. Synthetic CT image generation

The use of IGRT methods have raised the interest in plan adaptation based on daily images of the patient. However, in order to be used for advanced tasks such as dose calculation and adaptive treatment planning, the new patient image must contain correlation between pixel intensity information quantified in HU and ED which characterize the tissues (patient anatomy). The CT scan is the only patient image holding a direct correlation between HU and ED, and for that reason remains the patient reference image. At the same time, MRI images offer better soft tissue differentiation and therefore, they became the focus of many research groups aiming to generate sCT from MRI images and thus making them feasible for dose calculations [302]. The same interest goes towards the CBCT images which are frequently (daily or weekly) used in IGRT for accurate patient set up. The challenge of HU inaccuracy comes along other limitations related to CBCT images namely: image artefacts, scatter, poor soft-tissue differentiation and limited field of view. Generation of sCT images from CBCT scans is part of this thesis focus, notably the 4th contribution ([Chapter 5](#)), and therefore will be further elaborated with regards to HN cancer patients.

Initially, several methods have been proposed to reduce the scatter, metal, and beam-hardening artifacts on CBCT images [303–306]. Moreover, DL methods such as CNN have also been explored for image quality improvement [307,308]. Furthermore, rather than focusing on the correction for a specific artefact, methods were proposed for direct generation of sCT from CBCT images, thus enabling their use for both patient positioning and dose calculations [309,310]. QA guidelines and recommendations have recently been published for the validation of sCT solutions prior to clinical implementation [207]. To quantify the accuracy of the HU numbers, metrics such as Mean Error (ME), Mean Absolute Error (MAE), Peak signal to noise rating (PSNR) or structural similarity metrics (SSIM) can be used. Additionally, DICE coefficient can be used to assess the overlap of the bony structures which is relevant for both patient positioning and dose calculations. Furthermore, accuracy of radiation doses must be assessed by evaluating DVH points together with gamma index analysis that quantifies the dose differences in every image point. A table gathering the quantitative metrics for sCT image evaluation is summarized in Table 1.10-1.

From literature, four main approaches of sCT generation with varied complexity can be identified that have been applied to HN localization. Their principle, advantages, disadvantages and associated references are summarized in Table 1.10-2. The simplest method, is the use of CBCT-specific HU-ED conversion curve (Figure 1.10-1). Such a correspondence curve can be established either from phantom images, following the CT scan calibration process with several known ED inserts, or from one or multiple patient CBCT images, resulting in a patient-specific calibration curve. The established HU-ED curve must be applied to the CBCT image so dose calculation can be performed. Additionally, due to limited FOV of the CBCT images, water equivalent density should be assigned where patient information is missing so that radiation dose deposition can be accounted for.

Table 1.10-1 Quantitative metrics for evaluation of synthetic CT images (adapted from [207])

Metric	Principle	Advantages	Disadvantages
Mean Error (ME) / Mean Absolute Error (MAE)	Difference between HU values of sCT and ground truth CT. Paired/voxel-based comparison within a specified volume (e.g. body contour or another ROI)	<ul style="list-style-type: none"> • Frequently reported in studies • Relatively easy to calculate • Can be calculated within different structures/ROI 	<ul style="list-style-type: none"> • Does not show the spread of differences in the voxel • Difficult to compare between studies • Might not be clinically relevant
Peak signal to noise ratio (PSNR)	Ratio between maximum value of a signal and the power of distorting noise that affects the quality	<ul style="list-style-type: none"> • Easy to calculate • Gives some information about the relative error 	<ul style="list-style-type: none"> • No information about position of the error
Structural similarity metric (SSIM)	Used to measure the similarity between two images and is designed to help improve on metrics such as MAE or PSNR	<ul style="list-style-type: none"> • Carries information about interdependencies between pixel values 	<ul style="list-style-type: none"> • More difficult to calculate
DICE coefficient	Overlap of bony structures	<ul style="list-style-type: none"> • Relevant for patient positioning • Relevant for dose calculations 	<ul style="list-style-type: none"> • Dependent on the structure volume • Dependent on thresholding for bones • Gives no information on actual HU values
DVH dose differences	Compare DVHs from the same plan calculated on both sCT and reference CT using either the same structure set or structures deformed through registration	<ul style="list-style-type: none"> • Relatively easy to calculate • Care should be taken when transferring the structure sets. 	<ul style="list-style-type: none"> • Differences can be difficult to interpret; • Discrepancies can be caused by an 'error' in the sCT image, or in the contour or by a difference in the patient anatomy
Gamma Index	Calculate gamma value in every point in the image	<ul style="list-style-type: none"> • Used to give an overall representation of the dosimetric discrepancies • Analysis can be adjusted to preference (dose difference, distance to agreement and threshold) 	<ul style="list-style-type: none"> • Difficult to compare between studies due to chosen criteria

Abbreviations: ROI=region of interest, sCT=synthetic CT, MAE=mean absolute error, PSNR=peak signal to noise ratio, DVH=dose volume histogram, HU=Hounsfield units

Table 1.10-2 Principle, advantages and disadvantages of synthetic CT image generation methods from CBCT images and associated references with regards to studies on HN cancer patients

Method	Principle	Advantages	Disadvantages	References
CBCT-specific HU-ED curve (HU-ED curve)	Establishment of correspondence curve between HU CBCT numbers and densities (from phantom measurements or patient CBCT images)	<ul style="list-style-type: none"> Simple Acceptable precision in the absence of important mobbing artefacts 	<ul style="list-style-type: none"> Dependent on the acquisition protocol and the size of patient/phantom (often implying several curves) Error risk in choosing the appropriate curve Image artefacts (particularly for high density materials) 	<p>Phantom based:</p> <p>Giacometti et al. (2019) [311] Barateau et al. (2020) [310]</p> <p>Patient-specific :</p> <p>MacFarlane et al. (2018) [312]</p>
Density assignment methods (DAM)	ED override or HU override from the CT to CBCT images that are segmented into different classes or ROI	<ul style="list-style-type: none"> Simple Workflow automation Acceptable precision with only few densities classes 	<ul style="list-style-type: none"> Manual verification on CBCT class segmentation Segmentation impacted by image quality (artefacts) Precision dependent on localization 	<p>MacFarlane et al.(2018) [312] Giacometti et al.(2019) [311] Barateau et al.(2020) [310]</p>
Deformable image registration (DIR)	Deformable vector fields obtained between the reference CT and the CBCT image; Applied deformed field the CT to create the "pseudo-CT"	<ul style="list-style-type: none"> Quasi-independent of CBCT HU numbers Contour propagation possible Possibility of workflow automation Fast computation 	<ul style="list-style-type: none"> Evaluation and validation of the DIR algorithm Appearance/disappearance of anatomical marks 	<p>Giacometti et al.(2019) [311] MacFarlane et al.(2018) [312] Marchant et al.(2018) [313] Barateau et al.(2020) [310]</p>
Deep learning for synthetic CT generation (DL-sCT)	Training phase using database of CT and CBCT images Using DL models to directly generate a correspondent synthetic CT from CBCT	<ul style="list-style-type: none"> Quasi-independent of CBCT HU numbers Fast computation 	<ul style="list-style-type: none"> CBCT Image artefacts Computing power Network parameter optimization Number of samples for training 	<p>Supervised training:</p> <p>Li et al.(2019) [314] Chen et al.(2020) [315] Xue et al.(2021) [316]</p> <p>Unsupervised training:</p> <p>Liang et al.(2019) [82] Barateau et al.(2020) [310] Eckl et al.(2020) [317] Xue et al.(2021) [316]</p>

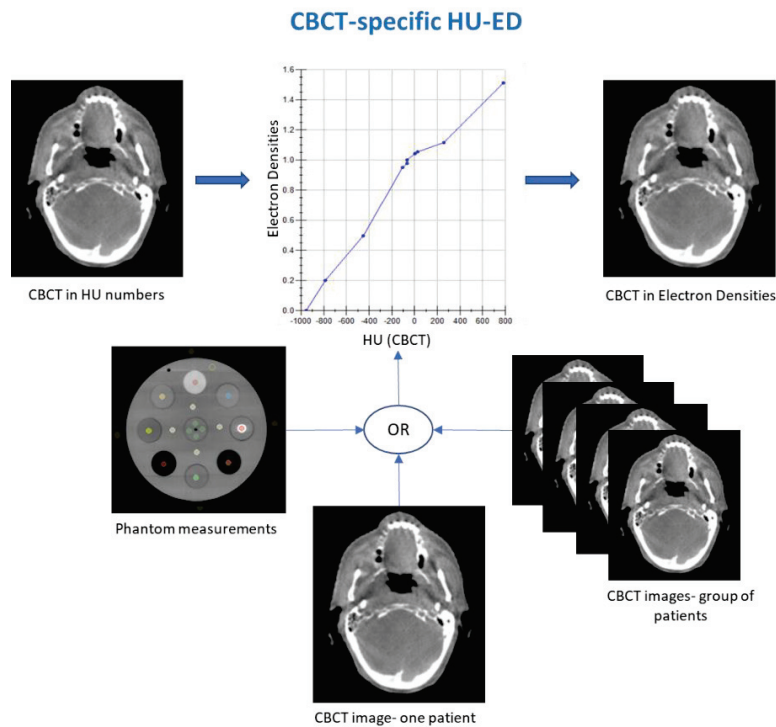


Figure 1.10-1 Establishment of CBCT-specific HU-ED curve (adapted from [309])

The study conducted by Giacometti et al. [311] included evaluation on 5 HN patients and demonstrated dose differences up to 5.4% when the CT standard calibration curve was applied on the CBCT image and 3.9% for a site-specific CBCT calibration curve. However, the gamma analysis (2%/0.1mm, 50% threshold) yielded >94% agreement to the reference dose distributions. Similarly, Barateau et al. [310] evaluated a HU-ED density curve from phantom measurements on 44 HN patients and obtained a MAE of 266.6 HU and a mean 3D gamma pass rate (2%/2mm, 30% dose threshold) of 91%. Contrarily, a patient specific calibration curve approach was evaluated by Macfarlane et al. [312] on 15 HN patients and resulted in average dose metric differences of -0.3% and average gamma pass rates of 95%.

Secondly, the bulk density assignment method (DAM) can be performed based on one or multiple tissue classes (water, bone, air, soft tissue, fat etc.) or based on several regions of interest (ROI) such that the ED or HU number information from reference CT is applied on the CBCT image (Figure 1.10-2). This involves an initial segmentation step on the CBCT image to create the volume that will correspond to the new anatomy of the day. For the density affectation step, several approaches can be followed: based on data from literature, based on a group of patients or individual patient CT scan. From literature, Giacometti et al. [311] evaluated a density override approach based on 7 densities (air, lung, adipose, muscle, soft and dense bone and metal) segmented on both reference CT and the CBCT image. The HU from the CT were introduced on the correspondent segments on the CBCT, leading to dose differences up to 3.2% and gamma pass rates >95%. The authors mention that this override method was found to be the most labor-intensive compared to other techniques, although some steps were automatized. Similarly, Barateau et al. [310] used a 3-class density override (bone, air and soft tissue), where CBCT and CT segmentations were based on HU thresholding. The obtained pseudo-CT images had a MAE of 113.2 HU and provided dose distributions similar to the reference with gamma pass rates of 98%. The other study, MacFarlane et al. [312], quantified an average dose metric difference of -1.1% and gamma pass rate of 94.4% for a density override method that was performed based on CT-CBCT rigid registration. More explicit, the anatomical differences observed on the CBCT were accounted for on the CT image and water or air equivalent

densities were assigned accordingly in the regions where the soft tissue had become air (e.g. from weight loss) or the air cavity has been replaced by soft tissue (e.g. closed air cavity).

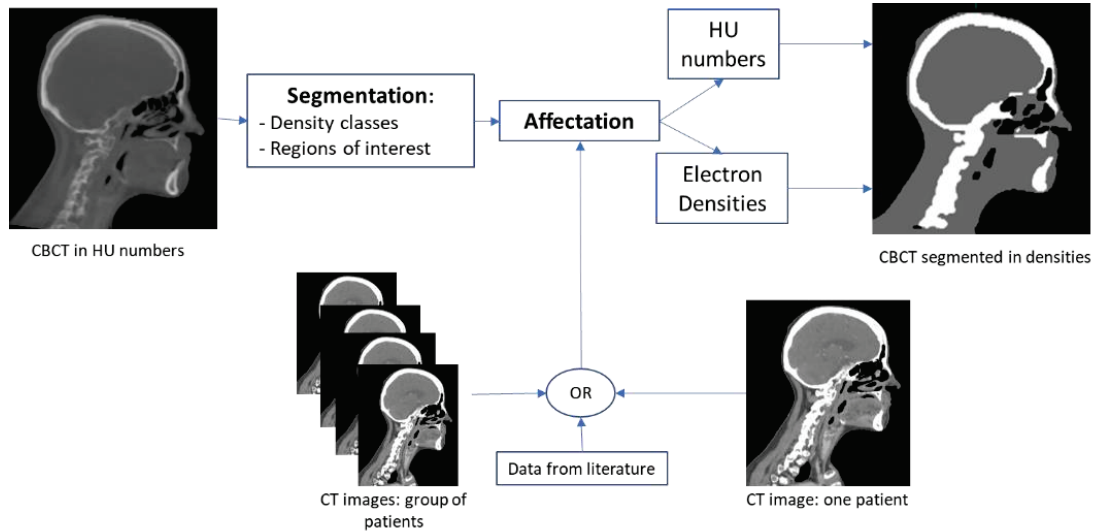


Figure 1.10-2 Density affection method (adapted from [309])

Furthermore, DIR algorithms can be used to deform the reference CT to the CBCT image, while keeping the appropriate CT HU numbers (Figure 1.10-3). The deformation field obtained through the image registration is applied to further deform the CT image and the associated contours, if desired. The new image will hold accurate HU values from the CT scan, and the same calibration curve used for the CT will be applied to this pseudo-CT to enable dose calculation. Results from literature show overall better dose accuracy for the DIR methods compared to the forth mention, HU-ED or DAM methods [310–312]. A MAE of 95.5 HU has been reported for the DIR method used in [310], which was better than for the HU-ED and DAM method in the study. Similarly, an evaluation conducted for Elastix and Niftireg DIR algorithms resulted in accurate dose recalculations with <1% dose error [313]. The main difficulty of this method remains the limited FOV of the CBCT image which for the HN cancer patients results in truncation of the shoulder region.

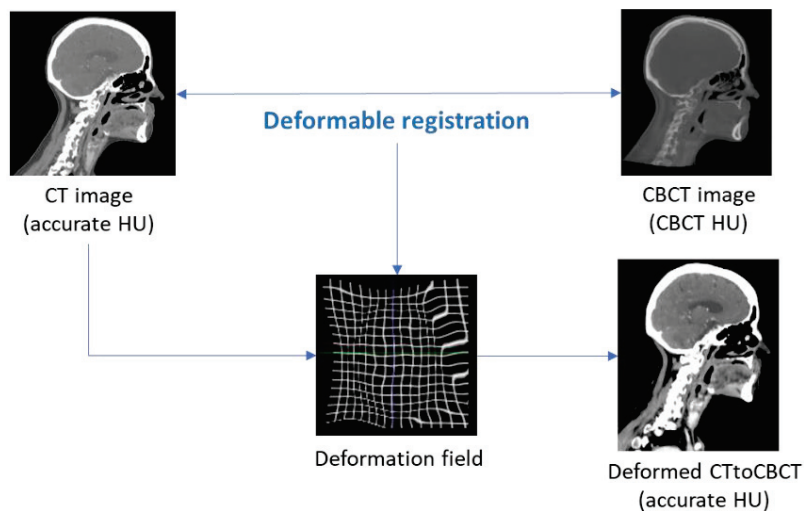


Figure 1.10-3 Deformable image registration between CT and CBCT images (adapted from [309])

Ultimately, DL methods offer the attractive possibility of directly generating sCT images with CT image quality in only few seconds. However, more work is required in the training phase. DL models can be trained either through supervised learning using CNNs and paired CT-CBCT training samples (Figure 1.10-4a), or through unsupervised learning using GAN architectures and unpaired CBCT and CT images for training (Figure 1.10-4b). The second method becomes more interesting when such paired CT-CBCT images are not available. By using GAN methods, unpaired data can be used to train generators that will generate images from the CBCT domain to the CT image domain, and discriminators that will be used to distinguish sCT images from real CT data based on the image distribution. In cycleGAN, 2 generator networks are trained in order to generate sCT images from CBCT to CT image domain and synthetic CBCT (sCBCT) images from CT to CBCT image domain. Then, 2 discriminators will be used to distinguish between the sCT and real CT data and between the sCBCT and the real CBCT images. Note that in this configuration, the sCBCT images are only a by-product used to calculate adversarial losses and only the CBCT to sCT generator will be used for further deployment. Based on this feedback loop, the accuracy of the generator networks is increased.

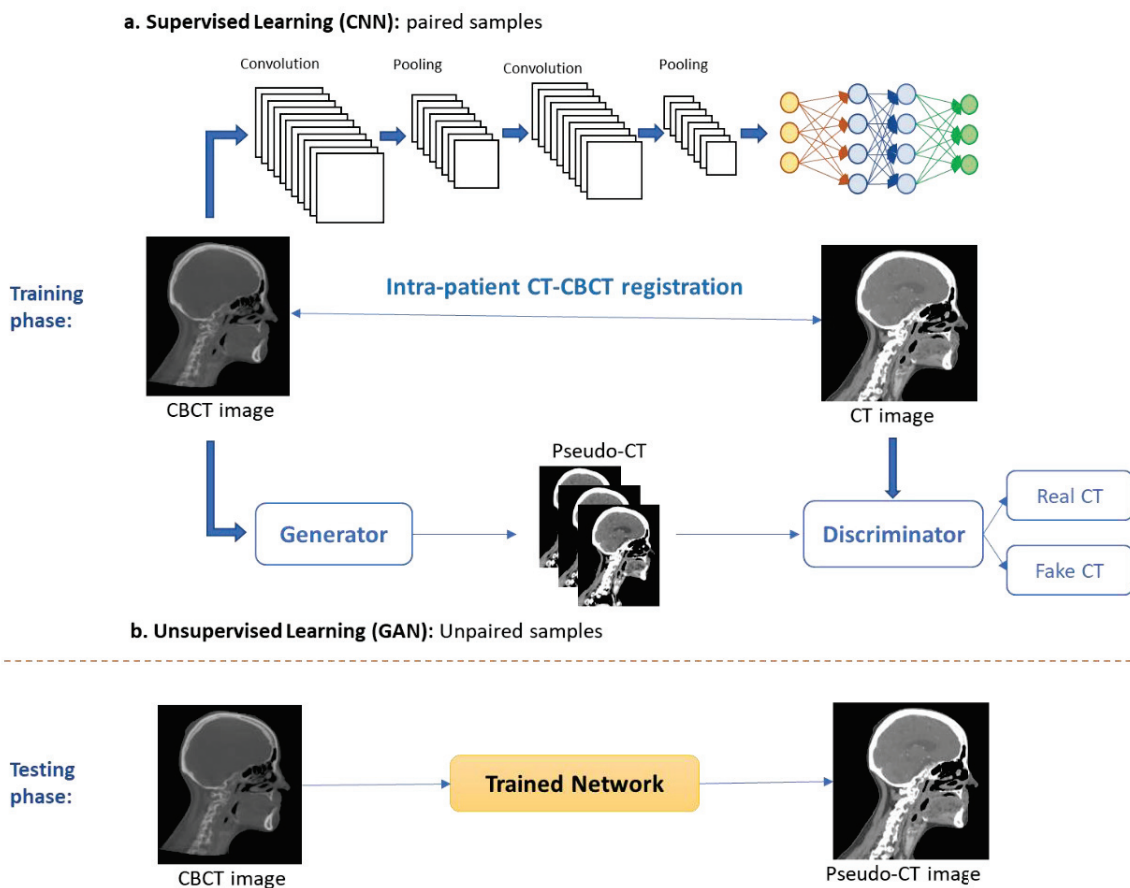


Figure 1.10-4 DL models training for synthetic-CT image generation

Using the supervised training approach, one study reported results from a 2D U-Net neural network trained on 50 CBCT/CT paired images of HN cancer patients [314]. Compared with the original CBCT images, the range of MAE between reference CT and the sCT images improved from (60; 120) HU to (6; 27) HU. Compared with the CT reference, the average DVH dose metric differences were 0.2% and the average gamma pass rates (1%/1mm) were 95.5%. The DCNN model demonstrated higher accuracy of

dose calculations when compared with that on the original CBCT images without any corrections. Another study that used U-net network architecture (with 37 HN patients for training), achieved similar accuracy (average MAE of 19 HU) and better to that of the CBCT images (MAE of 44 HU) [315]. However, no dosimetric study was performed to assess the dose-calculation feasibility.

Using unsupervised training, 4 studies on HN patients were found [82,310,316,317]. Barateau et al. [310] trained a GAN network using 2D slices from 30 paired datasets of HN patients and obtained sCT images with an average MAE of 82 HU which was significantly better compared with the other methods in the study (adapted HU-ED curve, DAM and DIR). While dose differences were not significant when compared to the other methods, mean gamma pass (2%/2 mm) rates of 98.1% were obtained by the DL method which were slightly lower than the ones obtained with the DIR method (98.8%). Similarly, Liang et al. generated sCT images using a cycle-consistent GAN (CycleGAN) framework (81 CBCT images for training) and obtained decreased MAE from 70 HU to 30 HU when compared with the original CBCT images [82]. Moreover, the gamma pass rates (1%/1mm) demonstrated higher accuracy of dose distributions calculated on sCT (96.26%) compared to CBCT images (86.92%). Additionally, a phantom study was conducted, to compare CycleGAN performance against two other unsupervised learning methods, namely deep convolutional generative adversarial network (DCGAN) and progressive growing of GAN (PGGAN), as well as a DIR algorithm. The results from similarity measures demonstrated the superiority of the CycleGAN method. Similarly, another study evaluated the performance of 3 different DL solutions (CycleGAN, Pix2pix and U-Net) based on supervised and unsupervised learning with a training database of 135 patients [316]. The image accuracy was quantified on a cohort of 34 patients based on MAE, RMSE, PSNR and SSIM and the dose calculation accuracy was assessed by comparing DVH and gamma passing rates. The results showed that all the sCT achieved better evaluation metrics than those of original CBCT, while the CycleGAN model was the best among the 3 methods (MAE of 24 HU and gamma rate of 97% for 2%/2mm criteria). Similar performance of a cycleGAN solution was demonstrated by Eckl et al. [317] on 3 localizations among which also the HN. The mean dosimetric differences of the target volumes were <1.7%. and the gamma pass rates >97.8% in all the cases.

To conclude, the limitations of HU-ED method is related to the CBCT image artefacts and patient scattering. While DAM methods may be used to compensate for this issue, the resulted sCT image will contain homogeneous tissues densities, and their definition depends on the accuracy of class segmentation. Ultimately DIR and DL methods seem to be the most attractive in terms of dose accuracy and computational time. Finally, all the CBCT-based dose calculation methods detailed above hold great importance for dose monitoring and treatment plan adaptation. Efforts remain needed for defining the thresholds for ART based on quantitative evaluation on daily CBCT images. A summary of studies proposing DL-based solutions for sCT generation from CBCT on HN localization can be followed in Table 1.10-3.

The last contribution of this thesis manuscript describes results of evaluating a DL solution trained with unpaired CBCT images from multiple centers and compare its image quality and dosimetric accuracy against previously described methods (adapted HU-ED curve, DAM, DIR).

Table 1.10-3 State-of-the-art DL-based solutions for sCT image generation from CBCT images on head-and-neck localization

Authors (year)	Type of network	Name of DL solution (Vendor)	Training cohort	Test cohort	Metrics for evaluation	Comments
Li et al. (2019) [314]	Supervised learning	2D U-Net neural network	60 CBCT/CT (50 +10)	N=10	MAE DVH Gamma analysis (1%/1mm)	MAE range: (6, 27) HU gamma>95%
Chen et al. (2020) [315]	Supervised learning	U-net	37 CBCT/CT (30+7)	N=7 patients	MAE	MAE=19 HU
Liang et al. (2019) [82]	Unsupervised learning	CycleGAN DCGAN PGGAN	90 CBCTs (81+9)	N=20	MAE, RMSE, SSIM, PSNR Gamma analysis (1%/1mm, 2%/2mm)	MAE=30 HU gamma=96%
Barateau et al. (2020) [310]	Unsupervised learning	GAN (ADMIRE-AI, Elekta AB)	30 CBCTs (2D slices used for training)	N=44 patients	MAE, ME DVH 3D Gamma analysis (2%/2 mm)	gamma=98.1%
Eckl et al. (2020) [317]	Unsupervised learning	CycleGAN	120 CBCTs (25 patients)	N=15	MAE, ME, Dice DVH 3D Gamma analysis (3%/3mm, 2%/2mm)	ME=1.4HU, MAE=77.2 HU Dose deviations <1.7% gamma(3%/3mm)>98.6% gamma(2%/2mm)>95%
Xue et al. (2021) [316]	Unsupervised learning	CycleGAN, Pix2pix U-Net	135 CBCTs	N=34 patients	MAE, RMSE, PSNR, SSIM DVH 2D Gamma (3%/3mm, 3%/2mm, 2%/2mm, 2%/3mm)	CycleGAN was the best

Abbreviations: MAE= mean absolute error, ME= mean error, PSNR = Peak signal to noise ratio, SSIM = structural similarity metrics, RMSE = root mean square error, DVH = dose volume histogram

1.11. Objectives of the study

In this clinical context, this thesis proposes to evaluate several automated solutions for different steps in the RT treatment workflow of HN cancer patients, that can enable the implementation of ART for this localization. The contributions of the work were divided into four axes:

1. Automatic treatment planning
 - We evaluated the quality of HN treatment plans when using an auto-planning solution vs manual VMAT and TomoTherapy treatment plans.
2. Automatic segmentation for OARs
 - We compared 4 atlas-based and 2 DL automatic segmentation solutions for the delineation of 10 OARs typically used in the treatment of HN cancers
 - We evaluated their performance with regards to resource demand, geometrical accuracy, the time needed for manual corrections and dosimetric impact on RT dose distributions calculated using auto-planning.
3. Automatic segmentation for CTVn
 - We compared the same 6 automatic segmentation solutions for the delineation of three lymph node levels (CTVn) in HN that usually are irradiated as secondary targets
 - We evaluated their performance with regards to resource demand, geometrical accuracy, the time needed for manual corrections and dosimetric impact on RT dose distributions calculated using auto-planning.
4. CBCT-based dose calculations for ART
 - We compared different methods for generating synthetic CT from CBCT images
 - We evaluated their potential application for ART in terms of dose calculation accuracy and image quality

An overview of the thesis objectives in function of the four axes is illustrated in Figure 1.11-1 together with their associated chapter in this manuscript.

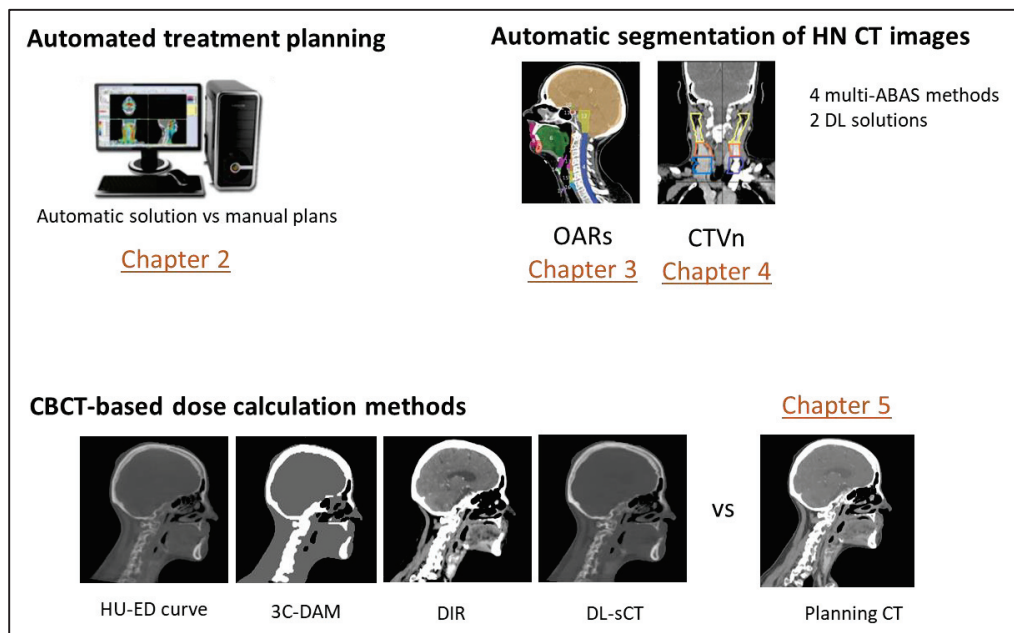


Figure 1.11-1 The objectives of the PhD study

Chapter 2. Evaluation of fully automated a priori MCO treatment planning in VMAT for head-and-neck cancer

The first investigation of adaptive methods for HN cancer treatment, was the evaluation of the performance of an a priori multicriteria plan optimization algorithm. The objective of the study was to investigate the research version of mCycle auto-planning solution (Elekta AB) against conventional manual planning using VMAT or Helical Tomotherapy (HT) for a cohort of HN patients. The results are presented in the article below (Biston et al. [2]) which has been published in the European Journal of Medical Physics in 2021. In this work, I contributed in the analysis of the results.

The clinical validation of the mCycle solution was done in several steps:

1. Plan quality index (PQI) calculated as a weighted sum of dose-volume objectives and constraints
2. Blind evaluation of manual vs automated plans done by 2 observers (low/high impact differences)
3. Deliverability of the plans including: number of control points (CP), number of MUs, modulation complexity score (MCS), and measured delivery times.



Contents lists available at ScienceDirect

Physica Medica

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

Original paper

Evaluation of fully automated a priori MCO treatment planning in VMAT for head-and-neck cancer



Marie-Claude Biston^{a,b,*}, Madalina Costea^{a,b}, Frédéric Gassa^a, Anne-Agathe Serre^a, Peter Voet^c, Randy Larson^c, Vincent Grégoire^a

^a Centre Léon Bérard, 28 rue Laennec 69373, LYON Cedex 08, France

^b CREATIS, CNRS UMR5220, Inserm U1044, INSA-Lyon, Université Lyon 1, Villeurbanne, France

^c Elekta AB, Stockholm, Sweden

ARTICLE INFO

Keywords:

Autoplanning

A priori MCO

VMAT

Head-and-neck cancer

ABSTRACT

Purpose: Automated planning techniques aim to reduce manual planning time and inter-operator variability without compromising the plan quality which is particularly challenging for head-and-neck (HN) cancer radiotherapy. The objective of this study was to evaluate the performance of an a priori-multicriteria plan optimization algorithm on a cohort of HN patients.

Methods: A total of 14 nasopharyngeal carcinoma (upper-HN) and 14 “middle-lower indications” (lower-HN) previously treated in our institution were enrolled in this study. Automatically generated plans (autoVMAT) were compared to manual VMAT or Helical Tomotherapy planning (manVMAT-HT) by assessing differences in dose delivered to targets and organs at risk (OARs), calculating plan quality indexes (PQIs) and performing blinded comparisons by clinicians. Quality control of the plans and measurements of the delivery times were also performed.

Results: For the 14 lower-HN patients, with equivalent planning target volume (PTV) dosimetric criteria and dose homogeneity, significant decrease in the mean doses to the oral cavity, esophagus, trachea and larynx were observed for autoVMAT compared to manVMAT-HT. Regarding the 14 upper-HN cases, the PTV coverage was generally significantly superior for autoVMAT which was also confirmed with higher calculated PQIs on PTVs for 13 out of 14 patients, whereas PQIs calculated on OARs were generally equivalent. Number of MUs and total delivery time were significantly higher for autoVMAT compared to manVMAT. All plans were considered clinically acceptable by clinicians.

Conclusions: Overall superiority of autoVMAT compared to manVMAT-HT plans was demonstrated for HN cancer. The obtained plans were operator-independent and required no post-optimization or manual intervention.

Introduction

Volumetric modulated arc therapy (VMAT) or Helical Tomotherapy (HT) have become the standard for head-and-neck (HN) cancer radiotherapy since, with these techniques, highly conformal dose distributions with steep dose gradients are obtained. HN treatment planning is reserved for experienced operators due to the complexity of planning target volume (PTV) coverage and homogeneity required while also sparing organs at risk (OARs). Even strict institution guidelines do not ensure that the lowest dose to the OARs has been achieved for each case. Hence, there is a need to automate the treatment planning optimization process.

Several studies reported the performance of autoplanning approaches implemented in treatment planning systems (TPS) [1]. One approach known as knowledge-based planning (KBP), commercialized as RapidPlan™ (Varian Medical Systems, Palo Alto, USA), uses available information from previously treated cases, and predicts the best achievable dose-volume histograms (DVHs) based on anatomical information of the current patient [2]. The algorithm generates constraints to assist optimization and produce plans similar to previous clinical cases. While working well for localizations with single PTVs, the results have higher variation for multiple targets [3]. On a cohort of 20 patients with nasopharyngeal carcinoma (NPC), Chang *et al* found that RapidPlan without manual adjustment could produce clinically acceptable plans

* Corresponding author at: Department of Radiation Oncology, Centre Léon Bérard, 28 rue Laennec, 69373 Lyon Cedex 08, France.

E-mail address: marie-claude.biston@lyon.unicancer.fr (M.-C. Biston).

<https://doi.org/10.1016/j.ejmp.2021.05.037>

Received 27 January 2021; Received in revised form 19 May 2021; Accepted 29 May 2021

Available online 8 June 2021

1120-1797/© 2021 Associazione Italiana di Fisica Medica. Published by Elsevier Ltd. All rights reserved.

for only 9 of the 20 patients. In another study knowledge-based treatment plans were comparable to clinical plans if the patient's OAR-planning target volume geometry was within the range of those included in the models [4].

Another approach is Pinnacle autoplanning (PAP) (Philips Medical Systems, Fitchburg, USA). Contrarily to the KBP approach, it requires no database of successful plans, but uses instead a template-based optimization tool that mimics the iterative optimization steps a skilled operator would undertake during manual planning [5]. PAP solution has been shown to yield good quality plans with respect to manual planning for several treatment sites, including HN cancer [5–8]. In one study (including 26 HN patients), PAP was able to produce clinically acceptable plans in all cases without manual intervention, with a score at least as high as manual plans in 94% of the evaluations [6].

Another solution is multicriterial optimization (MCO), as implemented in RayStation TPS (RaySearch Medical Laboratories AB, Stockholm, Sweden). The algorithm automatically generates Pareto-optimal plans according to typical DVHs constraints defined by the operator and stores them in a database for each patient. For each OAR constraint the system would require approximately 3–4 plans. The operator then explores the “Pareto surface” i.e. the different solutions obtained, and selects the solution which gives the best compromise between PTV coverage and OARs sparing [9]. Interestingly, MCO has been compared to PAP and KBP solutions for HN cancer [8]. All solutions were able to achieve all planning DVH constraints, with slightly better “parallel OARs” sparing obtained with PAP solution.

In this study we investigated the performance of auto-VMAT plans using mCycle, an *a priori*-MCO plan optimization algorithm implemented in a research version of Monaco TPS (version 5.59.11, Elekta AB, Stockholm, Sweden), which is currently not commercially available. While commercialized a posteriori MCO solutions upfront generate numerous Pareto-optimal plans and require the operator to select the best solution, a *a priori*-MCO algorithm directly and automatically generates a single Pareto-optimal plan for each new patient using a “wish-list”. This wish-list has predefined clinical dose objectives or constraints on OARs and PTVs based on the protocol. Each objective and constraint has a priority order, and the single treatment plan obtained features clinically favorable trade-offs between all treatment goals. Conversely to KBP that focuses on reproducing plan quality of previously treated patients, mCycle is expected to provide the best solution using one wish-list. The goal is then to have a single wish-list per clinical protocol providing consistent results over all the patients. This *a priori*-MCO approach was first developed and implemented in the Erasmus MC Cancer Center Institute in their iCycle software [10]. At that time, Erasmus-iCycle algorithm needed to be converted to Monaco plans for generating patient specific clinical plans. This *a priori*-MCO was evaluated for prostate, cervical and lung cancer [11–14], and for HN cancer for intensity modulated treatments (IMRT) [15]. In this study, automatically generated plans were preferred to plan manually generated in 97% of cases. In a recent multi-centric study involving 80 prostate cancer patients, the overall superiority of *a priori*-MCO compared to manual planning was demonstrated, resulting in a substantially reduced dose to the rectum with autoplanning while keeping the same delivered dose to PTV and bladder. Interestingly, important differences were found in some cases, pointing inconsistencies in manual planning.

In our study, mCycle (autoVMAT) compared to manual VMAT or HT planning (manVMAT-HT) was evaluated for the first time on 28 HN cancer patients treated with 2 PTV dose levels. The group was composed of 14 nasopharyngeal cancer (NPC) cases and 14 different “middle-lower” indications previously treated in VMAT or HT. Because the optic paths were closer to the PTVs for NPC patients, 2 different wish-lists were required to achieve the best quality plan for this cohort. We first assessed if mCycle plans were all clinically acceptable using these 2 wish-lists. Then the superiority of manual or mCycle plans was objectively evaluated using plan quality indexes (PQIs).

Materials and methods

Patients and clinical protocol

The data for this study consisted of 14 NPC and 14 “middle-lower indications” (larynx, oral cavity, oropharynx and hypopharynx) anonymized clinical plans from patients previously treated in our institution with VMAT or HT (TomoHD, Accuray, USA). VMAT treatments were delivered using VersaHD® (Elekta AB) Linac equipped with an Agility collimator with 5 mm leaves at isocenter. Among NPC cohort, half of the patients were treated with HT against only one patient of the second cohort. Target and OAR were contoured according to the international guidelines [16–18]. The PTVs were generated by adding 4 mm margins around the CTVs. The OARs under consideration are listed in Table 1. A 4 mm margin around the spinal cord and the brainstem, and a 3 mm margin around the optic nerves, chiasma and cochlea were added to define planning organ at risk volumes (PRV). A simultaneous integrated boost technique was used for delivering 70 Gy to the macroscopic tumor and 54.25 Gy to the nodes in 35 fractions.

Manual planning

Manual HT treatment plans were obtained with TomoHD™ TPS (V5.1.1.6, Accuray®, USA). HT plans were performed using a field size of 2.5 cm, a pitch of 0.43 to 0.3 and a modulation factor of 2.5 to 3, depending on the considered patients.

Manual VMAT treatment plans were performed using clinical Monaco TPS (V5.51, Elekta AB, Sweden). The optimization constraints were established on the basis of biological (serial/parallel) and physical cost functions. Values were adjusted from one patient to the other. In its clinical version, the TPS proposes 2 different planning modes: “constraints and Pareto”. In “constraint mode” priority is given to the OARs sparing versus PTV coverage. There is no ranking of the constraints to

Table 1

Dose-goal list and priorities/weighting factors for the different organs of interest;(*) The goals differed according to the macroscopic tumor level for the spinal cord and the brainstem, the lower HN having lower dose constraints compared to upper HN. The weighting factors (range 1–4) used to calculate the plan quality indexes (PQIs) were representative of the clinical dosimetric impact of the OAR. They were specific for each macroscopic tumor level (higher consideration for swallowing organs for lower HN cases and higher consideration for the spinal cord and the optic paths for upper HN cases).

Structure name	Goal	PQI weighting (OARs)	
		Lower HN	Upper NH
PTV 70 Gy	D _{95%} > 95% D _p		
	D _{90%} > 99% D _p		
	D _{7%} < 105% D _p		
	V _{95%} > 95% D _p		
PTV 54.25 Gy	D _{50%} = 70 Gy		
	D _{95%} > 95% D _p		
	D _{90%} > 90% D _p		
	D _{7%} < 105% D _p		
	V _{95%} > 95% D _p		
Parotids	D _{mean} < 25 Gy	4	3
PRV Cochlea	D _{mean} < 45 Gy	0	3
Oral Cavity	D _{mean} < 30 Gy	4	3
Constrictor Muscle	D _{mean} < 35 Gy	4	2
Submandibular glands	D _{mean} < 35 Gy	1	1
Esophagus	D _{mean} < 35 Gy	2	2
Trachea	D _{mean} < 35 Gy	2	2
Larynx	D _{mean} < 35 Gy	4	2
	D _{5%} < 55 Gy	4	2
PRV Spinal Cord	D _{2%} < 35 Gy/50 Gy*	1	4
PRV Brainstem	D _{2%} < 30 Gy/60 Gy*	1	4
PRV Optic Nerves	D _{2%} < 60 Gy	0	4
Mandible	D _{5%} < 70 Gy	2	1
PRV Chiasma	D _{2%} < 56 Gy	0	4

the OARs but a possibility to manually adjust weights on functions. However, this is not the philosophy of this TPS and this option is not used in our department. Hence, in the optimization stage, the PTV coverage will increase until dose-constraints have been reached to the adjacent OARs. If at the end of the segmentation stage the PTV coverage is insufficient, the Pareto mode can be turned on and can make it possible to gain few percentages of PTV coverage without “exploding” the dose constraints to the OARs. For manual VMAT plans, two 360° arcs were used.

mCycle autoplanning solution description

The mCycle solution is an integrated version of the Erasmus-iCycle methodology in Monaco TPSv5.59. This concept, using lexicographic MCO has been extensively described before [11,13,14,19]. To summarize, the first step is defining a “tumor and protocol” specific wish-list containing the protocol constraints and prioritized objectives. The wish-list is elaborated using the same optimization constraints as the one available in the clinical version of Monaco. However, the prescription template and the optimizer are different. The system can revisit an organ at a lower priority to reduce the dose in a multi-level approach for better balancing between OARs. The optimization stage has 2 passes. During the first pass, the TPS optimizes the individual objectives to meet the different requested goal values while respecting the priority order, and all defined constraints. If the goal value cannot be achieved, it is constrained at the achieved value and excluded from the second pass for further dose reduction. If the achieved value was below the goal dose, the goal dose is constrained, leaving optimization space for the lower prioritized objectives. The TPS optimizes all objectives in pass 1, and then moves to pass 2. During pass 2, the optimizer further reduces the objectives that were below the goal value in the first pass as low as possible. This is also possible to define sufficient goal values, which limits further dose reduction. Once the 2 passes are finished, the fluence map optimized plan will be segmented, followed by a segment weight and segment shape optimization.

The 2 wish-lists used for NPC and lower-HN cases are provided in [Supplementary data 1 and 2](#). The same arc input (two 360° arcs) and the same sequencing parameters as manual planning were used to perform the treatment plan.

A dosimetric contour was created for both manual and autoplanning for PTV54.25 Gy, which was a PTV54.25 Gy “ring” excluding PTV70Gy (-4 mm). Additional dosimetric contours were generated for the autoplanning: a structure in the neck which was an automatic posterior extension of the PRVspinal cord and, for NPC cases only, a PTV70Gy excluding all the PRVoptic paths (-3mm).

Dosimetric plan quality

Automated and manual plans were compared by assessing differences in PTV $V_{95\%}$, $D_{95\%}$, $D_{90\%}$, $D_{50\%}$, $D_{7\%}$, and HI, and D_{mean} , $D_{2\%}$ and/or $D_{5\%}$ for the OARs listed in [Table 1](#). Homogeneity index was calculated as $(D_{5\%}-D_{95\%})/D_{\text{mean}}$. The dosimetric goals were given to the planner by the physician before performing the clinical plans.

Two skilled physicians performed a blind dosimetric comparison between automatic and manual plans to assess clinical acceptability of all plans. This helped to determine the effectiveness of automated planning over manual planning.

We finally calculated a PQI to compare the achievement of the planning goals (in [Table 1](#)) in an operator independent manner. We

adopted the formalism proposed by Jorner et al. [20]. PQIs for PTVs and OARs were first considered independently and then a summation was performed to obtain a “global PQI” score for each patient, according to the formula below:

$$PQI(OARs) = \sum w^x \frac{D_{x\%,goal}(OAR) - D_{x\%,plan}(OAR)}{D_{x\%,goal}(OAR)}$$

$$PQI_{total} = PQI(PTV_{70Gy}) + PQI(PTV_{54.25Gy}) + PQI(OARs)$$

where D_x stands for the dose received by the $x\%$ of the volume of PTV or OAR, “plan” refers to the dose-volume indexes in the dose plan, “goal” refers to the dose objective and w refers to a weighting factor used as function of clinical relevance of the OAR ([Table 1](#)). For the lower-HN tumors, higher weight was given to the swallowing structures, while for the NPC HN cases, highest weights were attributed to the organs contained within the optic path. Higher values of PQI meant a better dose homogeneity for the PTVs and better OAR sparing, thus better overall plan quality.

Modulation complexity score and plan quality assurance

For VMAT plans only a modulation complexity score (MCS) was calculated from the segments details report (segment shape, area and weight) extracted from the treatment planning system following the formula introduced by Mc Niven et al. [21].

To verify deliverability of the plans, QA measurements were performed with the ArcCHECK device (Sun Nuclear Corporation) for both manual and automatic plans. This excluded patients having man-HT clinical plans. Hence a total of 26 and 14 QA measurements were performed for lower and upper HN cases, respectively.

Statistical analysis

For assessing the statistical significance ($p < 0.05$), all differences between manual and automatically generated plans were evaluated using paired two-sided Wilcoxon test performed in Jupyter notebook (Python version 3.8.3).

Results

For the 14 lower-HN patients, there was no significant difference between manVMAT-HT and mCycle regarding the metrics evaluating the dose distribution for the PTVs ([Table 2](#)). Furthermore, while no significant difference was observed in median PTV $\Delta V_{95\%}$ ($p \geq 0.39$) as shown in frequency histogram ([Fig. 1](#)), the mean doses to the oral cavity (24.4 Gy vs 30 Gy, $p < 0.001$), esophagus (8.6 Gy vs 14.2 Gy, $p < 0.001$), trachea (11.2 Gy vs 22 Gy, $p < 0.001$) and larynx (27.2 Gy vs 38.9 Gy, $p = 0.002$) were significantly higher for manVMAT-HT, meaning that, with equivalent PTV coverage, mCycle solution showed generally better OAR sparing.

In the 14 upper-HN cases, PTV70Gy $D_{95\%}$ (67.2 Gy vs 66.1 Gy, $p = 0.003$), $V_{95\%}$ (98% vs 94.9%, $p < 0.001$), HI (0.08 vs 0.10, $p = 0.008$) and PTV54.25 Gy $V_{95\%}$ (98.9% vs 98.2%, $p = 0.002$) were significantly better for mCycle versus manVMAT-HT, while no significant difference was observed on other PTV metrics ([Table 2](#)). Similarly, median PTV $\Delta V_{95\%}$ of both PTVs (-2.27% and -0.81%, for PTV70Gy and PTV54.25

$$PQI(PTV) = \sum \left(\frac{D_{95\%,plan}(PTV) - D_{95\%,goal}(PTV)}{D_{95\%,goal}(PTV)} + \frac{D_{90\%,plan}(PTV) - D_{90\%,goal}(PTV)}{D_{90\%,goal}(PTV)} + \frac{D_{7\%,goal}(PTV) - D_{7\%,plan}(PTV)}{D_{7\%,goal}(PTV)} + \frac{V_{95\%,plan}(PTV) - V_{95\%,goal}(PTV)}{V_{95\%,goal}(PTV)} \right)$$

Table 2

Plan parameters for manually generated plans (VMAT and HT plans grouped together), and automatic plans. Differences between manual and automatic plans are characterized by p-values. The homogeneity index (HI) was calculated according to the formula: (D95%-D5%)/Dmean.

Parameter	Lower HN – Average [range]			Upper NH - Average [range]				
	Automatic	Manual	p	Automatic	Manual	p		
PTV 70 Gy	D _{95%} [Gy]	67.8 [65.6,68.9]	67.6 [66,68.9]	0.09	67.2 [62.8,68.1]	66.1 [61.7,68.3]	0.003	
	D _{90%} [Gy]	68.8 [68.1,69.5]	68.4 [66.9,69.4]	0.14	68.1 [67.4,68.6]	68.1 [66.5,69.2]	0.76	
	D _{7%} [Gy]	72.2 [71.9,72.4]	72.2 [70.9,73.9]	0.86	71.9 [71.7,72.2]	72.4 [71.73,8]	0.12	
	V _{95%} [%]	97.9 [93.9,99.8]	97.6 [92.8,99.7]	0.39	98 [91.8,99.9]	94.9 [89.8,97.9]	p < 0.001	
	D _{50%} [Gy]	70.5 [70.1,70.8]	70.4 [69.3,71.5]	0.67	70 [69.9,70.2]	70.4 [69.5,71.5]	0.14	
PTV 54.25 Gy	HI	0.10 [0.05,0.30]	0.10 [0.05,0.27]	0.86	0.08 [0.05,0.18]	0.10 [0.05,0.23]	0.008	
	D _{95%} [Gy]	58.5 [53.5,61.2]	59.4 [52.4,65.2]	0.15	53.8 [52.7,59.2]	53.6 [51.8,60]	0.95	
	D _{90%} [Gy]	61.1 [54.2,64.9]	61.6 [53.7,67]	0.30	54.8 [53.3,62.1]	54.9 [52.3,64.8]	0.90	
	D _{7%} [Gy]	71.8 [70.6,72.2]	71.8 [70.6,73.7]	1.00	71 [69.7,71.7]	71.5 [69.4,73.3]	0.14	
	V _{95%} [%]	99.3 [98.4,99.9]	99.3 [96.7,100]	0.46	98.9 [98.3,99.7]	98.2 [96.6,99.3]	0.002	
Parotid_R	D _{50%} [Gy]	67.3 [56.5,70.2]	67.6 [58,71]	0.90	60.8 [55.8,69.4]	61.7 [53.9,70.9]	0.15	
	HI	0.51 [0.19,0.79]	0.52 [0.17,0.89]	0.50	0.45 [0.28,0.84]	0.46 [0.28,0.85]	1.00	
	D _{mean} [Gy]	25.7 [15.5,41.2]	26.4 [21,37.5]	0.43	31.8 [22.7,39.7]	30.1 [21.9,36.1]	0.04	
	Parotid_L	D _{mean} [Gy]	24.2 [17,34.8]	24.1 [19.9,30.3]	0.81	28.3 [23,40.1]	27.8 [23.2,30.4]	0.86
	PRV Cochlea_R	D _{mean} [Gy]	8.1 [1.1,41.6]	7.5 [1.3,25.6]	0.30	42.1 [28.5,70.4]	47 [33.3,70.4]	0.12
	PRV Cochlea_L	D _{mean} [Gy]	5.8 [1.1,29.3]	5.4 [1.3,24.5]	0.38	37.8 [13.8,50.9]	43.1 [25.7,54.7]	0.004
	Oral Cavity	D _{mean} [Gy]	24.4 [14.8,35.1]	30 [20.8,40.7]	p < 0.001	31.5 [24.5,43.1]	33.5 [25.3,42.7]	0.07
	Constrictor Muscle	D _{mean} [Gy]	41.8 [27.4,60.5]	44.2 [29.9,56.7]	0.10	51.9 [30.7,63.8]	54.8 [32.6,68.6]	0.02
	Submandibular_R	D _{mean} [Gy]	52.1 [44.3,62.6]	54.2 [46.7,65.3]	0.30	55.2 [41.8,69.8]	57.3 [44.3,69.8]	0.43
	Submandibular_L	D _{mean} [Gy]	51.9 [43,70.3]	54.6 [38.7,70.1]	0.22	53.1 [45.1,65.7]	55.4 [45.8,64.5]	0.19
	Esophagus	D _{mean} [Gy]	8.6 [2.5,15.2]	14.2 [5.6,22.6]	p < 0.001	13.3 [4.6,25]	17.1 [7.3,26.4]	0.010
	Trachea	D _{mean} [Gy]	11.2 [3.9,24]	22 [12.7,33.7]	p < 0.001	20.6 [5.3,27.3]	26.1 [10.1,40.3]	0.05
	Larynx	D _{mean} [Gy]	27.2 [16,61.2]	38.9 [27.7,63.5]	0.002	31.2 [16.7,50.1]	34.9 [25,42]	0.17
		D _{2%} [Gy]	52.3 [39,72]	58.5 [45.5,72.1]	0.06	51.7 [37,67.9]	52.3 [45.5,63.3]	0.95
	PRV Spinal cord	D _{2%} [Gy]	33.7 [12.2,38.5]	34.2 [29.5,37.7]	0.95	39.6 [34.1,42.9]	39.2 [32.5,55.2]	0.22
PRV Brainstem	D _{2%} [Gy]	21 [3,34.6]	22 [4.5,39.2]	0.33	58.3 [53.7,63.9]	53.3 [29.4,63.8]	0.001	
PRV Optic Nerve_R	D _{2%} [Gy]	1.9 [0.6,7.2]	2.6 [0.7,9.5]	0.08	48.8 [12.7,61.6]	47.2 [11.4,58.8]	0.03	
PRV Optic Nerve_L	D _{2%} [Gy]	2.7 [0.7,13.5]	3.9 [0.7,16.9]	0.008	47.1 [13.8,58.1]	45.3 [14.5,57.9]	0.06	
Mandible	D _{5%} [Gy]	55.1 [36.1,70.2]	52.4 [35.5,69]	0.11	56 [42,66.2]	58.6 [49.3,66.3]	0.27	
PRV Chiasma	D _{2%} [Gy]				39 [6.6,61.4]	41.7 [10.5,56.9]	0.58	

Gy, respectively) were also significantly different ($p = 0.002$) (Fig. 1). Significant advantage for mCycle was observed regarding the mean dose to the left cochlea (37.8 Gy vs 43.1 Gy, $p = 0.004$), constrictor muscle (51.9 Gy vs 54.8 Gy, $p = 0.02$) and esophagus (13.3 Gy vs 17.1 Gy, $p = 0.009$). Conversely, significantly lower mean dose to the right parotid (31.8 vs 30.1 Gy, $p = 0.04$) and D_{2%} to the brainstem (58.3 Gy vs 53.3 Gy, $p = 0.001$) and left optic nerve (48.8 vs 47.2 Gy, $p = 0.03$) was obtained for manVMAT-HT plans. Hence, with a significantly better target coverage, which can also explain higher D_{2%} to the brainstem for upper-HN, mCycle solution also better spared some OARs than manVMAT-HT, meaning that automatic plans were overall superior to manual ones.

Another important trend extracted from frequency histograms is that the differences between manVMAT-HT and mCycle in the mean doses to both parotids were globally symmetric around the Y axis (Fig. 1). Conversely, a large asymmetry in favor of mCycle was observed for the oral cavity and the larynx, meaning that, probably, the operator focused more its attention on sparing parotids than other organs.

After blind review, all plans were considered clinically acceptable by the 2 skilled physicians (Table 3). The physicians' score reflected the overall superiority of mCycle. For lower-HN cases mCycle was preferred in 11 and 14 of 14 patients for the first and second physician, respectively. PQIs for OARs were better for mCycle, for all the patients. Total PQIs were also found better for mCycle, with the exception of patients 7 and 9, due to a better scoring on PTVs for manual plans compared to mCycle (Fig. 2). Regarding patient 9, who also had the highest PQI score of the lower-HN cohort, the first physician considered that manual plan was superior to mCycle with high impact because of a better PTV70Gy coverage with the 95% isodose (98.8% vs 93.9%) whereas the second physician preferred mCycle solution because of too much overdose to the PTV54.25 Gy (D_{7%} = 61.5 Gy vs 57.5 Gy) and to the oral cavity (mean dose 25.9 Gy vs 16.6 Gy) on manual plan. Note that this "lack of PTV70Gy coverage" for mCycle compared to manVMAT-HT was mainly observed for larynx localizations because of the presence of air cavity

inside the PTV70Gy. Whereas manual planner tried to fill the air cavity at the price of a lower conformity, mCycle did not force the PTV coverage in the air. This is also why PTV70Gy coverage in mCycle solution was considered acceptable by the second physician. The same phenomenon also explains the better scoring obtained for manVMAT-HT for patient 7, which is also a larynx case but the 2 physicians agreed that, in this case, the PTV70Gy coverage was sufficient for both plans and chose mCycle solution because of better OARs sparing. Manual plan was also chosen by the first physician for patients 5 because mCycle was giving too much dose to the parotids, contrarily to the second physicist who considered that manual plan was under-covering PTV54.25 Gy. Finally, the first physician also preferred manual plan for patient 14 because of a slightly better PTV70Gy coverage compared to mCycle whereas the second physician preferred mCycle solution which provided better PTV54.25 Gy coverage and lower mean dose to the Larynx.

For upper-HN cases, total PQI scoring reflected the superiority of mCycle plans for all patients but patient 13, and showed that it relied on significantly higher PQIs for the PTVs, whereas few differences were observed on the PQIs for the OARs (Fig. 2). However, mCycle was preferred in only 10 and 9 of 14 patients for the first and second physician, respectively. Manual plan was preferred by the second physician for patient 13, in accordance to the total PQI result, and also for patients 1, 4, 5 and 6. The main reason for this choice was the balance between the maximum dose to the brainstem versus PTV70Gy coverage with the 95% isodose, which were both higher for mCycle. Choices were different for the first physician who preferred manual plans for patients 2, 7 and 8 which gave less dose to the parotids but to the price of lower PTV coverage. He also considered that mCycle plans were superior to manVMAT-HT with high impact for patient 3 (also having the largest difference in total PQI scoring between the 2 solutions) and for patient 14 where the PTV70Gy coverage with the 95% isodose is significantly higher for mCycle plan (97.7%) versus manVMAT-HT (93.9%).

Differences in the number of control points obtained for mCycle and manVMAT plans were not statistically significant whereas the number of

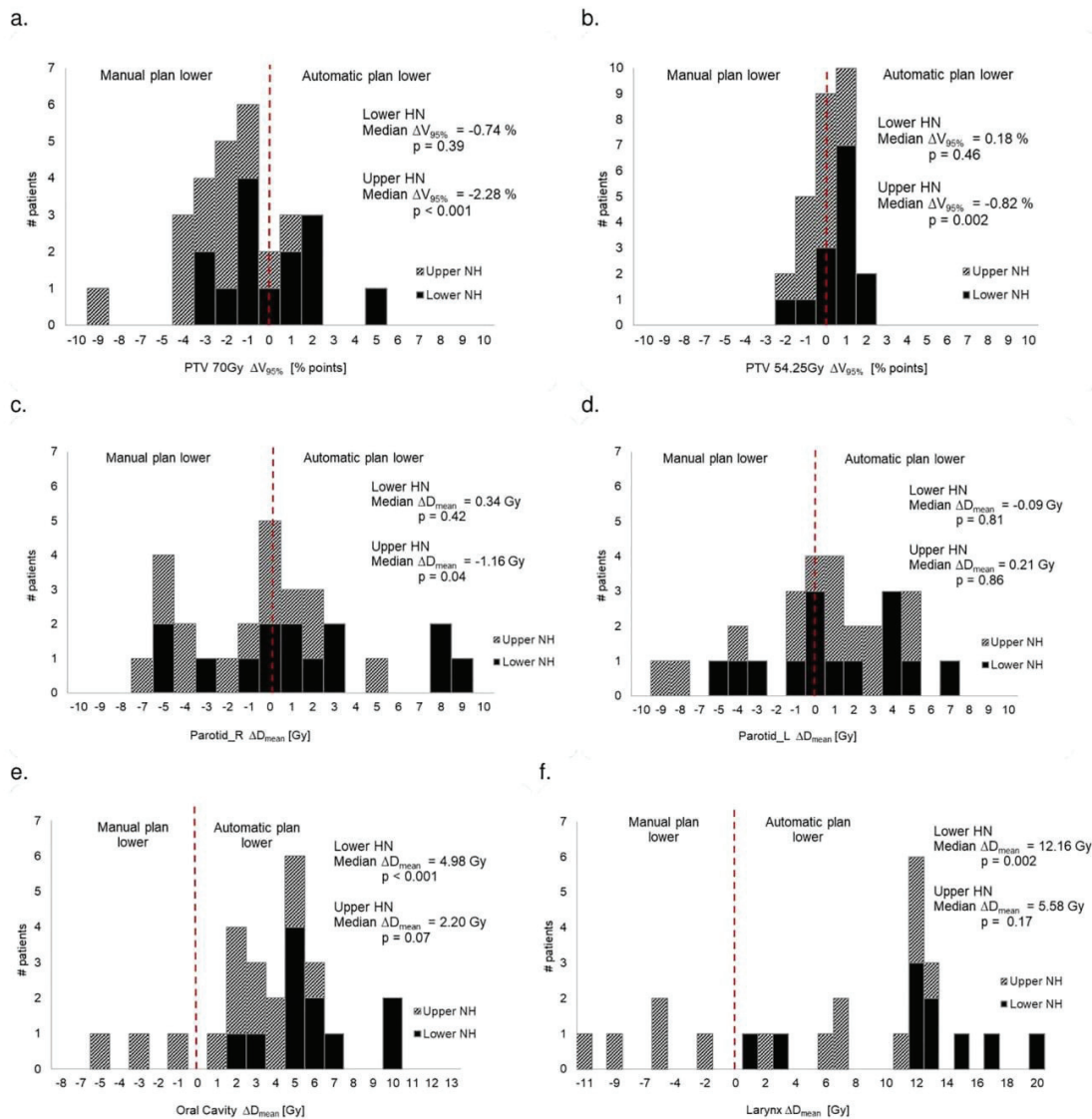


Fig. 1. Histograms showing the frequencies of observed differences between manual and automatic planning in $V_{95\%}$ for PTV70Gy (a) and PTV54.25Gy (b), parotids Dmean (c,d), oral cavity Dmean (e) and trachea Dmean (f).

MUs was significantly higher for mCycle plans whatever the HN localization ($p \leq 0.03$) (Table 4). The mean MCS decreased from 0.247 to 0.163 ($p < 0.001$) for lower-HN, and from 0.218 to 0.167 ($p = 0.13$) for upper-HN, for manVMAT versus mCycle, respectively. As a consequence, the mean delivery time was also significantly higher (mean 246 s versus 200 s) for mCycle plans independent of the HN location. QA measurements with the ArcCheck device showed a minimum pass rate of 98% versus 95.7% for lower-HN cases and 96.5% versus 97.5% for upper-HN cases, for mCycle and manVMAT plans respectively. Acceptance criteria were: 95% pass for 3%-3mm using a global analysis method.

Discussion

This study presented a comparison between manVMAT-HT and mCycle autoplanning solution on 28 HN cancer patients. All plans were

considered clinically acceptable, and both the physicians' scores and the PQIs calculations reflected the overall superiority of mCycle. Notably mCycle plans were considered better than manVMAT-HT plans by the physicians in 75% of cases. They generally agreed that, for lower-HN, mCycle better spared the OARs without compromising the PTV coverage. The presence of an air cavity in the target volume could lead to lower volume coverage with the 95% isodose for mCycle compared to manual planning. Indeed, contrarily to the manual planner, mCycle did not enforce PTV coverage in that region, to the detriment of dose conformity. This phenomenon was mainly observed for larynx localizations and explained why one physician preferred the clinical plan in one case. Note that this observation was found during plan scoring and that if the trade-off is preferred to be different and more in line with the manual plan, the wish-list could be adapted accordingly. The second physician was very sensitive to the dose homogeneity into the PTV54.25 Gy and attentive to the over-sparing of the parotids vs PTV54.25 Gy coverage.

Table 3
Blinded side-by-side comparisons of manually generated VMAT plans with automatically generated plans.

		Lower HN		Upper HN	
		Physician 1	Physician 2	Physician 1	Physician 2
Automatic plan better	- impact low	10	13	8	8
	- impact high	1		2	
Equal plan quality			1		1
Manual plan better	- impact low	2		4	5
	- impact high	1			
Total number of plans		14	14	14	14

Interestingly, the mean dose to the parotids were, in general, similar or better for manual vs automatic plans, reflecting that increased attention is given to this organ when creating manual plans. Conversely, this is difficult to find the best compromise between conformity for PTV5.4.25

Gy and sparing of the esophagus, trachea and larynx, and mCycle performed better on this aspect.

For upper-HN cases, while total PQI scoring reflected the superiority of mCycle plans compared to manVMAT-HT, medical point of views were much more divergent. Indeed 4 to 5 manVMAT-HT cases were preferred to mCycle by each physician. These cases were all different, thus reflecting a difference in sensitivity to the dosimetric parameters for both physicians. Upper-HN deviations appeared first because of the tradeoff between providing superior PTV coverage (mCycle) and higher dose gradients around the optic paths. Conversely the manual plans provided non-optimal PTV coverage but lower dose in optic paths. Notably, if the patient had been operated upon, physicians were more cautious not to take risks on the optic pathways and preferred the manual plan. Hence one more upper-HN wish-list with minor adjustments in the maximum dose to the optic paths may have been better adapted to patients previously operated upon. This is why explicit input from the protocol is needed for the wish-list. These results also show the limits of using PQIs, to objectively judge plan quality. The use of score card instead of PQIs, as available in Pinnacle TPS, would certainly have been more relevant to underline the differences in terms of plan quality on the doses delivered to the optic paths, which were more problematic for upper-HN case, as in some cases, the dose gradients was a determining factor for choosing the manual plans [22].

The results obtained are consistent with most dosimetric

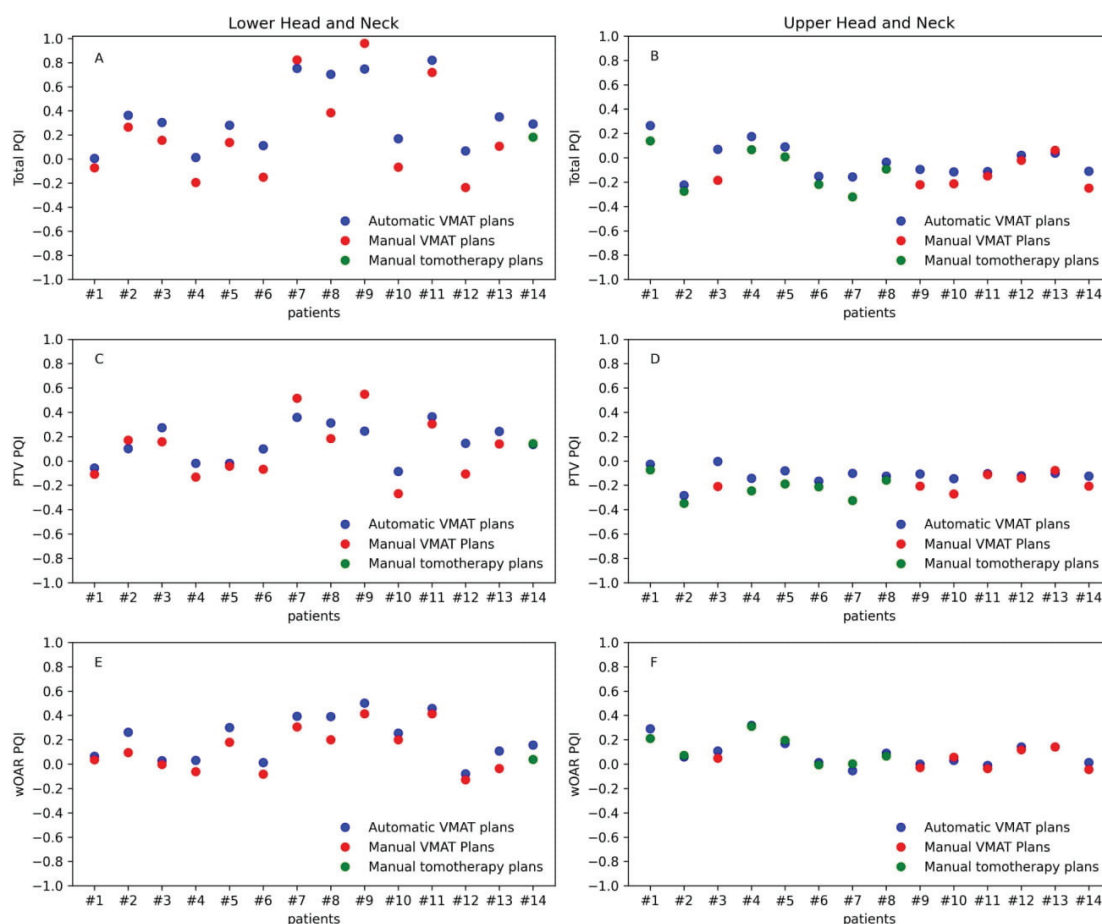


Fig. 2. Plan Quality Indexes (PQI) calculated for each patient of the 2 cohorts. Total plan PQI are represented on panel A and B for lower and upper HN cases, respectively; PQI calculated on PTVs are represented on panel C and D for lower and upper HN cases, respectively; PQI calculated on OARs are represented on panel E and F for lower and upper HN cases, respectively.

Table 4

Number of control points (CP), number of monitor units (MUs), modulation complexity score (MCS), and measured delivery times for manually and automatically generated plans. Differences between manual and automatic plans were characterized by p-values.

	Lower head and neck				Upper head and neck			
	Mean CP [range]	Mean MU [range]	Mean MCS [range]	Mean delivery time(s) [range]	Mean CP [range]	Mean MU [range]	Mean MCS [range]	Mean delivery time (s) [range]
Automatic	212 [197–228]	1274 [1135–1482]	0.163 [0.13–0.19]	244 [220–270]	215 [193–229]	1328 [1198–1413]	0.167 [0.15–0.18]	251 [240–275]
Manual	222 [168–291]	867 [655–1407]	0.247 [0.15–0.29]	192 [165–250]	244 [210–277]	991 [896–1100]	0.233 [0.21–0.26]	208 [190–230]
p	0.45	p < 0.001	p < 0.001	0.005	0.06	0.03	0.03	0.035

comparisons between manual planning and autoplanning for HN cancer reported with other commercial solutions [3,4,6,8,22–25,25,26]. In general, autoplanning was able to provide, at least, plans with the same quality as manual planners with, however, considerable differences between solutions regarding manual interventions. In *Chang et al* manual touch-up of the KBP solutions enabled to increase from 9 to 19 out of 20 the number of clinically acceptable plans [3]. Manual intervention is also necessary to select the best MCO solution proposed on the “Pareto surface”. Similarly to this study, with no manual intervention, PAP provided clinically acceptable solutions for all plans, with either a score at least as high as manual plans for 94% of the plans [26], or a preference for the autoplanning (80% vs 20% of the plans) [24].

The complexity of the plans was increased with mCycle compared to manVMAT but QA results were not impacted whereas the delivery time was significantly increased. The increased modulation using autoplanning algorithms was also reported for prostate cancer with Erasmus-iCycle, and with the PAP solution [12,27].

The gain in the treatment planning time with autoplanning compared to manual planning was also highlighted in some studies. This was generally >60% and even reaching 80% [3,24,25,28]. This was also very different between autoplanning solutions. In one study comparing PAP, KBP and MCO solutions, higher effective working time (which excludes targets and OARs contouring, and optimization time) was reported for the MCO solutions (116 min) than for other solutions (range 1–15 min), as well as higher optimization time (218 min for MCO vs 83 min for PAP and 30 min for KBP) [8].

In Monaco TPS, the effective working time is similar to KBP and PAP solutions for both automatic and manual planning (i.e. 5 min). The optimization time for manual planning is at least 30 min to obtain one solution for a HN case. Most of the time 3–4 optimizations are required. This means that at least 2 h are needed for an acceptable plan. This time can be double when dealing with more complex upper-HN cases. With mCycle, once the wish-lists were established, the optimization time was 50–60 min for all HN cases. This was substantially less than PAP and MCO solutions, but slower than KBP (28 min) [8]. Thus taking into account all the planning stages, the gain in time with autoplanning compared to manual planning was similar to other solutions.

The mCycle algorithm requires an intensive iterative tuning process to obtain a consistent wish-list. The time needed to build a robust wishlist depends on how explicit a protocol is defined. The more detailed the protocol the easier mCycle will deliver expected results. To be expert in Monaco planning is an undeniable plus since a good starting point for the wish-list is to use a prescription template issued from manual planning and to discuss the priority order of the different functions and dose objectives with the clinicians. The first aim is to mimic the clinical plans. An evaluation of the current wish-list is made after applying it on a restricted number of patients (typically 4–5) per protocol. The wish-list is then updated until an optimal solution is obtained for all cases. At this step, the goal is to improve the clinical plans. The process is stopped when it is considered that further improvements are impossible. Then it is applied on a larger cohort (10 patients). If no failure is observed on that cohort the wish-list is validated.

In this study, to obtain plans of the same quality as clinical plans few

wish-list iterations were required [3–4] for lower-HN cases. We performed additional 2–3 iterations to better balance the dose delivery between oral cavity and parotids and to improve again the dose conformity for lower-HN. This final wish-list was then applied to all the cohort of the study. For upper-HN, we started from the lower-HN wish-list, increased the maximum dose constraints tolerance to the brainstem and added strict maximum dose constraints to the chiasma and optic nerves. Initial iterations were performed on more cases (5–8 patients) because failure to respect strict maximum dose constraints to all the critical organs were observed.

Hence this might take time to have a robust wish-list but once validated for one clinical protocol and for a localization, it can be very quickly adapted to another dose protocol provided that dose constraints to the OARs do not change drastically (i.e. HN localization with 3 doses levels...). Similarly, the time spent to elaborate a wish-list can be drastically reduced for localizations with small number of critical OARs with clearly established dose constraints (i.e. breast).

Another aspect of autoplanning solutions is the user’s impact on the quality of treatment plans. Even if during the optimization there is no or little manual intervention for all the algorithms to obtain a robust clinical plan, the degree of expertise of the planner will impact the quality of the results. Hence, similarly to us, the quality of plans generated by PAP has been reported to be dependent on the input from experience treatment planners to set the initial user settings and define good clinical protocols [22]. For KBP solution, an apparent limitation is that the models can only be as good as the training data that has been input in the first instance. This means that the plans can be clinically acceptable but not the optimal plans. RapidPlan attempts to get around this by always placing the optimization objectives for OARs lower than the predicted DVH such that it always tries to improve on the prediction [1].

To conclude, this study demonstrated overall superiority of mCycle compared to manVMAT-HT plans for HN cancer. This was observed through the calculation of PQIs and blind clinical evaluation performed by 2 skilled physicians. No manual adjustment of the 2 wish-lists was required to obtain robust clinical plans although the cohorts were very heterogeneous in terms of anatomy, volume and tumor location. Finally, another advantage of mCycle is that it does not stop at the set objective goals but always minimizes dose to OARs and unspecified tissues. Further investigations are in progress in our hospital to establish wish-lists for other treatment sites and protocols.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: This work was performed in the framework of a research cooperation agreement with Elekta AB.

Acknowledgments

Elekta AB is acknowledged for having involved the CLB team in this research project. Two co-authors (PV and RL) are employed by Elekta

and had an advisory role in the study design. They were also partly involved in data collection and analysis. They also contributed in the writing of the manuscript. This work was performed within the framework of the SIRIC LYriCAN Grant INCa-INSERM-DGOS-12563, and the LABEX PRIMES(ANR-11-LABX-0063) of Université de Lyon, within the program Investissements d'Avenir (ANR-11-IDEX-0007) operated by the ANR.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.ejmp.2021.05.037>.

References

- [1] Hussein M, Heijmen BJM, Verellen D, Nisbet A. Automation in intensity modulated radiotherapy treatment planning—a review of recent innovations. *Br J Radiol* 2018; 91(1092):20180270. <https://doi.org/10.1259/bjr.20180270>.
- [2] Zhang J, Ge Y, Sheng Y, Yin F-F, Wu QJ. Modeling of multiple planning target volumes for head and neck treatments in knowledge-based treatment planning. *Med Phys* 2019;46(9):3812–22.
- [3] Chang ATY, Hung AWM, Cheung FWK, Lee MCH, Chan OSH, Philips H, et al. Comparison of planning quality and efficiency between conventional and knowledge-based algorithms in nasopharyngeal cancer patients using intensity modulated radiation therapy. *Int J Radiat Oncol Biol Phys* 2016;95(3):981–90.
- [4] Tol JP, Delaney AR, Dahele M, Slotman BJ, Verbakel WFAR. Evaluation of a knowledge-based planning solution for head and neck cancer. *Int J Radiat Oncol Biol Phys* 2015;91(3):612–20.
- [5] Janssen TM, Kusters M, Wang Y, Wortel G, Monshouwer R, Damen E, et al. Independent knowledge-based treatment planning QA to audit Pinnacle autoplanning. *Radiother Oncol* 2019;133:198–204.
- [6] Hazell I, Bzdusek K, Kumar P, Hansen CR, Bertelsen A, Eriksen JG, et al. Automatic planning of head and neck treatment plans. *J Appl Clin Med Phys* 2016;17(1): 272–82.
- [7] Wang J, Chen Z, Li W, Qian W, Wang X, Hu W. A new strategy for volumetric-modulated arc therapy planning using AutoPlanning based multicriteria optimization for nasopharyngeal carcinoma. *Radiat Oncol* 2018;13:94.
- [8] Krayenbuehl J, Zamburlini M, Ghandour S, Pachoud M, Tanadini-Lang S, Tol J, et al. Planning comparison of five automated treatment planning solutions for locally advanced head and neck cancer. *Radiat Oncol* 2018;13(1). <https://doi.org/10.1186/s13014-018-1113-z>.
- [9] Craft DL, Hong TS, Shih HA, Bortfeld TR. Improved planning time and plan quality through multicriteria optimization for intensity-modulated radiotherapy. *Int J Radiat Oncol Biol Phys* 2012;82(1):e83–90.
- [10] Voet PWJ, Breedveld S, Dirkx MLP, Levendag PC, Heijmen BJM. Integrated multicriterial optimization of beam angles and intensity profiles for coplanar and noncoplanar head and neck IMRT and implications for VMAT. *Med Phys* 2012;39(8):4858–65.
- [11] Voet PWJ, Dirkx MLP, Breedveld S, Al-Mamgani A, Incrocci L, Heijmen BJM. Fully automated volumetric modulated arc therapy plan generation for prostate cancer patients. *Int J Radiat Oncol Biol Phys* 2014;88(5):1175–9.
- [12] Heijmen B, Voet P, Fransen D, Penninkhof J, Milder M, Akhlat H, et al. Fully automated, multi-criterial planning for Volumetric Modulated Arc Therapy - An international multi-center validation for prostate cancer. *Radiother Oncol* 2018; 128(2):343–8.
- [13] Sharfo AWM, Voet PWJ, Breedveld S, Mens JWM, Hoogeman MS, Heijmen BJM. Comparison of VMAT and IMRT strategies for cervical cancer patients using automated planning. *Radiother Oncol* 2015;114(3):395–401.
- [14] Della Gala G, Dirkx MLP, Hoekstra N, Fransen D, Lanconelli N, van de Pol M, et al. Fully automated VMAT treatment planning for advanced-stage NSCLC patients. *Vollautomatische VMAT-Behandlungsplanung für Patienten mit fortgeschrittenem NSCLC. Strahlenther Onkol* 2017;193(5):402–9.
- [15] Voet PWJ, Dirkx MLP, Breedveld S, Fransen D, Levendag PC, Heijmen BJM. Toward fully automated multicriterial plan generation: a prospective clinical study. *Int J Radiat Oncol Biol Phys* 2013;85(3):866–72.
- [16] Grégoire V, Evans M, Le Q-T, Bourhis J, Budach V, Chen A, et al. Delineation of the primary tumour Clinical Target Volumes (CTV-P) in laryngeal, hypopharyngeal, oropharyngeal and oral cavity squamous cell carcinoma: AIRO, CACA, DAHANCA, EORTC, GEORCC, GORTEC, HKNPCSG, HNCIG, IAG-KHT, LPRHHT, NCIC CTG, NCRI, NRG Oncology, PHNS, SBRT, SOMERA, SRO, SSHNO, TROG consensus guidelines. *Radiother Oncol* 2018;126(1):3–24.
- [17] Grégoire V, Ang K, Budach W, Grau C, Hamoir M, Langendijk JA, et al. Delineation of the neck node levels for head and neck tumors: a 2013 update. *DAHANCA, EORTC, HKNPCSG, NCIC CTG, NCRI, RTOG, TROG consensus guidelines. Radiother Oncol* 2014;110:172–81.
- [18] Brouwer CL, Steenbakkers RJHM, Bourhis J, Budach W, Grau C, Grégoire V, et al. CT-based delineation of organs at risk in the head and neck region: DAHANCA, EORTC, GORTEC, HKNPCSG, NCIC CTG, NCRI, NRG Oncology and TROG consensus guidelines. *Radiother Oncol* 2015;117(1):83–90.
- [19] Breedveld S, Storch PRM, Voet PWJ, Heijmen BJM. iCycle: integrated, multicriterial beam angle, and profile optimization for generation of coplanar and noncoplanar IMRT plans. *Med Phys* 2012;39(2):951–63.
- [20] Jorret N, Carrasco P, Beltrán M, Calvo JF, Escudé L, Hernández V, et al. Multicentre validation of IMRT pre-treatment verification: comparison of in-house and external audit. *Radiother Oncol* 2014;112(3):381–8.
- [21] McNiven AL, Sharpe MB, Purdie TG. A new metric for assessing IMRT modulation complexity and plan deliverability. *Med Phys* 2010;37(2):505–15.
- [22] Speer S, Klein A, Kober L, Weiss A, Yohannes I, Bert C. Automation of radiation treatment planning: Evaluation of head and neck cancer patient plans created by the Pinnacle (3) scripting and Auto-Planning functions. *Automatisierte Bestrahlungsplanung: Auswertung von mit Pinnacle3 via Scripting und Auto-Planning erzeugten Bestrahlungsplänen von Patienten mit Kopf-Hals-Tumor. Strahlenther Onkol* 2017;193(8):656–65.
- [23] Fogliata A, Reggiori G, Stravato A, Lobefalo F, Franzese C, Franceschini D, et al. RapidPlan head and neck model: the objectives and possible clinical benefit. *Radiat Oncol* 2017;12(1). <https://doi.org/10.1186/s13014-017-0808-x>.
- [24] Krayenbuehl J, Norton I, Studer G, Guckenberger M. Evaluation of an automated knowledge based treatment planning system for head and neck. *Radiat Oncol* 2015;10:226.
- [25] Kierkels RG, Visser R, Bijl HP, Langendijk JA, van 't Veld AA, Steenbakkers RJ, et al. Multicriteria optimization enables less experienced planners to efficiently produce high quality treatment plans in head and neck cancer radiotherapy. *Radiat Oncol* 2015;10:87.
- [26] Hansen CR, Bertelsen A, Hazell I, Zukauskaitė R, Gyldenkerne N, Johansen J, et al. Automatic treatment planning improves the clinical quality of head and neck cancer treatment plans. *Clin Transl Radiat Oncol* 2016;1:2–8.
- [27] Cilla S, Ianiro A, Romano C, Deodato F, Macchia G, Buwenge M, et al. Template-based automation of treatment planning in advanced radiotherapy: a comprehensive dosimetric and clinical evaluation. *Sci Rep* 2020;10(1). <https://doi.org/10.1038/s41598-019-56966-y>.
- [28] Kusters JMAM, Bzdusek K, Kumar P, van Kollenburg PGM, Kunze-Busch MC, Wendling M, et al. Automated IMRT planning in Pinnacle: a study in head-and-neck cancer. *Automatisierte IMRT-Planung mit Pinnacle: eine Studie zu Kopf-Hals-Tumoren. Strahlenther Onkol* 2017;193(12):1031–8.

Synthesis

In this study, the performance of the mCycle auto-planning solution was evaluated against manually optimized VMAT or HT plans. The comparison was performed based on PQIs calculations, blinded evaluation by 2 skilled physicians, number of CP, number of MUs, MCS, and QA measurements. Based on a cohort of 14 nasopharyngeal carcinomas (upper-HN) and 14 “middle-lower indications” (lower-HN), the superiority of mCycle solution was demonstrated. Moreover, mCycle plans were considered better than manual plans in 75% of the cases. This result is clinically meaningful because manual optimization of a complex HN case requires at least 3-4 optimizations and takes considerably longer time (>2h) when compared to the automatic solution proposed (<1h). Moreover, a user-free solution enables increased consistency among planners. Another big advantage of using mCycle solution was that it allowed better sparing of the OARs while maintaining the desired coverage to the PTVs. This was most of the time the reason why it was preferred when compared to manual plans. As a consequence, mCycle plans had higher complexity compared to manual VMAT plans, which significantly increased the delivery time, however without any negative impact on the QA measurements.

In my opinion, this evidence motivates the clinical integration of auto-planning solutions for complex cases such as HN, where better OARs sparing can be achieved by automatic iterations than by a human operator. Nevertheless, if further improvements of the plan are wanted, the plan proposed by mCycle may be a good starting point for additional adjustments. It must be mentioned that an excessive amount of time is required to have a robust wish-list and that the relatively long computational time remains a limitation for ART. To get a robust wish-list, an intensive iterative tuning process must be made that involves a collaborative work between the team of clinicians for deciding the priority order of the different functions and dose objectives. The complexity of the task increases with the number of critical OARs considered, and few iterations are required in order to well balance the dose objectives and constraints. Nevertheless, once validated for a clinical protocol and for a localization, it can be quickly adapted to another dose protocol having similar dose constraints (i.e. HN localization with 3 PTV dose levels).

As future perspectives, Elekta company is working on providing a faster auto-planning solution (<15min for HN planning) which will open doors for the ART for HN cancer patients. However, manual delineation of organs on the anatomy of the day is still a limitation that challenges the promising results of AS solutions. This topic will be further discussed in the next two chapters of the thesis.

Chapter 3. Comparison of atlas-based and deep learning methods for organs at risk delineation on head-and-neck CT images using an automated treatment planning system

This chapter represents the work of an article that has been published in the Radiotherapy and Oncology Journal in November 2022 [3]. The objective was to evaluate and compare the performances of different AS methods for OARs segmentation on HN CT images.

3.1. Abstract

Background and purpose: To investigate the performance of head-and-neck (HN) organs-at-risk (OAR) automatic segmentation (AS) using four atlas-based (ABAS) and two deep learning (DL) solutions.

Material and Methods: All patients underwent iodine contrast-enhanced planning CT. Fourteen OAR were manually delineated. DL.1 and DL.2 solutions were trained with 63 mono-centric patients and >1000 multi-centric patients, respectively. Ten and 15 patients with varied anatomies were selected for the atlas library and for testing, respectively. The evaluation was based on geometric indices (DICE coefficient and 95th percentile-Hausdorff Distance (HD_{95%}), time needed for manual corrections and clinical dosimetric endpoints obtained using automated treatment planning.

Results: Both DICE and HD95% results indicated that DL algorithms generally performed better compared with ABAS algorithms for automatic segmentation of HN OAR. However, the hybrid-ABAS (ABAS.3) algorithm sometimes provided the highest agreement to the reference contours compared with the 2 DL. Compared with DL.2 and ABAS.3, DL.1 contours were the fastest to correct. For the 3 solutions, the differences in dose distributions obtained using AS contours and AS+manually corrected contours were not statistically significant. High dose differences could be observed when OAR contours were at short distances to the targets. However, this was not always interrelated.

Conclusion: DL methods generally showed higher delineation accuracy compared with ABAS methods for AS segmentation of HN OAR. Most ABAS contours had high conformity to the reference but were more time consuming than DL algorithms, especially when considering the computing time and the time spent on manual corrections.

3.2. Introduction

Manual contouring of organs-at-risk (OAR) is a time-consuming task that suffers from large intra- and inter-observer variations (IOV), especially for HN cancer patients, because of the complex anatomy and the number of OARs [26,127,128,318]. Contour variations may also result in important dosimetric differences [142]. Therefore, automatic segmentation (AS) methods are strongly sought after to increase contouring accuracy, improve the inter-observer variability, reduce delineation time, and facilitate treatment plan adaptation [132,133].

Among the different methods, atlas-based segmentation (ABAS) uses one or more representative patients with carefully delineated OAR as reference atlas library for contouring new patients [152]. Those methods are widely spread because they require minimum of resources, but they do have several drawbacks: atlas selection strategy (single vs multi-atlas) [152]; performance plateau reached after 10-20 atlases [153]; poor performance for small and low contrast soft tissue structures [154]; increased computational time with each added atlas [155].

Data from multiple atlases (multi-ABAS) can be combined with the help of a fusion algorithm in order to reduce the risk of anatomical variability between the atlas and the new patient [156]. Additionally, hybrid approaches are developed to combine multi-ABAS with machine learning features [169,170,172–

174]. Despite a higher computational time, multi-ABAS studies have consistently demonstrated improved conformity to the reference contours over the single atlas methods, with consequent reduction of the post-editing time [157,319]. By adding image intensity information, other studies have shown improved accuracy for model-based methods particularly on large organs such as brainstem and spinal cord but lacking precision for tiny structures like cochlea [133,169,172,173].

Another method issued from AI research and challenging ABAS is the use of deep learning (DL) techniques [132–134,143,163,175,191,320]. DL contouring typically implies the training of a convolutional neural network (CNN) directly from a set of annotated reference data. Although the training phase requires excessive GPU computing power and work in data gathering and curation, once trained, the segmentation is very fast. Different network architectures are continuously investigated to reach the best predictions for multiple organ segmentation. While some models are accurate on most volumes, they may have difficulties in segmenting small volumes such as optical nerves or cochlea, or organs with low image contrast such as constrictor muscles. Comparison between different DL models is rather difficult due to differences in the data sets used. From the few studies analyzing the performance of different DL models trained and tested on the same data sets, Chen et al. examined one multi-ABAS and three similar DL models following U-Net-like network architectures with distinctive differences in the configuration and loss functions [191]. While nnU-net [187] is a self-configuring network based on the training dataset, AnatomyNet [186] follows a defined scheme with squeeze-and-excitation residual blocks for better feature representation and a combination of two loss functions (DICE and Focal Loss). By using Ua-Net [192] for the HN model, that first performs an OAR detection module and then considers image features only within the detected regions, WBNet was superior to the other methods for most organs. Apart from the in-house developed models, several commercially available solutions have reported good agreement with physicians' manual contours and considerable time savings on the delineation task [143,163,206,217,218].

Most studies showed that DL methods outperformed ABAS methods [143,163,191]. However, there is still room for improvement in the AS of computed tomography (CT) images for small organs or with limited image contrast such as optic nerves, optic chiasm or cochlea [134,154,191,218]. Generally, AS methods comparisons are based on geometric indices calculations only (DICE; HD) to compare the volume overlap between the reference and the automatically generated contour [135]. However, it is highly recommended to perform additionally a dosimetric evaluation by generating treatment plans with the AS contours [131,133,144,145]. Nevertheless, this involves excessive time in generating treatment plans, and may also introduce inter or intra-planner variability [321,322].

In this context, the objective of the present study was to evaluate the performances of 4 atlas-based algorithms and 2 DL solutions for the AS of 14 HN OAR. Three multi-ABAS algorithms and one DL solution are commercially available while one hybrid-ABAS algorithm and one center-specific DL solution were investigated for the first time on HN CT images. All six solutions were evaluated based on geometrical accuracy and computational time. The time spent for correcting the contours was measured for the most accurate three AS methods and an auto-planning solution based on a priori multicriteria optimization (MCO) algorithm was used for the first time to derive doses from AS contours with and without manual correction.

3.3. Materials and methods

Patient data

Seventy-eight non-operated HN cancer patients treated with radiation therapy between 2018 and 2021 and who underwent iodine contrast-enhanced planning CT, were selected for this study, which was approved by the hospital ethics committee. The contrast agent protocol followed a 2-phase administration of 90mL iodine solution of 2mL/s with 45s pause in between 2 doses of 45mL. The CT scan acquisition was

done 10s after the second injection. Fourteen OAR (i.e. parotids, submandibular glands, oral cavity, constrictor muscle, larynx, esophagus, trachea, thyroid, eyes, optical nerves, cochlea, brainstem, spinal cord, mandible) were manually delineated by a single expert physician (>30 years of experience), on 512x512 and 2mm-thick CT slices following HN delineation guidelines [26]. An overview of the study design is provided in Figure 3.3-1. For the multi-ABAS approach, 10 patients from this database were selected based on their body mass index (BMI) (from 18.9 to 30.7) to form a heterogeneous library of atlases with various representative patient anatomies. The same 10 atlases were used to create a library in MIM-Maestro (MIM Software; Cleveland, USA) and in research version of the ADMIRE software (ADMIREv3.41, Elekta AB; Stockholm, Sweden). A mono-centric DL.1 model was trained using 63 patients with the same set of OAR excluding optical nerves and cochlea. Conversely, DL.2 model was trained on a large database of patients (>1000) collected from multiple centers including ours. Fifteen patients having a BMI ranging from 12.1 to 34.7 were reserved for the testing phase. Characteristics of the test cohort are detailed in Table 3.3-1.

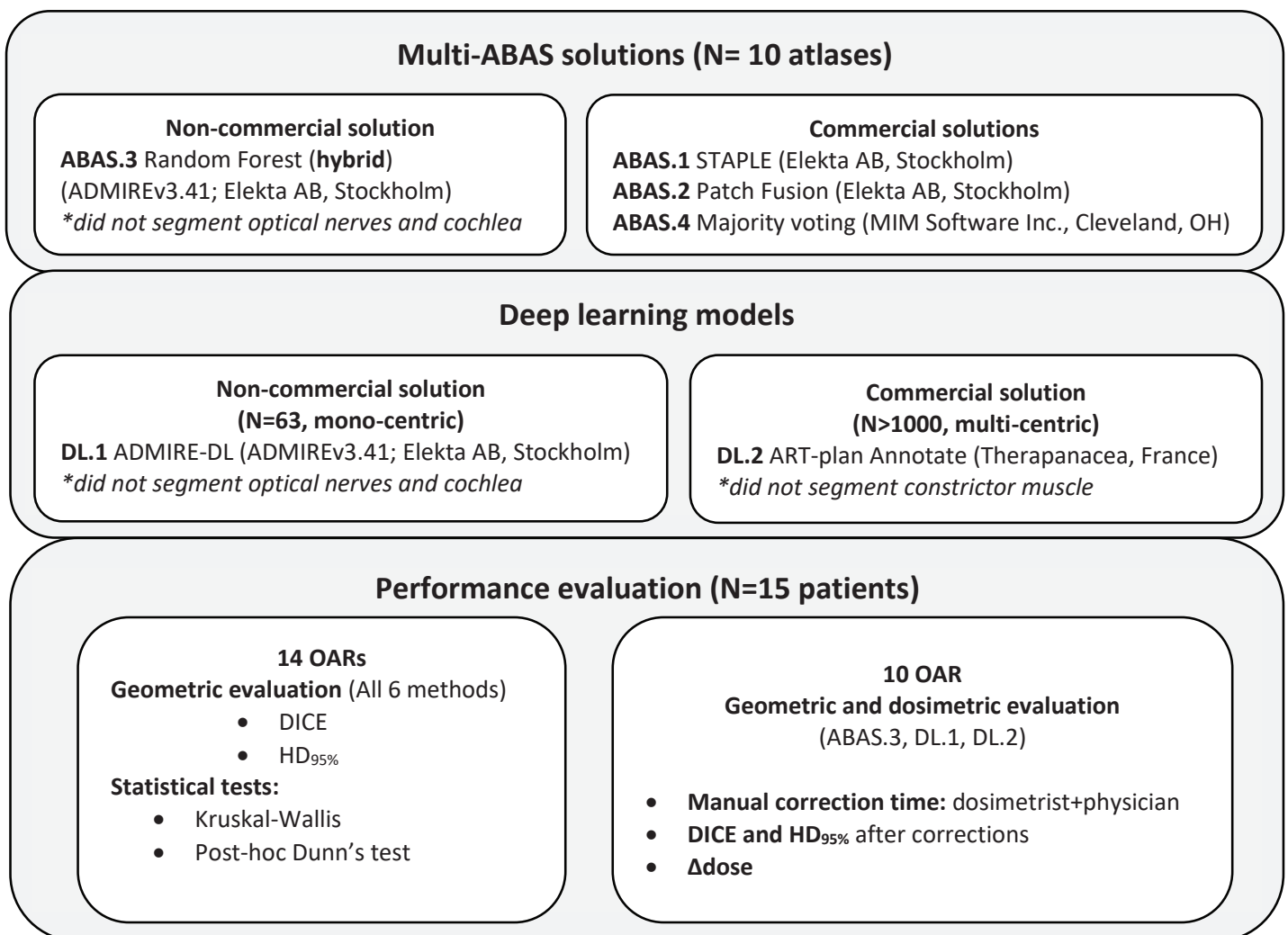


Figure 3.3-1 Overview of the study design and performance evaluation methods; OAR = organs-at-risk, HD_{95%}=95th percentile-Hausdorff Distance; * indicates the OAR that were not segmented by certain methods; Δdose = difference between the reference plan created with corrected OAR contours and the plan created with AS contours only;

Table 3.3-1 Characteristics of the testing cohort used for evaluation of the AS solutions

	Tumor localization	TNM	BMI	Gender	Age
Patient 1	Oral cavity	T4aN0M0	24.1	M	88Y
Patient 2	Hypopharynx	T3N2aM0	12.1	M	57Y
Patient 3	Nasal cavity	T2N0M0	31	F	83Y
Patient 4	Tonsils	T2N0M0	20.8	M	75Y
Patient 5	Hypopharynx	T0N3M0	24	M	45Y
Patient 6	Rhinopharynx	T3N1M0	19	F	58Y
Patient 7	Rhinopharynx	T3N0M0	19.8	F	69Y
Patient 8	Rhinopharynx	T2N2M0	21.4	F	71Y
Patient 9	Hypopharynx	T1N1M0	23.4	M	75Y
Patient 10	Larynx	T2N0M0	30.4	M	60Y
Patient 11	Tonsils	T2N0M0	24	F	69Y
Patient 12	Unilateral ganglion	T4N3M1	34.7	M	54Y
Patient 13	Parapharynx	T2N1M0	25.5	M	65Y
Patient 14	Larynx	T4aN0M0	21.5	M	58Y
Patient 15	Hypopharynx	T4bN0M0	19	M	75Y

Multi-ABAS and DL methodologies

Three multi-ABAS solutions integrated in the research version of Monaco TPS [323] (Monaco 5.59.11 with ADMIREv.3.41) and another one available in MIM-Maestro (MIM Software Inc., Cleveland, OH) were investigated:

- ABAS.1: STAPLE consists in estimating the optimal combination of the atlases segmentations by weighting each segmentation upon the estimated performance level based on expectation-maximization algorithm [156].
- ABAS.2: Patch Fusion algorithm computes the final probability of a voxel to belong to a structure as a weighted average of the atlases' contours based on voxel intensity information [165].
- ABAS.3: Random Forest (RF) is a supervised learning algorithm which constructs a voxel classifier for each structure using the registered atlases as training data [174].
- ABAS.4: Majority voting (MIM) [162].

For the ADMIRE software, out of the 10 atlases used, a reference patient was selected for each test patient based on the closest BMI of the atlas and the underlying patient. No individual atlas selection was required for MIM, but a general *template scan* (patient having an anatomy close to the mean BMI of the atlas cohort) was registered with all the atlases in the library.

Two DL models were investigated:

- DL.1: ADMIRE-DL (ADMIREv.3.41, Elekta AB, Stockholm) trained with N=63 patients from one center. It is a fully connected deep convolutional neural network (DCNN) with 3D U-net architecture and short-range residual connections developed from the ResUnet3D network [124]. While the encoding part is responsible for learning multi-scale multi-dimensional image features in multiple levels, the combination of long and short-range connections allows the decoding part to preserve the high-resolution image features and produce a label map corresponding to the input image size [124,190].

- DL.2: ART-plan Annotate (Therapanacea, France) trained on a large database with N>1000 patients obtained from several clinical sites. The model uses anatomy preserving DL ensemble networks that first detects organs through DL-based registration to a collection of whole-body annotated volumes. Then, the delineation of each anatomical structure is performed through an original combination of data-driven and decisional artificial intelligence that enforces anatomical consistency [206,217].

Geometric evaluation of auto-segmentation solutions

To quantitatively evaluate the segmentation results, we used two geometric indices: volumetric DICE and 95th percentile-Hausdorff distance (HD_{95%}) [135]. DICE is a measure of the volumetric overlap between the ground truth contour (A) and the predicted segmentation (B), leading to a value between 0 (no overlap) and 1 (perfect overlap):

$$DICE = \frac{2x|A \cap B|}{|A| + |B|}$$

However, DICE is limited to the pixels overlap without considering the shape differences. Therefore, a second metric was used to indicate the magnitude of mislocalization of the prediction. The HD is a boundary-based metric that measures the surface distances between the predicted contour and the ground truth segmentation. To eliminate the possible outliers, we used HD_{95%}:

$$HD_{95\%} = \max_{k \in 95\%} [d(A, B), d(B, A)]$$

$$d(A, B) = \max_{a \in A} \min_{b \in B} \|a - b\|$$

where d(A,B) is the directed HD and A and B are the set of non-zero pixels in the images. HD metric has its own limitation that does not focus on the object itself therefore does not punish a prediction with a large hole inside or with a spotted pattern within the contour [44]. For the elongated organs (i.e., esophagus, trachea, constrictor muscle and spinal cord) the results were calculated only on the slices where both contours were present to avoid situations where the reference ground truth was missing.

Time needed for manual corrections

Three of the automatic solutions (ABAS.3, DL.1 and DL.2) were clinically reviewed and corrected by a dosimetrist and validated by a skilled physician on Monaco contouring workstation following the regular clinical routine. The time spent on correcting and validating each structure was recorded independently.

Dosimetric evaluation – automatic treatment plans

For each patient, and for ABAS.3, DL.1 and DL.2 solutions, 2 different plans were generated: one using the AS contours and another one using AS+manually corrected contours. The differences in dose distributions were then evaluated on the corrected contours. In total, 90 VMAT treatment plans were calculated with mCycle auto-planning solution (Monaco 5.59.11, Elekta AB, Stockholm). The software uses a lexicographic MCO which has been extensively described before [2]. All plans were performed using two 360° arcs. A simultaneous integrated boost technique was used for delivering 70Gy to the planned target volume (PTV) associated to the primary tumor and 54.25Gy to the PTV associated to prophylactic nodal target, in 35 fractions of 2Gy. Clinically relevant dosimetric endpoints for target volumes (V_{95%}) and OAR (D_{mean}, D_{2%}, D_{5%}) were considered upon the clinical protocol and according to the recommendations of the French Society of Radiation Oncology [25].

Statistical analysis

Per organ and per algorithm, statistical differences between methods were assessed using the non-parametric Kruskal-Wallis test. Subsequently, to detect between which pairs of algorithms the differences were significant, the post-hoc Dunn's test with Bonferroni correction was applied. Similarly, the differences between radiotherapy doses derived from AS contours with or without corrections were tested for statistical significance. P-values <0.05 were considered significant. The statistical analysis was performed using the libraries (scipy 1.6 and scikit-posthocs 0.7) in Python 3.8.

3.4. Results

Computational time per patient was in average 10.3 ± 1.6 min, 10.5 ± 0.6 min and 12.1 ± 0.6 min for ABAS.1, ABAS.2 and ABAS.3, respectively. For ABAS.4, it was under 1min while the atlas registration took approximately 6min for a library of 10 atlases. DL.1 and DL.2 provided a solution in less than 1min and 2min, respectively. Per algorithm and per OAR, DICE scores and HD_{95%} distance results of all solutions are summarized in Figure 3.4-1 and Figure 3.4-2, respectively.

Overall, both DICE and HD_{95%} results indicated that DL algorithms were more accurate than ABAS algorithms for AS of HN OAR. The Kruskal-Wallis statistical test identified significant differences between the 6 AS methods. However, the post-hoc paired test showed no statistical difference in terms of DICE and HD_{95%} between the DL.1 and DL.2 and between ABAS.3 and the 2 DL solutions. With 11 common OAR, DL.1 had overall better contour overlap compared with DL.2 with a DICE average of 0.85 ± 0.32 vs 0.82 ± 0.06 and 11 vs 9 OAR having DICES ≥ 0.8 . Per organ differences however did not reach a statistically significant level.

Regarding ABAS solutions, ADMIRE ABAS algorithms had overall better DICES and HD_{95%} than ABAS.4, which had the lowest DICE results for 7 out of 14 OAR. While all the ADMIRE solutions had DICE results ≥ 0.8 for 7 OAR, ABAS.3 contours were closer to the reference contours. Per OAR statistics revealed however no significant differences in DICE and HD_{95%} between ABAS.2 and ABAS.3 and, compared with ABAS.1, both ABAS.2 and ABAS.3 performed significantly better for the eyes ($p < 0.02$). Moreover, compared with ABAS.4, ABAS.3 performed significantly better for parotids ($p < 0.003$), mandible ($p < 0.01$) and submandibular glands ($p < 0.02$). Note that ABAS.3 did not segment optic nerves and cochlea owing the algorithm's limitation for such small structures.

Compared with DL.1, ABAS.3 had significantly better DICE for the mandible ($p = 0.02$). Compared with DL.2, ABAS.3 had significantly better DICE for the eyes ($p = 0.01$) and for the mandible ($p = 0.01$). On the opposite, DL.2 had significantly better DICE for the esophagus ($p = 0.04$) and significantly better HD_{95%} for the thyroid ($p = 0.03$). Finally, the superiority of DL. 1 over ABAS.3 was not statistically demonstrated.

An example of AS contours from ABAS.3, DL.1 and DL.2 in contrast with the physicians' manual delineations is provided in Figure 3.4-3.

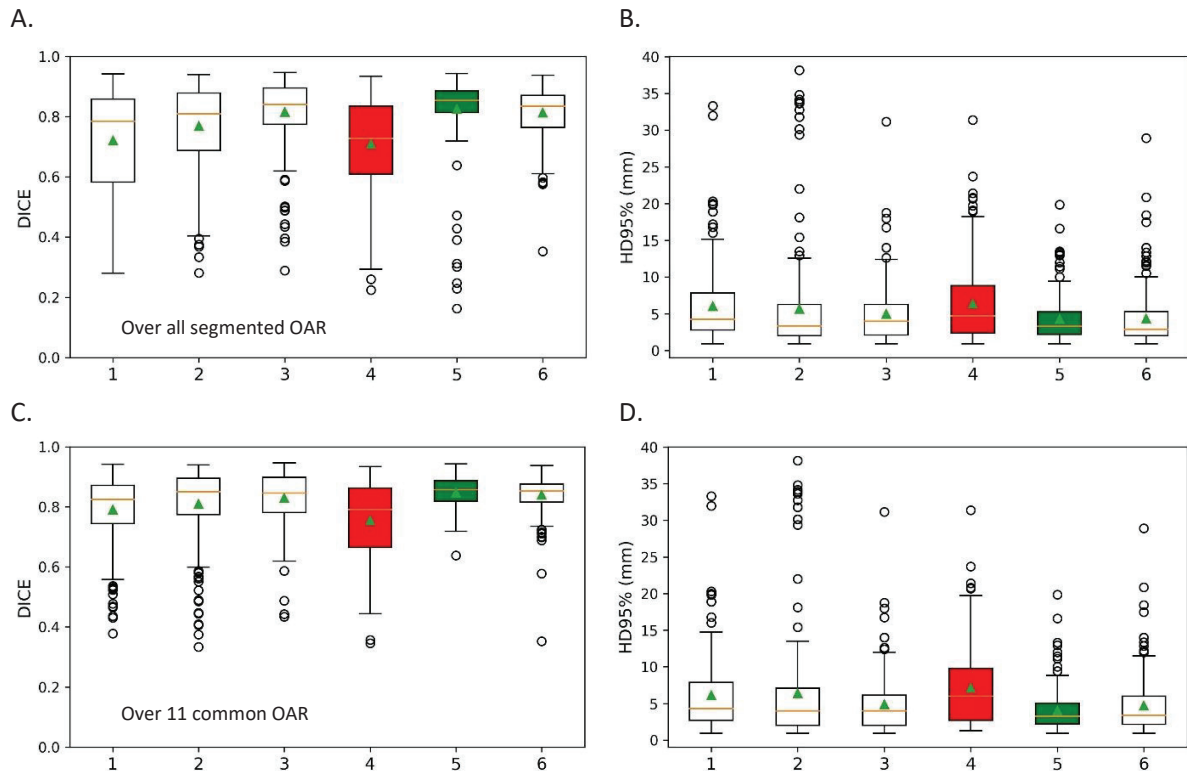


Figure 3.4-1 Geometric evaluation of the 6 automatic solutions. 1 = ABAS.1, 2 = ABAS.2, 3 = ABAS.3, 4 = ABAS.4, 5 = DL.1, 6 = DL.2; Panels A and B: analysis was performed over all the OAR. Panels C and D: analysis was performed over 11 common OAR. In red and in green are highlighted the worst and the best results, respectively determined by the mean value of DICE/HD_{95%}; in the boxplots, the orange line represents the median, the green triangle indicate the mean value and the circles represent outliers.

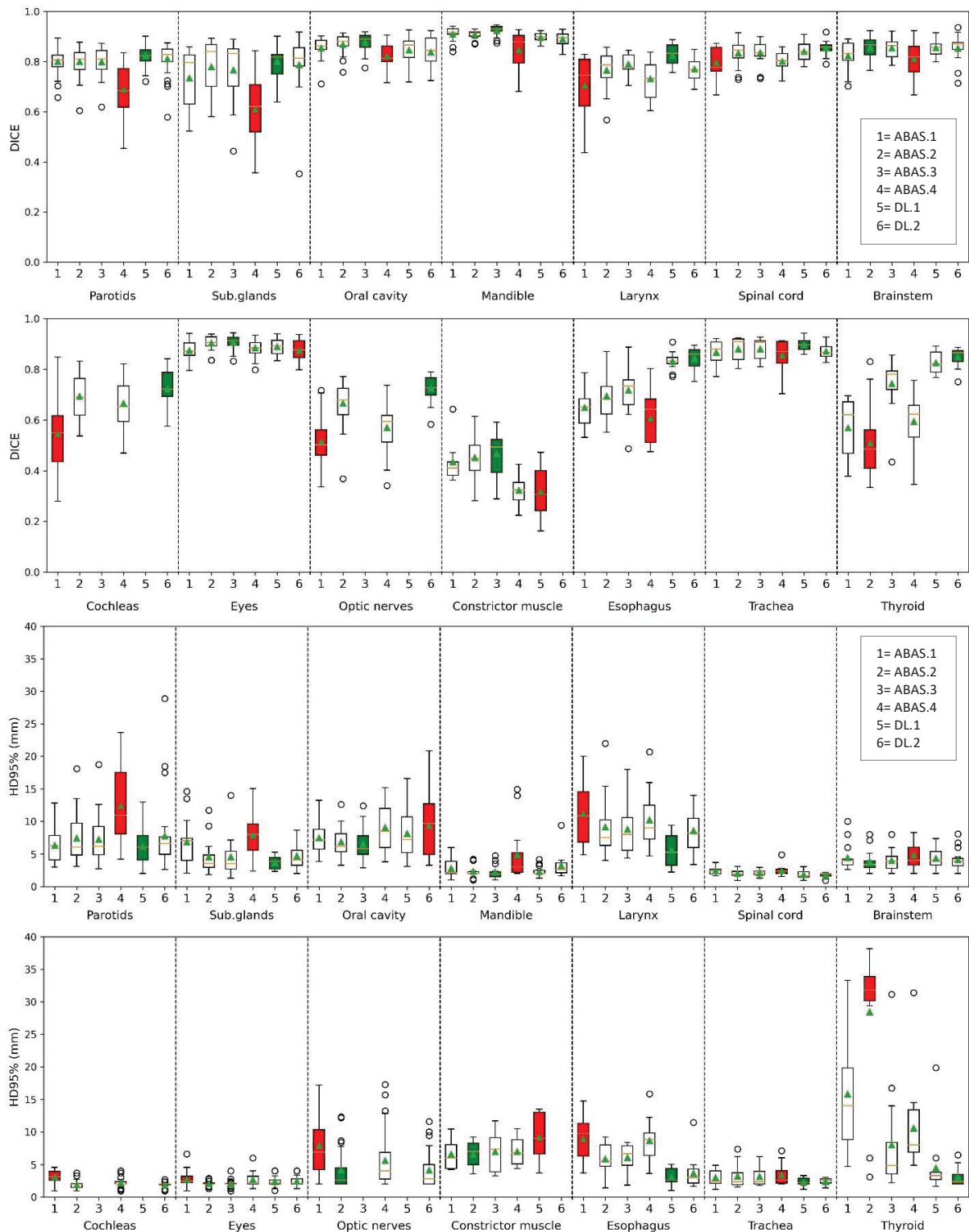


Figure 3.4-2 Geometric evaluation per OAR of the 4 multi-ABAS and 2 DL solutions; 1 = ABAS.1, 2 = ABAS.2, 3 = ABAS.3, 4 = ABAS.4, 5 = DL.1, 6 = DL.2; in red and in green are highlighted the worst and the best results, respectively determined by the mean value of DICE/HD95%; in the boxplots, the orange line represents the median, the green triangle indicate the mean value and the circles represent outliers; Abbreviations: Sub.glands = submandibular glands;

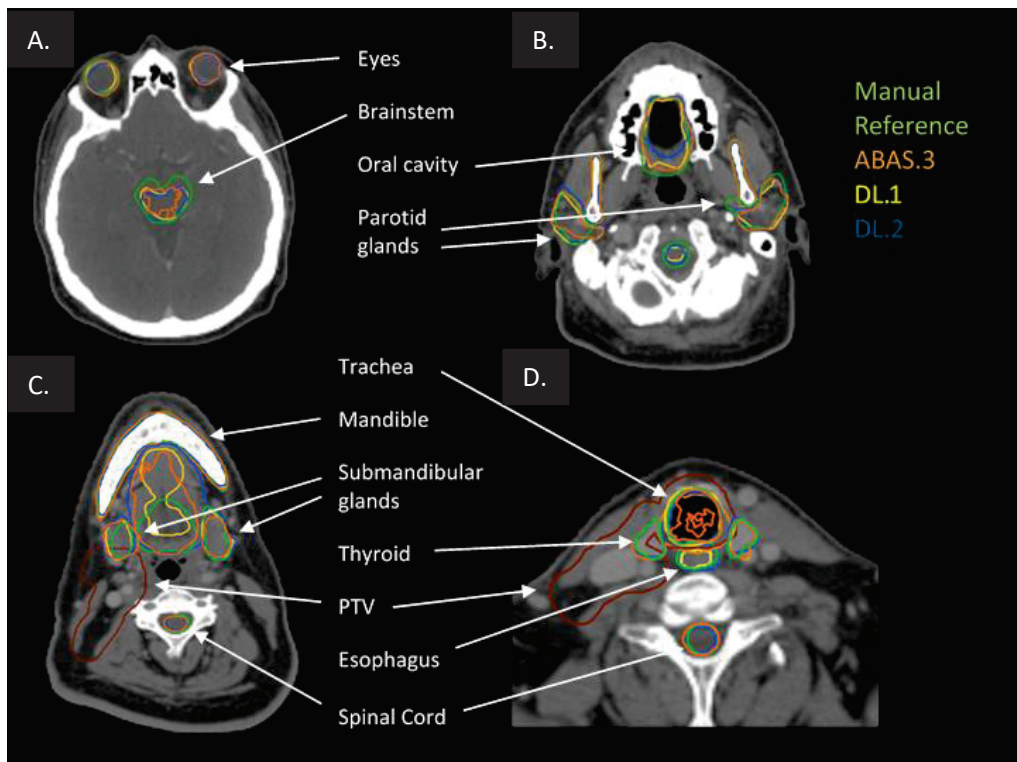


Figure 3.4-3 Example of automatic segmentation uncertainties compared with manual delineations. OAR position relative to the PTV can be observed in C and D. A good agreement was generally observed in simple geometry structures such as eyes or spinal cord. Large contour discrepancies were noticed compared with the manual reference in the cranial and caudal slices for some structures such as oral cavity, trachea or brainstem. to illustrate OAR position relative the target, PTV volume is displayed

Ten of the OAR obtained with ABAS.3, DL.1 and DL.2 (best solutions graded based on the geometric accuracy results) were thereafter carefully corrected by a dosimetrist and checked by a physician. Manual corrections were done organ by organ on all the CT slices. The targets were never displayed, to not influence the observers. The manual correction time per patient was in average 36min34sec, 17min54sec and 26min57sec for ABAS.3, DL.1, and DL.2, respectively. The contours generated by DL.1, were the fastest to correct. In general, manual corrections of eyes, spinal cord and brainstem were <2min for the 3 solutions while for oral cavity and esophagus correction times were >3min depending on the AS algorithm used (Figure 3.4-4).

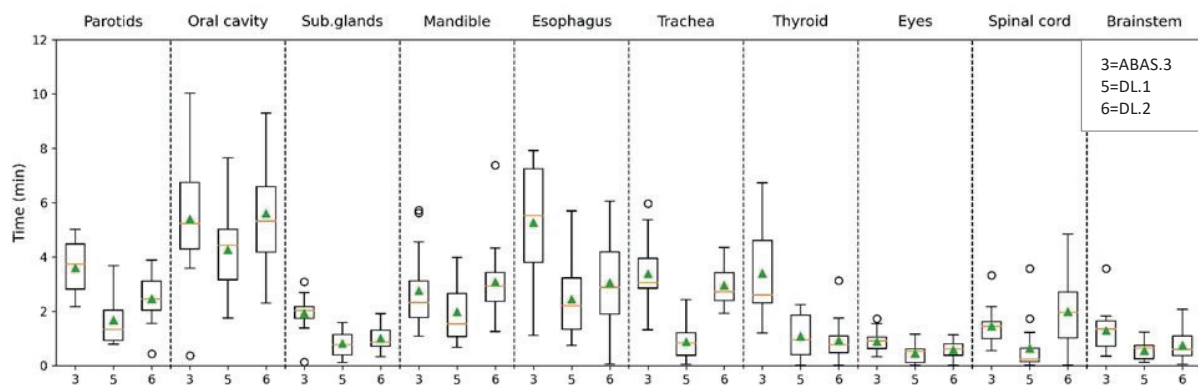


Figure 3.4-4 Time spent on manual corrections for each OAR automatically generated. 3 = ABAS.3, 5 = DL.1, and 6 = DL.2; Abbreviations: Sub.glands = submandibular glands;

After manual corrections, the DICE scores of all OAR were improved, except for the oral cavity on all 3 solutions, and for the spinal cord on DL.2 solution (Table 3.4-1), thus highlighting inter-observer variability in contouring the oral cavity between the expert physician providing the reference contours, and the other physician performing manual corrections. At the same time, the HD_{95%} did not decrease consistently for all the structures after the manual corrections, confirming, once more, the variability in manual delineation between observers. While performing correction on the DL.1 contours did not significantly improve DICE and HD_{95%} results, for DL.2 contours, results were significantly improved for the trachea ($p < 0.001$). For ABAS.3, the improvements were statistically significant for esophagus ($p < 0.001$) and thyroid ($p < 0.001$).

The differences in doses on corrected OAR, between treatment plans generated using the AS contours, with or without manual corrections are presented in

Table 3.4-2. No statistically significant difference was found between doses for the 3 solutions. For each patient, a minimum distance between each OAR and the targets was calculated. Among OAR having a maximum dose constraint, the mandible had the largest dose difference when it overlapped with the PTV. For the brainstem and spinal cord, the largest dose differences occurred when the OAR was located at a larger distance to the PTV (>30mm). For the parotids and for the submandibular glands, maximum differences occurred when the OAR overlapped with the PTV. For the oral cavity, for the eyes and for the esophagus, the maximum differences were generally observed at distances <20 mm from the PTV. However, for the esophagus, there were some outliers at larger distances from the target (>60mm) for DL.2. For the trachea, only in one patient case, and for DL.2, a large difference was observed but at a high distance from the target (80mm). Some illustrations of dose distributions with regard to corrected/non-corrected contours and PTV position are available in Figure 3.4-5 and a summary of studies on AS methods for OAR on HN CT images is presented in Table 3.4-3.

Table 3.4-1 Geometric evaluation after manual corrections of 10 OAR for the three best solutions; with * are marked the differences that are statistically significant ($p < 0.05$)

	ABAS.3		DL.1		DL.2	
	without corrections	after manual corrections	without corrections	after manual corrections	without corrections	after manual corrections
DICE						
Parotids	0.8 ± 0.05	0.84 ± 0.04	0.82 ± 0.04	0.84 ± 0.04	0.81 ± 0.06	0.85 ± 0.03
Oral cavity	0.87 ± 0.04	0.81 ± 0.06	0.85 ± 0.06	0.79 ± 0.08	0.84 ± 0.07	0.79 ± 0.08
Sub.glands	0.77 ± 0.13	0.83 ± 0.1	0.8 ± 0.07	0.84 ± 0.07	0.79 ± 0.13	0.82 ± 0.14
Mandible	0.92 ± 0.02	0.93 ± 0.02	0.9 ± 0.02	0.9 ± 0.02	0.89 ± 0.03	0.9 ± 0.03
Esophagus	0.72 ± 0.1	0.86 ± 0.04*	0.83 ± 0.04	0.86 ± 0.03	0.84 ± 0.05	0.87 ± 0.03
Trachea	0.88 ± 0.05	0.91 ± 0.04	0.9 ± 0.02	0.9 ± 0.06	0.87 ± 0.03	0.91 ± 0.04*
Thyroid	0.74 ± 0.11	0.85 ± 0.03 *	0.83 ± 0.04	0.85 ± 0.04	0.85 ± 0.04	0.86 ± 0.03
Eyes	0.91 ± 0.03	0.91 ± 0.03	0.89 ± 0.03	0.9 ± 0.03	0.87 ± 0.04	0.9 ± 0.03
Spinal cord	0.84 ± 0.05	0.84 ± 0.05	0.84 ± 0.04	0.85 ± 0.04	0.85 ± 0.03	0.84 ± 0.04
Brainstem	0.85 ± 0.04	0.86 ± 0.05	0.85 ± 0.03	0.86 ± 0.04	0.85 ± 0.06	0.86 ± 0.06
HD _{95%} (mm)						
Parotids	7.2 ± 3.4	7.9 ± 7.2	6.2 ± 2.6	8.4 ± 8.1	7.8 ± 5.2	7.0 ± 7.1
Oral cavity	6.5 ± 2.6	11.2 ± 3.9	8.1 ± 3.9	11.4 ± 5.1	9.4 ± 5.1	12 ± 5.3
Sub.glands	4.5 ± 3.1	4.0 ± 2.7	3.7 ± 0.9	2.9 ± 1.3	4.7 ± 2.2	3.8 ± 2.1
Mandible	2.2 ± 1.1	1.5 ± 1	2.3 ± 0.8	2.1 ± 0.9	3.7 ± 2	2.1 ± 1.4
Esophagus	6.1 ± 2.3	2.3 ± 0.9*	3 ± 1.3	2.7 ± 0.7	3.6 ± 2.6	1.9 ± 0.4
Trachea	3.2 ± 1.6	1.9 ± 0.5	2.3 ± 0.7	1.8 ± 0.7	2.4 ± 0.5	1.8 ± 0.6*
Thyroid	8 ± 8.3	2.5 ± 1.2 *	4.5 ± 4.7	2.2 ± 0.6	3.0 ± 1.4	2.5 ± 1.3
Eyes	2. ± 0.5	1.9 ± 0.5	2.4 ± 0.8	2 ± 0.4	2.4 ± 0.8	2.2 ± 0.7
Spinal cord	2.1 ± 0.5	2 ± 0.5	1.8 ± 0.6	1.8 ± 0.5	1.7 ± 0.4	2 ± 0.5
Brainstem	3.9 ± 1.6	3.9 ± 1.9	4.4 ± 1.6	4.1 ± 1.7	4.1 ± 1.7	3.9 ± 1.8

Abbreviations: Sub.glands=submandibular glands

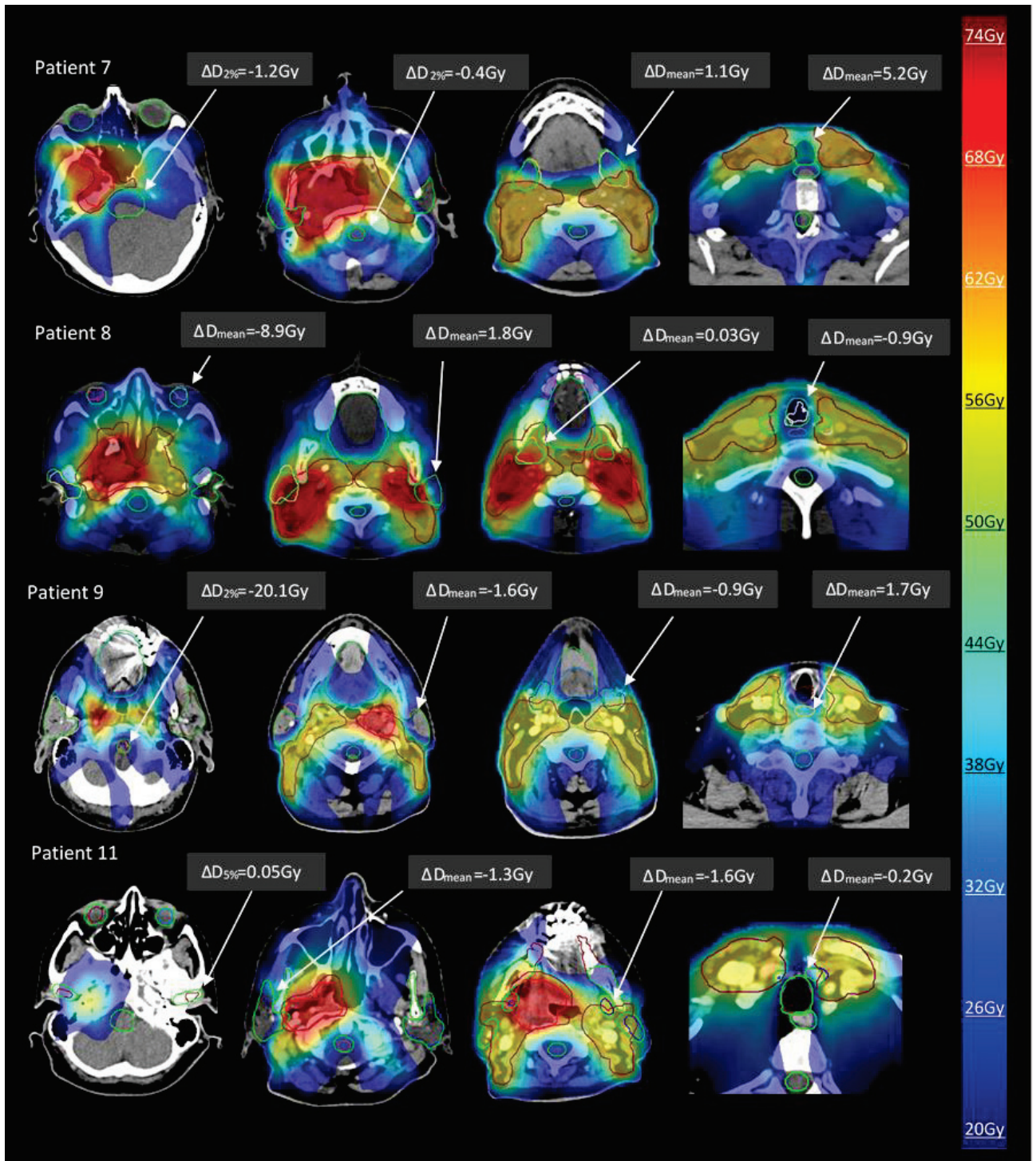


Figure 3.4-5 Illustrations of radiation dose distributions of plan generated with the AS contours. With green are highlighted AS+manually corrected contours; the PTV volumes are in red and dark red so their position relative to the OAR can be observed. The arrows point OAR and their correspondent dose difference in the corrected contour between plans created with AS+corrections and plans created with AS contours ($\Delta D_{AS+corr} = D_{AS+corr} - D_{AS}$)

Table 3.4-2 Dosimetric differences between doses generated with manually corrected contours and automatic contours, analyzed on the corrected contours; impact on the target volumes is evaluated in $V_{95\%}$ dose coverage, for the spinal cord and brainstem in $D_{2\%}$, and for the mandible in $D_{5\%}$ while for the rest of the OARs mean doses are calculated.

Structure	ABAS.3			DL.1			DL.2		
	$V_{95\%}$ [%]	$\Delta V_{95\%}$ [%]	$\Delta V_{95\%}$ ranges [%]	$V_{95\%}$ [%]	$\Delta V_{95\%}$ [%]	$\Delta V_{95\%}$ ranges [%]	$V_{95\%}$ [%]	$\Delta V_{95\%}$ [%]	$\Delta V_{95\%}$ ranges [%]
PTV_70Gy	98.3	-0.1	[-0.9, 0.7]	98.4	0.2	[-1.4, 1.3]	98.3	-0.04	[-0.8, 2.1]
PTV_54.25Gy	99.2	-0.03	[-0.2, 0.1]	99.2	0.01	[-0.2, 0.3]	99.2	0.05	[-0.1, 0.2]
	D_{mean} [Gy]	ΔD_{mean} [Gy]	ΔD_{mean} ranges [Gy]	D_{mean} [Gy]	ΔD_{mean} [Gy]	ΔD_{mean} ranges [Gy]	D_{mean} [Gy]	ΔD_{mean} [Gy]	ΔD_{mean} ranges [Gy]
Parotid R	16.2	-0.1	[-3.4, 3.3]	16.8	0.3	[-1.8, 3.4]	17.6	0.3	[-2.2, 3.6]
Parotid L	12.9	-0.04	[-4.3, 3.3]	15.1	0.4	[-2.4, 4]	15	1.1	[-0.8, 6]
Oral cavity	21	-0.02	[-2.7, 3.2]	21.1	-0.6	[-2.2, 0.9]	21	-0.7	[-3.8, 1.3]
Sub.gland R	32.5	-0.1	[-0.9, 1.7]	38.7	-0.5	[-2.3, 1.2]	39.1	-0.4	[-7.4, 3.1]
Sub.gland L	34.1	-0.4	[-2.8, 1.2]	40.6	-0.3	[-1.8, 1.3]	40.9	0.1	[-1.6, 1.1]
Esophagus	7.9	0.2	[-0.5, 1.7]	9.2	0.5	[-1.1, 2.4]	10.2	0.9	[-0.3, 5.5]
Trachea	10.1	-0.2	[-0.9, 0.2]	11.9	-0.1	[-0.9, 0.4]	13.5	0.2	[-1.1, 5.2]
Thyroid	29.4	-0.2	[-4.8, 1.1]	35.5	-0.5	[-3.7, 1.9]	35.4	-0.3	[-2, 1.3]
Eye R	5.2	0.5	[-1.5, 4.2]	4.9	0.4	[-4.1, 10.3]	5.1	-0.4	[-6.1, 0.8]
Eye L	4.5	-0.4	[-8.9, 2.5]	5.4	0.8	[-0.4, 8.8]	4.4	-0.8	[-8.9, 1.2]
	$D_{2\%}$ [Gy]	$\Delta D_{2\%}$ [Gy]	$\Delta D_{2\%}$ ranges [Gy]	$D_{2\%}$ [Gy]	$\Delta D_{2\%}$ [Gy]	$\Delta D_{2\%}$ ranges [Gy]	$D_{2\%}$ [Gy]	$\Delta D_{2\%}$ [Gy]	$\Delta D_{2\%}$ ranges [Gy]
Spinal cord	25	2.5	[-4.6, 24.4]	23.7	-0.5	[-16.3, 10.1]	24.2	-2.1	[-13.6, 16.2]
Brainstem	16.9	-1.5	[-20.2, 13.1]	18	-1.0	[-7.1, 4.9]	18.9	-0.6	[-9.7, 6.9]
	$D_{5\%}$ [Gy]	$\Delta D_{5\%}$ [Gy]	$\Delta D_{5\%}$ ranges [Gy]	$D_{5\%}$ [Gy]	$\Delta D_{5\%}$ [Gy]	$\Delta D_{5\%}$ ranges [Gy]	$D_{5\%}$ [Gy]	$\Delta D_{5\%}$ [Gy]	$\Delta D_{5\%}$ ranges [Gy]
Mandible	43.3	0.1	[-3.6, 2.1]	43.4	-0.2	[-2.9, 6.3]	44.4	0.35	[-2.5, 4.9]

Abbreviations: sub.glands=submandibular glands; R = right, L = left;

Table 3.4-3 Studies on automatic segmentation (AS) methods for Head-and-Neck (HN) organs-at-risk (OAR)

Authors	Automatic segmentation approach	Nr. of atlases/ Nr. of patients	Nr. of test samples	Nr. of segmented OAR	Performance evaluation	Time record	Comments
Han et al. (2008) [152]	Single ABAS (MI similarity) Multi-ABAS (STAPLE)	9 atlases	10	7	DICE	Computation time: 7min/atlas	5/7 DICE>0.8 for multi-ABAS (exception spinal cord and submandibular glands)
Teguh et al. (2010) [157]	Single-ABAS (MI similarity) Multi-ABAS (STAPLE)	10 atlases	12	20	DICE MD	Computation time: 7min/patient Mean manual editing time: 35min/patient	100% of the auto-contours scored as “minor deviation, editable” or better
La Macchia et al. (2012) [154]	Single ABAS (VelocityAI 2.6.2) MIIM 5.1.1 ABAS 2.0)	1 atlas	5	16	DICE Volume Sensitivity Specificity CoM	Time gain: ABAS 1h23 min MIIM 1h4 min VelocityAI 45min	ABAS = underestimation of the volumes MIIM and VelocityAI = overestimation of the volumes
Fortunati et al. (2013) [173]	Multi-ABAS (majority voting) Hybrid-ABAS (intensity modeling)	17 atlases	18	11	DICE HD _{mean} HD	Computational time: 10min/atlas registration (170min) ABAS majority voting 45s/patient Intensity modeling: 6min/patient (total ~3h)	Hybrid ABAS = more robust than Multi-ABAS with 8/11 OARs close to inter-observer agreement
Walker et al. (2014) [170]	Hybrid ABAS (SPICE, Pinnacle9.4)	Unknown	40	16	DICE	Time savings ranged from 2.3–15.7 min concluding in 30.9% average time reduction	Optic nerves, cochlea, oral cavity and larynx poorly segmented (DICE<0.71) Other contours within 8 resident physicians’ agreement
Gresswell et al. (2017) [155]	Multi ABAS (in house ABAS)	9 atlases	10	16	DICE MDA	Atlas registration time: 23s/patient	DICE increased with increasing atlas number, while MDA did not decrease after 6 atlases.
Ibragimov et al. (2017) [193]	Deep Learning (13 separate CNN)	48 patients (max 48 samples)	50	13	DICE	/	Lower performance compared to literature for segmentation of parotid glands, submandibular glands, and optic chiasm
Willems et al. (2018) [194]	Deep learning (DeepVoxNet-3D CNN)	70 patients	20	16	DICE HD ASSD	Manual delineation 45-120min Segmentation time <4min Average correction time: 15min/patient	Small corrections needed except for the upper esophagus and glottic area.

Van Rooij et al. (2019) [144]	Deep learning (3D CNN U-Net)	142 patients (max 118 samples)	15	9	DICE Sensitivity PPP D_{max} D_{mean}	Segmentation time: <10s/patient Hyperparameter optimization time (~24h/ tuning loop/OAR)	Lower geometric performance for structures with smaller number of training samples (esophagus; brainstem) but this did not translate into inferior OAR dosimetry; For noncritical OARs, checking can be omitted.
Zhu et al. (2019) [186]	Deep Learning (Anatomynet U-net)	261 patients	10	9	DICE $HD_{95\%}$	Segmentation time: 0.12s	Multi-institutional training data Better DICE results versus results from MICCAI 2015 competition Lower performance for small organs
van Dijk et al. (2019) [143]	Deep learning (DLCExpert-CNN) ABAS majority voting (WorkflowBox 1.4, Mirada Medical Ltd.)	589 patients 30 atlases	104	22	DICE HD Volume ΔD_{mean} ΔD_{max}	For 7 OARs, manual delineation time: <90min/patient; Average correction time: <36min for an expert and <60min for a beginner observer	DLCExpert resulted in equally or better performance in 19/22 OARs versus ABAS.
Chen et al. (2021) [191]	Deep Learning (WBNet on whole body CT)	150 patients (in house) 70 patients (public dataset)	30 (in house) 70 (public dataset)	28	DICE $HD_{95\%}$ ΔD_{mean} ΔD_{max}	Manual delineation time: 33.60 ± 2.55min Manual editing time: 13.10 ± 3.14min; 61% time reduction for HN OARs	Using the same data for training, WBNet had superior performance for most of the OAR compared to ABAS, AnatomyNet and nnU-Net, on in-house test set as well as on public data set
Costea et al. (2022) [3]	Multi-ABAS (MIM, STAPLE, PF) Hybrid-ABAS (RF) Deep Learning (ADMIRE-DL, ART-plan Annotate)	10 atlases 10 atlases N=63 N>1000	15	14	DICE $HD_{95\%}$ $\Delta V_{95\%}$ ΔD_{mean} $\Delta D_{2\%}$ $\Delta D_{5\%}$	Manual correction time for 10 OARs: 36min34sec (Hybrid ABAS) 26min57sec (ART-plan Annotate) 17min54sec (ADMIRE-DL)	RF and DL solutions superior to MIM, STAPLE and PF solutions Improved workflow efficiency with the DL solutions

Abbreviations: ABAS = atlas-based AS; DL = deep learning; CoM = center of mass; MI = mutual information; HD = maximum Hausdorff distance; MSHD = mean slice-wise Hausdorff Distance; $HD_{95\%}$ = 95th percentile-HD; $HD50\%$ = medianHD; HD_{mean} =mean HD; STSD =surface-to-surface distances; MD =mean absolute distance; MDA = mean distance to agreement; MSD = mean surface distance; DTA = distance to agreement; ASSD = average symmetric surface distance; PFP = proportion of false positives; Δ = dose difference compared to manual contours; dosimetric criteria: ΔD_{mean} , ΔD_{max} , $\Delta D_{2\%}$, $\Delta D_{5\%}$, ΔV_{95}

3.5. Discussion

We showed in this study that, overall, both DICE and HD_{95%} results indicated that DL algorithms performed better compared with the ABAS algorithms for automatic segmentation of HN OAR. Concerning the 2 DL solutions, out of 12 contours, DL.1 outperformed DL.2 solution in terms of DICE for 7 OARs, with, however, no statistically significant differences. Contrarily to DL.2, DL.1 was not tailored to automatically contour optic nerves and cochlea. Nevertheless, the correction of the AS contour of small organs generally takes more time than starting from scratch [123,194]. Conversely, DL.2 was not trained to contour the constrictor muscle. However, the DL.1 results were highly inaccurate, showing the difficulty to get satisfying results for such organs with high anatomical variations and low image contrast. Therefore, consistent with the literature, OAR with good CT contrast were better segmented by ABAS and DL solutions compared with small and/or thin OAR such as optic nerves or cochlea, and OAR which do not have well-defined boundaries like constrictor muscles [143,163,186,191,324,325].

Before this study, DL.1 and DL.2 algorithms had not been explored on HN site. The DL.1 algorithm was trained exclusively with manual delineations coming from one expert physician, providing uniformity of the training data. Ideally, there should be a consortium for the contour delineation between physicians working in a radiotherapy department, which should rely on internationally published guidelines [26]. In this study, with a limited training dataset (N=63), we showed that a model can achieve consistent results for most of the structures in HN. Hence, with a minimum of work, centers can adapt a model to their standard delineation's practices. Similarly, high accuracy segmentation results were obtained with the DeepVoxNet and another CNN with networks trained on N=70 and N=50 samples, respectively [193,194]. Other studies demonstrated that organs' pattern depends on the training sample size [326] and yet similar results can be obtained when training on a small set of carefully curated data compared with a larger set of more easily available routine-level clinical annotations [327]. On the opposite, DL.2 solution was trained with more than 1000 samples per organ collected from multiple centers and can segment 50 OAR and target volumes in HN. Despite this, highly accurate contours were obtained in this study. Proving that a multi-center study approach includes combination of manual contours from different physicians (easier to obtain), DL.2 results presented good conformity to new datasets and comparable performance to a model train with data from a single center.

We also showed that, using a carefully selected atlas of patients, ADMIRE multi-ABAS methods achieved good agreement with manual contours (DICE \geq 0.8)[152,157] and, for some organs, similar or better agreement with the reference contours compared with DL models (i.e. oral cavity, mandible, eyes). Conversely, ABAS.4 had overall inferior performance. Among multi-ABAS algorithms, ABAS.3, which had not been explored before, produced the best results and had significantly better DICEs than DL.1 and DL.2 solutions for mandible and eyes, respectively. Therefore, with only 10 carefully selected atlases composed of non-operated patients with a wide range of BMI, ABAS.3 algorithm may serve as an AS solution easy to implement clinically. Note that using an enlarged library of 20 patients (data not shown) did not considerably improve the performances of ABAS.3 but drastically increased the computation time, demonstrating that the performance plateau phenomena still exists with this new ABAS method.

Many studies have reported the performances of different algorithms for HN OAR segmentation on CT images (Table 3.4-3). All studies underlined limited performance on small organs, and the importance of both manual contours' quality, and training data size to obtain accurate segmentations and clinically acceptable treatment plans. It was also mentioned that, for noncritical OAR (i.e. far from PTV), manual corrections could be omitted [144]. Moreover, AS has shown to reduce inter-observer variability when observers performed manual editing on the automatically generated contours, which improves the consistency of manual delineation [142].

According to the recently published guidelines, together with geometric accuracy, studies should ideally report benefit in time saving and clinical acceptability in terms of patient dose evaluation, for

assessing the benefit of an automatic segmentation method [135]. Both tasks involve exhaustive labor and are not systematically conducted first because of the time requested to be completed, and secondly because of the intra-observer factor, which could introduce a bias in the observations. In this study, both tasks were completed for the three best algorithms, and an auto-planning solution was used to perform treatment plans based on AS contours with or without corrections. This was a strength of this study, and an efficient way to isolate the consequences of contour variations on the radiotherapy doses and reveal more precisely which contours require greater attention [131,144] Among other methods, some authors proposed to superpose the original clinical plan onto the automatically delineated contours [142,143], to use automated planning strategies such as knowledge-based planning (KBP) [131,144] or to conserve the original beam configuration parameters [145]. To our knowledge, this is the first time that an *a priori* MCO auto-planning solution is used for contour evaluation.

We observed in our study that, for most structures, the correction time for DL.1 and DL.2 solutions was <1min (e.g. eyes, brainstem, submandibular glands) and <2min (e.g., mandible, parotid glands) demonstrating significant time saving versus starting from scratch, particularly for the dosimetrist, whose work represented, depending on the AS solution, from 60% to 70% of the total manual editing time. Correcting DL.1 contours was 18min and 9min faster compared with ABAS.3 and DL.2 contours, respectively. Generally, the oral cavity and esophagus took more time to be corrected. For the oral cavity, this may be correlated with the inter-observer variability since the DICE results were consistently smaller for all 3 solutions after the manual corrections. We finally observed that all dose-volume constraints and target objectives were respected in all plans and that manual corrections of the AS contours had no statistically significant impact on the dose distributions. The ΔD_{mean} for the investigated structures were <0.9Gy. Generally, the range of the $\Delta D_{2\%}$ were the highest for the spinal cord and for the brainstem for all the solutions, which may be an important factor in physician's decision when validating the treatment plan. Similar to other studies, for most organs, the difference in the delivered dose was not significant [144][145]. The dose constraints and objectives were respected for all the plans automatically generated and thus, manual correction could be omitted.

Considering the organ position relative to the PTV, high dose differences could be observed when the OAR contour overlapped with the target volumes or was located in their short vicinity. However, this was not always interrelated. This was true for the parotid glands, but for the spinal cord and brainstem, the highest $\Delta D_{2\%}$ were located at a larger distance between the OAR and the PTV (>35mm and >15mm relative to PTV 70Gy and PTV 54.25Gy, respectively). At the same time, at short distances from the PTV (<5mm), the $\Delta D_{2\%}$ in brainstem and spinal cord was <2Gy. One possible reason is that, closer to targets, the AS contours were highly accurate. Although spinal cord and brainstem presented generally good agreement with the manual reference contour, the manual corrections which were nevertheless fast, proved clinically meaningful in certain patients.

Note that this study was deliberately focused on a center-specific approach. The goal was to investigate which of the 6 AS solutions available in our department were more accurate and required less resources in terms of patient data and manpower. In particular, the objective was to evaluate whether, with a relatively small database of homogeneous contoured patients, a center could easily implement an AS solution conformed to its own contouring practices, which, nevertheless, should respect international contouring guidelines. At the same time, we evaluated a solution that was trained on a multi-centric database of contours. Note that the reference contours used in this study belonged to only one expert physician, and also, the manual corrections were done by only one dosimetrist and one physician, both trained by the reference expert. Although the study could benefit from multiple observers involved in manual corrections of the contours, this was, nevertheless, reproducing the clinical workflow of our department. Moreover, the relatively small cohort of the test patients was composed of heterogeneous patients' anatomies and tumor locations, in order to challenge the different AS solutions. Including more patients will definitely strengthen the study, in particular, the statistical analysis. Finally, these findings

suggest that, acknowledging their strengths and limitations, the investigated hybrid ABAS and DL methods improved our clinical workflow.

3.6. Conclusions

DL methods generally showed higher delineation accuracy compared with ABAS methods for AS segmentation of HN OAR. We showed that a DL model can provide accurate contours with a limited training dataset, provided that data comes from a single hospital, and if possible, only one expert physician is involved. Most ABAS contours had high conformity to the reference but were more time consuming than DL algorithms, especially when considering the computing time and the time spent on manual corrections. Finally, even if manual checks and modifications must not be ignored, all AS solutions allow reducing inter-observer variability when physicians perform manual editing of the AS contours.

3.7. Synthesis

The present study provided an extensive comparison between 4 atlas-based and 2 DL solutions for OARs delineation on HN CT images. Their performance was evaluated with regard to several aspects that are relevant when considering an AS solution, namely: patient data resources demand, computational time, geometric accuracy (volumetric overlap and surface distance metric), manual correction time, and dosimetric impact (using auto-planning). The results showed that DL solutions had overall higher accuracy when compared to the ABAS methods. It was also demonstrated that hybrid-ABAS contours had a good agreement with the reference contours and were sometimes better than the DL-based contours. However, when considering the computational time and the time spent on manual corrections, DL solutions were more efficient.

Regarding the data needed for training a DL solution, similarly to other studies in the literature, our study showed that with a limited but more uniform training dataset, a model can achieve consistent results for most of the structures in HN. This can be of great interest for centers that wish to adapt a model to their standard delineation practices. At the same time, we also demonstrated that a model trained with larger amount of multi-centric data can provide good generalizability to new datasets. Performing manual corrections was most efficient on the mono-centric DL model contours (on average 18min). This is particularly relevant for the workload of the dosimetrists. These considerations can help a department choose the best AS solution for their needs in function of their available time and resources. With regards to the dosimetric impact, no statistical differences were observed between the plans created with AS or AS+manually corrections. This was consistent with literature from results that attest that manual corrections could potentially be omitted. Additionally, in our study we investigated the correlation between the organ's position relative to the PTVs and dose differences. The aim was to understand in what situations the correction of an AS contour is particularly important. Unfortunately, we were not able to identify a clear trend of this relationship since high dose differences were not systematically observed in the target's vicinity. The study was original because it evaluated several commercial and non-commercial AS solutions, among which 3 of them have not yet been investigated on HN localization. Another novelty was the use of an MCO auto-planning solution in the dosimetric study for eliminating the planner bias.

We acknowledge limitations of the study in terms of the small cohort of test patients and reference contours coming from one expert physician only. However, the heterogeneous dataset for testing was able to challenge the different AS algorithms. Moreover, this study deliberately focused on a center-specific approach, with the goal to investigate which solution available in the department was more accurate and required fewer resources in terms of patient data and manpower. However, the statistical results could benefit from more patients included in the testing cohort and more observers involved in the manual

corrections task. Future perspectives of the work include increasing the database for training and incorporating cases of both operated and non-operated patients.

Finally, these findings are of great interest for the development of ART workflows for HN patients because they prove increased workflow efficiency when using AS for OAR delineation combined with treatment plans generated using an auto-planning solution. Manual delineation of the primary target remains a time limitation, whereas AS of the lymph node levels that are usually irradiated as secondary target are being discussed in the next chapter of the manuscript.

Chapter 4. Evaluation of different algorithms for automatic segmentation of head-and-neck lymph nodes on CT images

4.1. Abstract

Purpose: To investigate the performance of 4 atlas-based (ABAS) and 2 deep learning (DL) solutions for head-and-neck (HN) elective nodes (CTVn) segmentation on CT images.

Material and Methods: One expert physician delineated bilateral CTVn levels 2, 3 and 4 on 70 HN patients. Ten and 49 patients were used for multi-ABAS atlas library and for training of a mono-centric DL model respectively. Additionally, a commercial multi-centric DL solution was considered. Remaining 21 patients were used for testing. Quantitative evaluation was assessed using volumetric DICE and 95-percentile Hausdorff distance ($HD_{95\%}$). Subjective evaluation was performed on 3 solutions by 4 physicians. One recorded manual correction time. A dosimetric study was conducted using an auto-planning solution.

Results: Overall DL solutions had better DICE and $HD_{95\%}$ results than multi-ABAS methods. No statistically significant difference was found between the 2 DL solutions. However, the contours provided by multi-centric DL solution were clinically better rated by all physicians and were also faster to correct (1min06sec vs 4min10sec, on average). The time needed for manual corrections was larger for ABAS contours (6min31sec). For all methods, decreased accuracy was observed from CTVn2 to CTVn4, and using the contours in treatment planning resulted in underdosage of the elective target volume.

Conclusion: The commercial DL solution provided the best contours when compared to mono-centric ABAS and DL methods for the segmentation of CTVn on HN CT images. Particularly for the CTVn4, manual corrections remain necessary to avoid target underdosage. Finally, AS contours help reducing the manual delineation time.

4.2. Introduction

High accuracy in radiation doses implies highly conformal dose distribution and steep dose gradients, achieved today by using intensity modulated radiation therapy (IMRT) techniques. For head-and-neck (HN) cancer, the fluence optimization is complex because of a large number of organs with strict dose-objectives. To achieve optimal target coverage with minimum normal tissue toxicity, accurate delineation of both organs-at-risk (OARs) and target volumes is a crucial step. Manual contouring is time-consuming and although international guidelines exist [27,328,329], large inter and intra-observer variation are observed [128,318,330] that can negatively impact patient doses [137,142]. To help in organs differentiation and increase the images' contrast, the patients should be injected with an iodine solution before the simulation computer tomography (CT) scans[25]. To reduce the delineation time, improve consistency and accuracy of volume definition, automatic segmentation (AS) solutions received great interest[132,133,331]. AS methods for HN OARs contouring were investigated in many studies, whereas fewer studies were focused on the clinical target volumes (CTV). Whereas important anatomical variations make gross tumor volumes difficult candidates for AS, the HN lymph nodes levels (LN) have well-established anatomical borders [27,329] and are often irradiated as secondary prophylactic nodal target (CTVn).

Among the AS solutions, atlas-based AS (ABAS) methods are attractive as they require only one or few (multi-ABAS) patients as prior information (in form of an atlas library), but they are limited to the range of patient anatomical representation. Few studies have demonstrated the superiority of multi-ABAS vs single-ABAS strategies for CTVn segmentation [152,332–334]. It was shown that using 11 vs 1 atlas enabled decreasing the manual delineation time by 20% [332]. In another study, the range of reported overlap between ABAS and reference contours, was 29%-78% depending on the CTVn level considered [334]. One multi-ABAS study (N=10 atlases) evaluated dosimetric plan quality when using AS contours obtained with

a commercially available solution (ABAS, Elekta AB), and demonstrated that despite >80% contour overlap, non-edited CTVn contours can cause large underdosage in target volumes [137]. Hybrid approaches combining multi-ABAS and machine learning features have also been explored. Qazi et al. evaluated a model-based algorithm (N=15 atlases) and achieved expert contours overlap of 74% (LN level 1-4) [169]. Their results were superior to Chen et al. [335] who created an active shape model (N=14 atlases) that reached 69% overlap (one nodal volume covering LN 2-4) and superior to that of Gorthi et al. [336] with an active contour-based model (N=9 atlases) that reached maximum 58% overlap (individual CTVn levels 1-6).

Alternatively, deep learning (DL) solutions should increase accuracy and efficiency in AS at the cost of more efforts involved in gathering and curating manual contours databases for training. Promising results were obtained particularly for OARs in HN patients and several solutions are commercially available [3,133,143,163,190,217,218]. From the few studies evaluating their accuracy in segmenting HN CTVn, Wong et al. investigated a commercial DL-based segmentation software (Limbus Contour build 1.0.22) trained with publicly available annotated data (on average 328 CT scans/organ) [218]. One single CTVn volume including 6 LN was auto-segmented. The overlap with the experts' contours was 72% which was inferior to the inter-observer variability (IOV) assessed (79%). Another study investigated a 3D-convolutional neural network (CNN) trained on 69 patients (mono-centric data), for segmenting 10 separated CTVn levels, with the exception of levels 2-4 which formed one volume [337]. The contour overlap against 2 experts ranged between 46-82%, in function of the considered CTVn level. The manual delineation time was reduced from 52 to 35min using AS contours. Moreover, it was shown that using the AS solution enabled to significantly improve the IOV (92.2% vs 79.8%). Lastly, Strijbis et al. [338] trained 3 different Unet networks on 70 patients for segmenting individual volumes for CTVn 1-5. They showed that an ensemble of networks provided the best results with >85% contour overlap for the CTVn 1, 2 and 3, but <72% for CTVn 4 and 5.

In this context, the objective of the present study was to evaluate the performance of 4 ABAS and 2 DL solutions for the individual segmentation of 6 CTVn volumes, explicitly the left (L) and right (R) LN levels 2 (CTVn2), 3 (CTVn3) and 4 (CTVn4). This follows on from the formerly performed work on HN OARs segmentation [3]. Five of the solutions were investigated for the first time on HN CTVn segmentation. One hybrid-ABAS and one center-specific DL solution not commercially available were compared to three and one, commercially available multi-ABAS and DL solutions, respectively. All 6 solutions were evaluated based on geometrical accuracy. A clinical scoring of the contours was then performed by 4 expert physicians on the 3 most accurate AS solutions. For one physician, the time spend on correcting the contours was measured. Lastly, an auto-planning solution based on a priori multicriteria optimization (MCO) algorithm was used to generate treatment plans when using manual and AS CTVn contours.

4.3. Materials and methods

Patient data

Seventy HN cancer patients treated with radiation therapy between 2018 and 2022 were included in the study, which was approved by the hospital ethics committee. For each patient, CT scan acquisition was performed after 2 injections of iodine contrast agent following national recommendations [25]. Bilateral CTVn 2, 3 and 4 were then manually delineated according to international delineation guidelines [27] by a senior expert physician, on 512x512 and 2mm-thick CT-slices. Forty-nine non-operated patients were used to train a mono-centric DL model to automatically segment the 6 target volumes. Based on their body mass index (BMI, from 19.9 to 26), 10 of these patients were subsequently used to form an atlas library for the multi-ABAS solutions, that covers a large variety of patient anatomies. Identical atlas libraries were

created within MIM-Maestro (MIM-Software; Cleveland, USA) and the research version of ADMIRE software (ADMIREv3.41, Elekta AB; Stockholm, Sweden). Conversely, the training of the second DL solution, was exclusively handled by the vendor (Therapanacea, France), incorporating big amount of multi-centric patient data (>1000). The remaining 21 patients with different tumors and anatomies (BMI 17.9 – 33.7), were used for testing of the 6 AS solutions. In addition to reference contours for CTVn, the test cohort (Table 4.3-1) included expert delineations for OARs and primary tumor volumes.

Table 4.3-1 Characteristics of the testing cohort used for evaluation of the AS solutions

	Tumor localization	TNM	Treated CTVn	BMI
Patient 1	Rhinopharynx	T3 N1 M0	2 – 4 L/R	19
Patient 2	Rhinopharynx	T2 N2 M0	3-4 L/R	21.4
Patient 3	Rhinopharynx	T1 N1 M0	2-4 L, 3-4 R	33.7
Patient 4	Hypopharynx	T4 N3b M0	2-4 L	18.6
Patient 5	Hypopharynx	T1 N1 M0	2 – 4 L/R	23.4
Patient 6	Hypopharynx	T4b N0 M0	2 – 4 L/R	19
Patient 7	Larynx	T2 N0 M0	2 – 4 L/R	24
Patient 8	Larynx	T2 N0 M0	2-4 R	30.4
Patient 9	Larynx	T3 N0 M0	2 – 4 L/R	21.8
Patient 10	Larynx	T4a N0 M0	2 – 4 L/R	21.5
Patient 11	Tonsils	T2 N0 M0	2-4 L	24
Patient 12	Tonsils	T2 N1 M0	2 – 4 L/R	25.5
Patient 13	Tonsils	T1 N1 M0	2-4 L, 3-4 R	17.9
Patient 14	Tonsils	T2 N1 M0	2-4 R	23.8
Patient 15	Tonsils	T2 N0 M0	2 – 4 L	32.4
Patient 16	Tonsils	T3 N0 M0	2 – 4 L/R	31.5
Patient 17	Tonsils	T2 N1 M0	2 – 4 L/R	32.6
Patient 18	Oral cavity	T4a N2c M0	2-4 R, 4 L	30.2
Patient 19	Oral cavity	T3 N1 M0	2-4 L, 3-4 R	24.2
Patient 20	Oral cavity	T2 N0 M0	2 – 4 L/R	22.1
Patient 21	Oral cavity	T3 N2a M0	2 – 4 R	21.0

Automatic segmentation solutions

Three multi-ABAS solutions integrated in the research version of Monaco treatment planning system (TPS) (Monaco 5.59.11 with ADMIREv3.41) and another one available in MIM-Maestro (MIM Software Inc., Cleveland, OH) were investigated (Table 4.3-2). Two DL solutions were considered, one mono-centric (data from this study) solution and one commercially available multi-centric solution (Table 4.3-2).

All AS solutions have been fully described in a previous work [3]. Briefly, ABAS.1 uses a traditional method for atlas fusion based on expectation-maximization algorithm. ABAS.2 uses voxel intensity information to obtain a weighted average of the atlases' contours. ABAS.3 algorithm trains a voxel classifier on the fly using the registered atlases as training data. Lastly, ABAS.4 performs the voxel annotation based on labels predicted by majority of the atlases. For the 3 ABAS solutions used in ADMIRE software, for each test patient, a reference atlas was selected from the library, upon the closest BMI. Distinct process was followed in MIM-Maestro, where to create the atlas library, one atlas was chosen as template patient (based on BMI) and was registered to the 9 remaining atlases.

Among the DL solutions, DL.1 is a CNN where the high-resolution image features captured in the encoding part are preserved with the help of short-range connectors in the decoding part for a label map corresponding to the input image size. The DL.2 solution uses a set of organ-specific networks with an

original combination of data-driven and decisional artificial intelligence that enforces anatomical consistency. Its training included annotated patients from multiple centers and was handled by the vendor.

Table 4.3-2 Characteristics of the 6 automatic segmentation solutions investigated in the study.

	Solution name	Software Vendor	Nr. of Atlases/ Nr. of training patients	Commercially available
1. ABAS.1	STAPLE [156]	ADMIREv3.41 (Elekta AB, Stockholm, Sweden)	N=10	Yes
2. ABAS.2	Patch Fusion [339]			Yes
3. ABAS.3	Random Forest [174]			No
4. ABAS.4	Majority Voting [162]	MIM Maestro 7.0 (MIM Software Inc., Cleveland, OH)	N=10	Yes
5. DL.1	ADMIRE-DL [124,190]	ADMIREv3.41 (Elekta AB, Stockholm, Sweden)	N=49 mono-centric patient data	No
6. DL.2	ART-plan Annotate [206,217]	ART-plan (Therapanacea, France)	N>1000 multi-centric patient data	Yes

Abbreviations: STAPLE = Simultaneous Truth and Performance Level Estimation

Geometric evaluation

The quantitative evaluation of the 6 AS solution was performed per CTV_n level and per their union, based on volumetric DICE coefficient and 95-percentile Hausdorff Distance (HD_{95%}) [135], similar to a previous work [3].

Clinical acceptability assessment and time required for manual editing

The union of the bilateral CTV_n contours (CTV_n_union) was further examined on 12 patients for the most accurate 3 solutions, in terms of clinical acceptability and manual correction time. First, a blinded evaluation was made by 4 physicians choosing for each CTV_n_union one of the following options:

- a) clinically acceptable without corrections
- b) clinically acceptable with minor corrections
- c) clinically acceptable with major corrections
- d) not acceptable for clinical use

Then the AS contours were manually adjusted by one of the physicians on Elekta ProKnow® (Elekta AB, Stockholm) platform and the time spent on corrections was recorded for each of the 3 solutions.

Dosimetric end-points using auto-planning solution

For the 12 patients, 4 treatment plans were calculated automatically using *mCycle* auto-planning solution (Monaco 5.59.11, Elekta AB; Stockholm, Sweden). All plans were designed using 2 arcs and a simultaneous integrated boost technique to deliver 70Gy to the primary planned target volume (PTV_70Gy) and 54.25Gy to the prophylactic nodal target (PTV_54.25Gy), in 35 fractions of 2Gy. The reference plan was created using exclusively manually delineated contours of OARs and PTVs. The 3

experimental plans were created by replacing the manual CTVn contour with CTVn contours obtained by ABAS.2, DL.1 and DL.2 solutions. The PTV_54.25Gy was created for each plan from the union of CTVn levels and the prophylactic target plus additional 4mm margins. The resultant 4 dose distributions were all analyzed on the reference manual contours. From the dose-volume histograms (DVHs) clinically relevant dosimetric endpoints were extracted according to the French Society of Radiation Oncology recommendations [25].

Statistics

For all 6 solutions and for each lymph node level, Kruskal-Wallis test was performed to assess if the methods were statistically different. Furthermore, post-hoc Dunn's with Bonferroni correction for multiple testing was performed to detect between which pairs of algorithms the differences were statistically significant. Similarly, dose differences between the treatment plans were statistically evaluated. The statistical tests were performed using Python 3.8 with level of significance set <0.05 .

4.4. Results

Computational time for one patient was about 6min, 9min and 10min for ABAS.1, ABAS.2 and ABAS.3, respectively. For ABAS.4, the segmentation time was approximately 1 min, whereas creation of the atlas library (registering of the atlases to the library) took around 7min. DL.1 and DL.2 provided segmentation in <1 and <2 min, respectively.

DICE and $HD_{95\%}$ results obtained for each CTVn level and for CTVn_union are presented in Figure 4.4-1. Overall DL solutions (DICE: 0.63-0.87) were more accurate than ABAS methods (DICE: 0.49-0.79), with no statistically significant difference between DL.1 and DL.2 ($p>0.2$). However, DL.1 generally performed better on CTVn4 than DL.2, whereas DL.2 had the lowest $HD_{95\%}$ distances ($\leq 6.4\text{mm} \pm 5.1$) among all the methods.

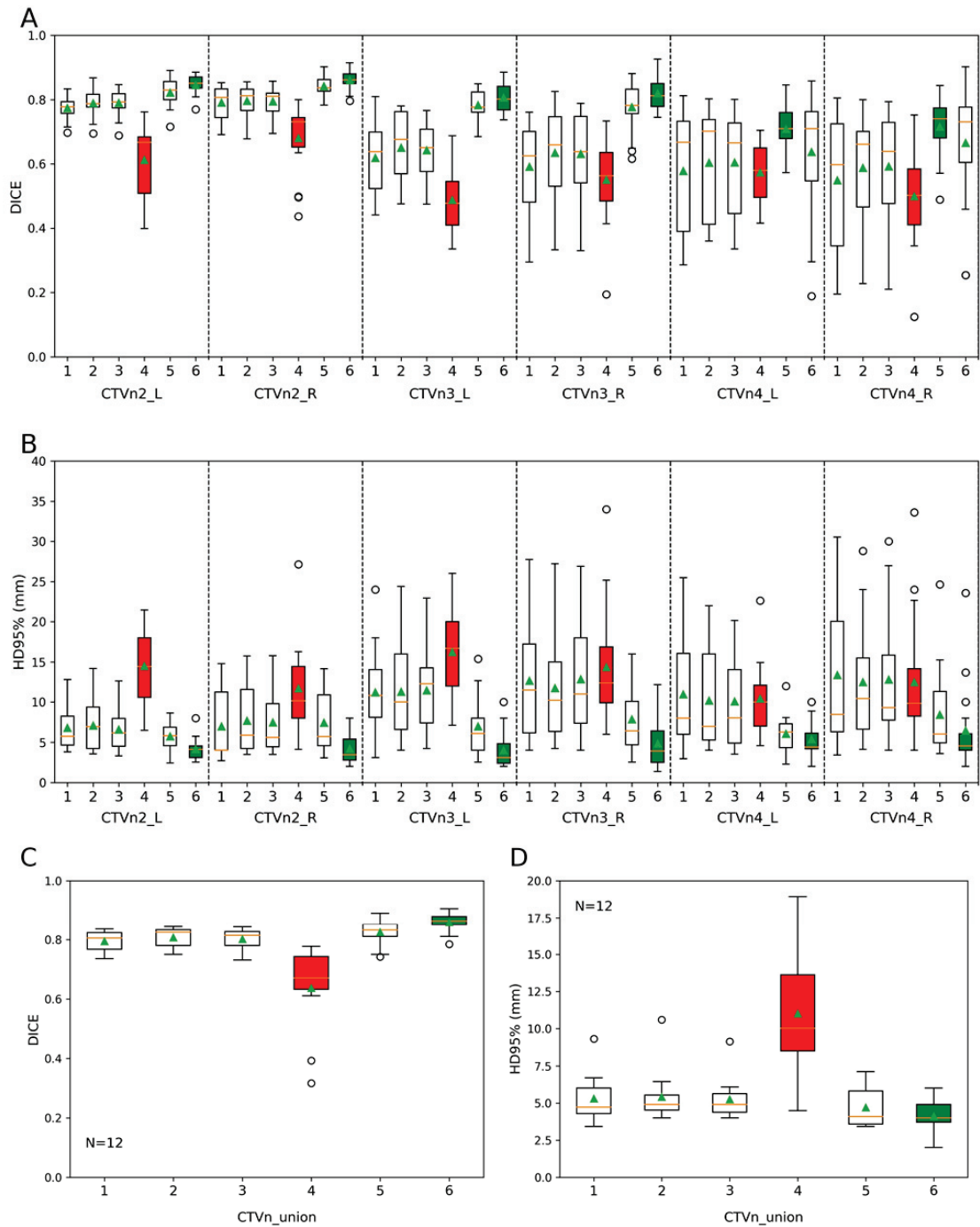


Figure 4.4-1 Geometric evaluation results of the four ABAS and two DL solutions; 1 = ABAS.1, 2 = ABAS.2, 3 = ABAS.3, 4 = ABAS.4, 5 = DL.1, 6 = DL.2. Panel A and B show DICE and HD_{95%} results per individual CTVn level; Panel C and D show DICE and HD_{95%} results of the CTVn union volume; in red and in green are highlighted the worst and the best results, respectively determined by the mean value of DICE/HD_{95%}; in the boxplots, the orange line represents the median, the green triangle indicate the mean value and the circles represent outliers.

Considering the multi-ABAS solutions, no statistically significant difference was observed between ABAS1, ABAS.2 and ABAS.3 ($p=1$). ABAS.4 provided the worst results but differences in both DICE and $HD_{95\%}$ compared to other ABAS methods were statistically significant only on CTVn2_L contours ($p<0.005$).

Differences were statistically significant between DL.1 algorithm and ABAS.1, ABAS.2 and ABAS.3 solutions only in DICE for the CTVn3_L/R ($p<0.03$). Compared with ABAS.4, DL.1 had statistically better DICE results for all the CTVn levels ($p<0.04$) and statistically better $HD_{95\%}$ results for CTVn2_L ($p<0.001$) and CTVn_3L/R ($p<0.04$).

Similarly, DL.2 had significantly better DICE results compared with ABAS.1, ABAS.2 and ABAS.3 for CTVn2_L/R ($p<0.03$) and CTVn3_L/R ($p<0.001$) and significantly better $HD_{95\%}$ for CTVn3_L/R ($p<0.002$) and CTVn4_R ($p=0.02$). Moreover, compared with ABAS.4, DL.2 had significantly better DICE results for all levels ($p<0.02$) but CTVn4_L ($p=0.5$) and significantly better $HD_{95\%}$ for all node levels ($p<0.007$).

An additional geometric analysis of the CTVn_union resulted in increased conformity to the manual reference, particularly for the multi-ABAS solutions, for which the contour unification enabled DICE results to reach values up to 0.81 (ABAS.2). Finally, DL.2 solution obtained the best conformity to the union of the reference contours (meanDICE: 0.86 ± 0.03 ; mean $HD_{95\%}$: $4.1 \pm 1.2\text{mm}$) (Figure 4.4-1 Panel C and D). Moreover, the blinded study results (Figure 4.4-2) showed that all physicians rated DL.2 contours as clinically acceptable without or with only minor corrections. Contrarily, none of ABAS.2 contours were accepted without corrections, and only one physician accepted few contours from DL.1 without corrections. Moreover, some ABAS.2 and DL.1 contours were also rejected by two physicians.

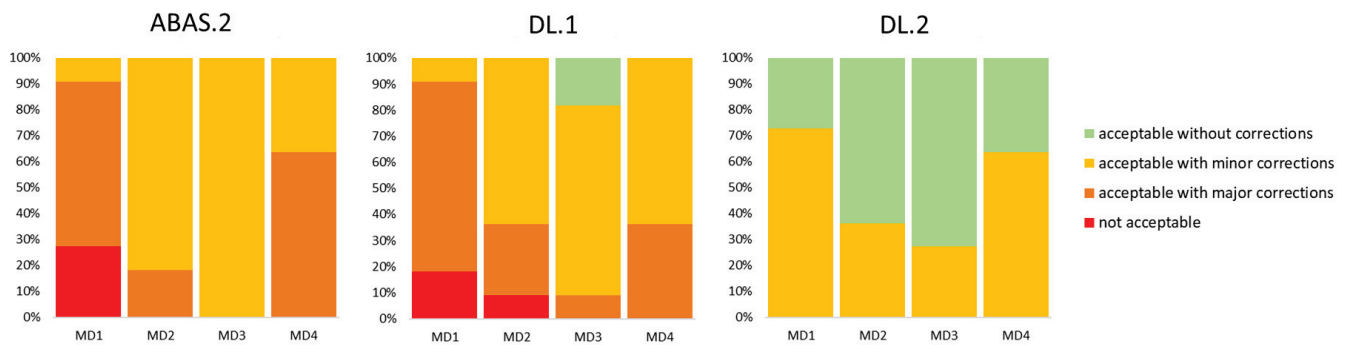


Figure 4.4-2 Expert evaluation on the CTVn_union from ABAS.2, DL.1 and DL.2 solutions

Furthermore, manual correction time was in average 6min31sec, 4min10sec and 1min06sec for ABAS.2, DL.1 and DL.2 respectively, and contours' accuracy improved significantly only for ABAS.2 solution ($p<0.001$) (Figure 4.4-3).

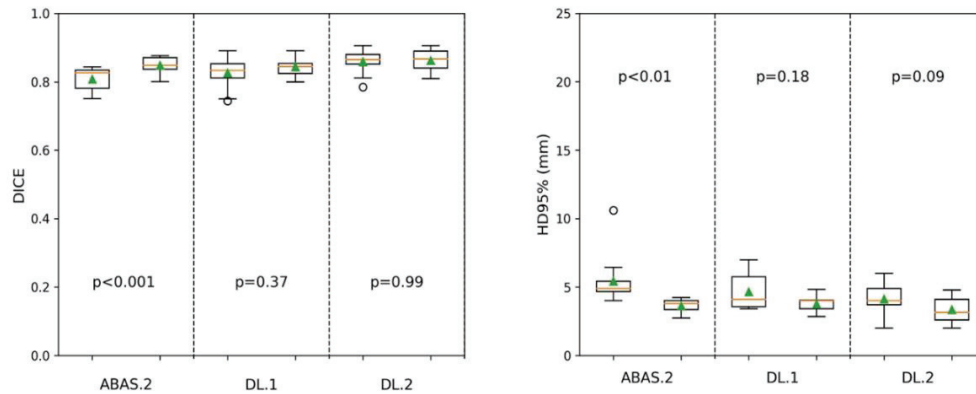


Figure 4.4-3 DICE and HD_{95%} results before and after performing manual corrections

Regarding the dosimetric study, in all the plans, the dose objectives for PTV_70Gy and the dose constraints for the OARs were achieved with no statistically significant dose difference observed ($p > 0.1$). Conversely, PTV_54.25Gy which contained the CTV_n experimental volumes, experienced significant loss in coverage compared to reference plans for all solutions ($p < 0.01$). Overall, the largest dose variations were observed on CTV_{n4} where ABAS.2 and DL.1 achieved sufficient coverage ($V_{95\%} > 95\%$) while DL.2 did not (92.3%). Figure 4.4-4 illustrates results of the dosimetric study per PTV and per CTV_n level. Between the reference and both ABAS.2 and DL.1 experimental plans, statistically significant dose differences were identified for CTV_{n2} (in both $V_{95\%}$ and $D_{98\%}$, $p < 0.002$), CTV_{n3} ($V_{95\%}$, $p < 0.001$) and CTV_{n4} ($V_{95\%}$ and $D_{98\%}$, $p < 0.006$). Similarly, dose differences were significant between reference and DL.2 experimental plan for the dose distribution to CTV_{n2} ($V_{95\%}$, $p = 0.001$) and CTV_{n4} ($V_{95\%}$, $D_{98\%}$ and $D_{50\%}$, $p < 0.007$).

One patient case that exhibited large dose differences (Patient 4) is illustrated in Figure 4.4-5, with a visual representation of the CTV_n_union contours variation presented in panel A, and important loss in PTV_54.25Gy coverage illustrated in panel B and C. Furthermore, a summary of studies from literature on AS for HN CTV_n volumes is presented in Table 4.4-1.

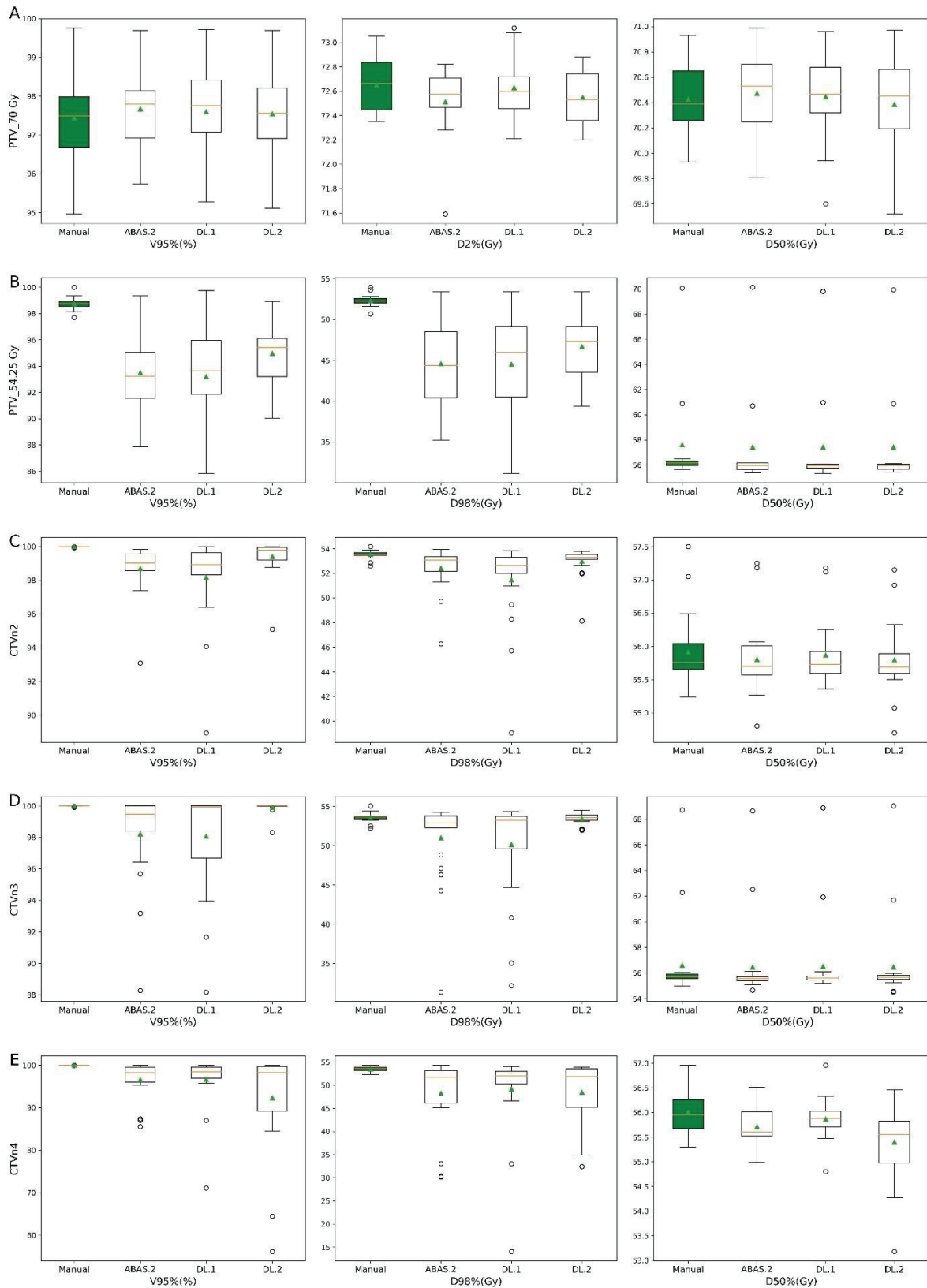


Figure 4.4-4 Dosimetric comparison between reference plan and the 3 experimental plans. Panel A: dosimetric impact on the primary target (PTV_70Gy) evaluated in terms of $V_{95\%}$, $D_{2\%}$ and $D_{50\%}$; Panel B: dosimetric impact on the nodal target (PTV_54.25Gy) evaluated in terms of $V_{95\%}$, $D_{98\%}$ and $D_{50\%}$; Panel C, D and E: dosimetric impact on the bilateral CTVn levels 2, 3 and 4 respectively, evaluated in terms of $V_{95\%}$, $D_{98\%}$ and $D_{50\%}$. With green are highlighted the reference results from the reference plan created exclusively with manual contours.

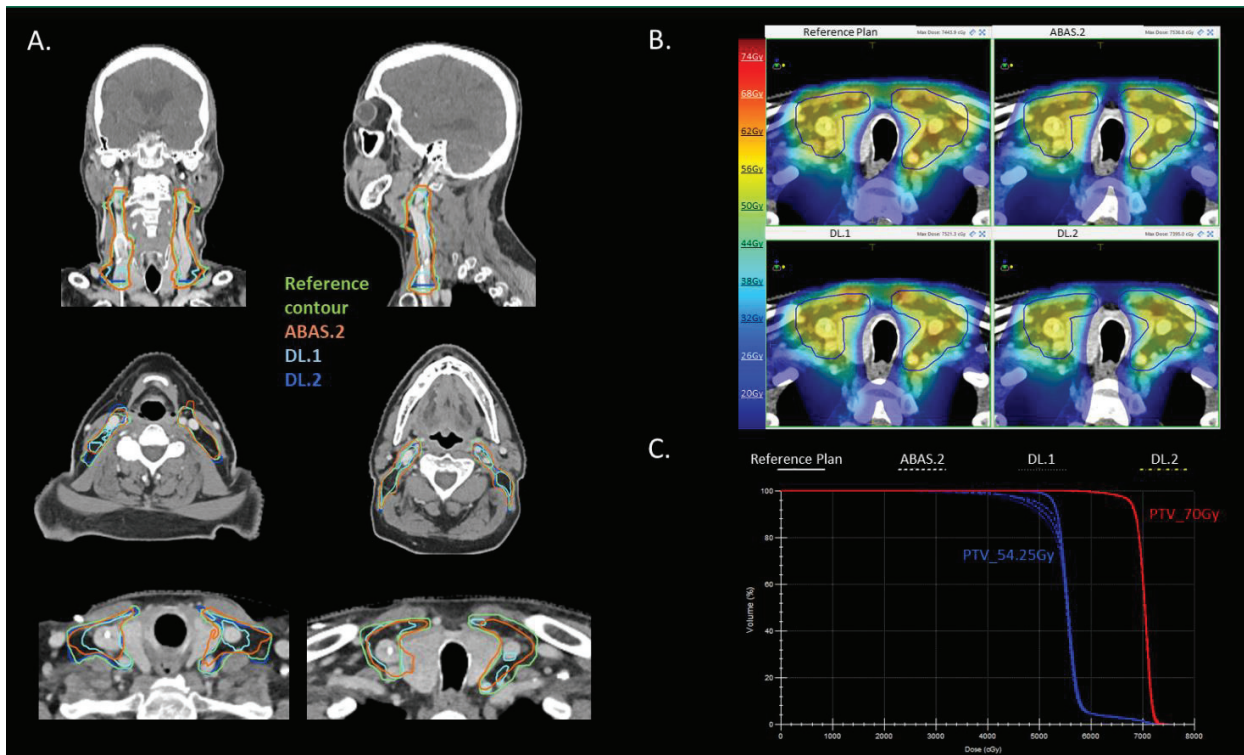


Figure 4.4-5 Illustration of a patient case (Patient 4). Panel A: Visualization of AS contours (CTV_n_union) from the 3 solutions: ABAS.2 (orange), DL.1 (light blue) and DL.2 (dark blue) solutions in contrast with the manual reference (green). Panel B: experimental radiotherapy plans (ABAS.2, DL.1 and DL.2) compared to the reference plan; the reference manual contour of the PTV_{54.25Gy} is displayed; significant underdosage to the secondary target is observed particularly for the plan created with DL.2 contours. Panel C: From dose volume histogram, a negative impact on elective target is observed in the plans generated with CTV_n AS contours

Table 4.4-1 Studies on automatic segmentation for CTVn volumes

Author (year)	Approach	Algorithm/Solution	CTVn levels	Nr of atlases/ Nr of training data	Test cohort	Manual correction time/ time reduction	Comments
Daisne et al. (2013) [334]	Single-ABAS	(iPlan® Net, Brainlab, AG, Germany)	Levels 1-6 (separate volumes)	N=1	N=20	22min for manual corrections, 29%-time reduction	image registration and recognition by anatomical key structures
Voet et al. (2011) [137]	Multi-ABAS	STAPLE (ABAS Elekta)	Level 1-5 (one volume)	N= 10 atlases	N=9 patients	/	editing of neck is essential to avoid underdosage in target volumes
Stapleford et al. (2010) [160]	Single and Multi-ABAS	VelocityAI and STAPLE fusion	Levels 1-6 (one volume)	N=5	N=5	35% time reduction	DICE=0.76 32% of the contours clinically acceptable without corrections, overall AS contours were larger IOV= 0.79
Sjoberg et al. (2013) [332]	Single and Multi-ABAS	Cross correlation and probabilistic weighting (VelocityAI)	One volume lymph nodes region	N =1 atlas N=11 atlases	N=10	29% and 49% time reduction with single ABAS and multi-ABAS, respectively	DICE=0.65 single atlas DICE=0.71 fused atlases After corrections DICE 0.82-0.84
Teguh et al. (2011) [157]	Single and multi-ABAS	MI and STAPLE (ABAS Elekta)	Level 1-5 (in one volume)	N=1 N=10 atlases	N=12	from 180min to 66 min time reduction (OARs and CTVn)	DICE ≤ 0.67
Gorthi et al. (2009) [336]	Hybrid-ABAS	active contour-based atlas registration	Level 1-6 (separate volumes)	N=9	N=10	/	DICEs ≤ 0.58
Chen et al. (2010) [335]	Hybrid-ABAS	registration and active shape models (ASM)	Level 2-4 (one volume)	N=14	N=15	/	DICE=0.63 using average patient DICE=0.69 using ASM method with structure constrain
Qazi et al. (2011) [169]	Hybrid-ABAS	Model-based	Level 1, and level 2-4 (one volume)	N=15	N=10	/	DICE=0.74 (about 10 min/patient to generate the contours)

Men et al. (2017) [200]	Deep Learning	DCNN vs VGG-16	One nodes volume including level 2	N=184	N=46	/	DCNN > VGG-16 DICE = 0.83 for CTVn volume
Wong et al. (2021) [340]	Deep Learning	Limbus Contour	Levels 1-5 (one volume)	N=328 on average/organ (public databases)	N=20	0.6 vs 26.6 min (OAR and CTVn)	DICE = 0.72, HD _{95%} : 10.93mm IOV = 0.79 with HD _{95%} : 6.75mm
Van der Veen et al. (2020) [337]	Deep Learning	CNN	10 Levels (separate volumes, Level 2-4 one volume)	N=69 training patients Monocentric data	N= 16	35 vs 52 min over all CTVn 8 vs 15 min for case-relevant CTVn	IOV = 0.79 AS+corrections IOV = 0.92
Strijbis et al. (2022) [338]	Deep Learning	3D-UNet	Levels 1-5 (separate volumes, and unified levels)	N=70	N=10	/	DICE: 0.74 – 0.86
Costea et al. (2022)	Multi-ABAS	ABAS.1(Elekta AB) ABAS.2 (Elekta AB) ABAS.4 (MIMSoftware)	Level 2-4 (separate volumes)	N=10 atlases	N=21	6min31 for ABAS.2 4min10 for DL.1 1min06 for DL.2	DL>multi-ABAS DL.2 > DL.1 ABAS.2>ABAS.3>ABAS.1>ABAS.4
	Hybrid-ABAS	ABAS.3 (Elekta AB)	And their union	N=10 atlases			DICE CTVn2 and CTVn3 > 0.81
	Deep Learning	DL.1 (Elekta AB) DL.2 (ART-plan)		N=49 (monocentric) N>1000 (multicentric)			DICE CTVn4 <0.72 Underdosage in CTVn4

Abbreviations: ASM= active shape models, MI = mutual information, IOV=interobserver variation

4.5. Discussion

In this study, we evaluated for the first time the performance of ABAS.2, ABAS.3, ABAS.4, DL.1 and DL.2 solutions. We observed that overall DL solutions had better accuracy compared with multi-ABAS methods for CTVn segmentation on HN CT images. With regard to the geometric indexes, the 2 DL solutions were not statistically different. In general DICE results were better for DL.2 on CTVn2 and CTVn3, and better for DL.1 on CTVn4. When evaluating CTVn_union, DL.2 provided better conformity to manual reference and 61% were considered clinically acceptable without correction while the rest requested only minor corrections which took about 1min/patient for one of the physicians involved. Conversely, DL.1 contours needed minor or major correction in 94% of the cases which resulted in more important manual correction times (4min43sec/patient). DL.1 model was trained with relatively small number of patients (N=49) delineated exclusively by one expert physician, which ensured data uniformity. Similar with the previous work on OARs, we showed in this study that accurate results can be obtained with a limited but uniform training database, which can encourage centers to create custom-made models adapted to their standard delineation practices. A similar mono-centric training data approach (N=69) was followed by van der Veen et al. for the segmentation of 10 CTVn levels [337]. For the union of LN 2-4, they found 76% and 82% overlap to manual contours from 2 observers, whereas an 83% overlap was obtained with DL.1 solution in our study. They showed that the IOV was improved when corrections were made on the DL-based contours. Contrarily, DL.2 solution was trained with much more patients coming from multiple centers, which also included segmentations from the same reference physician involved in the training database of DL.1. Compared to the mono-centric solution, DL.2 obtained better overlap for CTVn_union (0.86), which indicated a good generalizability of the model.

Regarding multi-ABAS methods, this study showed that good results can be obtained when using a library of only 10 patients, and atlas selection strategy based on the closest BMI. Moreover, performing the union of the CTVn levels, enabled an overlap ≥ 0.80 which suggested that most of the AS contour discrepancies happen at the junction of the level. When considering both the computational and the manual correction time, approximately 15min were needed to segment a new patient with ABAS.2 solution. Contrary to the results observed previously on OARs segmentation, the superiority of the new ABAS.3 solution over the commercial ABAS.1 and ABAS.2 solutions was not demonstrated for the CTVn segmentation [3].

According to recently published guidelines, studies should also report findings from dosimetric evaluation of treatment plans created with a new AS methods [207]. From the limited literature, only one study performed a dosimetric evaluation and attested that editing ABAS contours of neck CTVn was required to avoid large underdosage in the target volumes [341]. Moreover, the heterogeneity of studies' design makes comparison rather difficult. Since big amount of work is required to create reference data bases, some studies considered a total volume of the CTVn whereas others considered independent contours per CTVn.

To our knowledge, the present study investigated for the first time 5 AS methods for segmenting 3 CTVn levels separately. Additionally, auto-planning was used to assess dosimetric consequences of using AS contours from one multi-ABAS and 2 DL solutions. This allowed decreased labor and IOV, and to focus on the dosimetric effect coming from the CTVn contour only. Overall the results were similar among the AS methods, and showed no significant impact on the primary PTV and OARs. However, despite the use of a CTV-to-PTV margin of 4mm, significant underdosage on the nodal PTV was observed for all the AS solutions, which was consistent with the literature [341]. The effect was more pronounced on the CTVn4 level which could be related with the higher discrepancies previously identified in the geometrical overlap. Moreover, the blinded study showed that majority of AS contours were clinically acceptable with only minor corrections. When considering both computational and manual correction time, substantial time savings can be achieved by using DL solutions. In our study, one single physician performed the reference

contours and another physician performed the manual corrections. A good contour overlap (DICE=0.85) between the two experts was observed when manual corrections were performed on the AS contours. IOV between manual delineations among the experts was not assessed in this study. However, other study showed that performing manual adjustments on AS contours enabled to improve IOV [337]. While differences in DICE and HD_{95%} were not statistically significant between DL.1 and DL.2, DL.2 contours were better rated by all the 4 physicians, and time for correcting the contours was significantly lower. Therefore, the interplay between the training cohort size and a DL model architecture could be further investigated by training DL.1 on a larger cohort (N>50 patients). In the previous work, 63 patients were used for training the same model on OARs, which provided consistent result over the majority of structures. While on CTVn delineation, DICES_{≥0.82} were obtained for CTVn2, more training data could potentially improve the accuracy on CTVn3 and CTVn4. At the same time, DL.2 was trained on large database of patients and the overlap for CTVn4 was inferior to DL.1 model. Overall, both multi-ABAS and DL results showed decreased accuracy from CTVn2 to CTVn4 which is consistent with the literature [334,338].

4.6. Conclusion

DL solutions were faster and more accurate than multi-ABAS methods for CT-based AS of HN CTVn levels. The multi-centric DL model provided high quality contours leaving only 1min for manual corrections. Similar contours were obtained with the mono-centric model trained with <50 patients, but 4min were required for manual adjustments. With only 10 atlases, ABAS methods can provide good conformity to reference contours, but with decreased workflow efficiency. A decrease in contour accuracy was observed from CTVn2 to CTVn4. Finally, manual corrections are still needed to preserve the coverage of the elective target.

4.7. Synthesis

The present study represents the third contribution of this PhD project and is a continuation of the previous work on AS of HN OARs. We compared the 4 ABAS and the 2 DL solutions for the delineation of separate CTVn levels 2,3 and 4, that are typically irradiated as secondary targets in HN localization. Additionally, we evaluated the accuracy of their union as this volume is generally used in for the treatment planning. We analyzed the contours in terms of: computational time, spatial accuracy, clinical acceptability (assessed by 4 physicians), time needed to perform manual corrections (performed by one of the observers) and dosimetric consequences when using the AS contours in treatment planning.

Similar to the previous study on OARs, we observed that DL solutions had better accuracy when compared with ABAS methods for CTVn segmentation on CT images. However, ABAS methods were also able to reach DICES_{>0.80} when considering the union of the levels. This revealed an important observation, the fact that significant contour inaccuracies happen at the border between the levels, at the upper and lower extremities. The main advantage of an ABAS method was the small amount of resources required. However, the drawback was the computational time (6-10min per patient, for a library of 10 atlases). Contrarily, DL methods were trained on larger databases (49 patients for DL.1 and >100 patients for DL.2 solution) but the contours were generated faster (<1min and <2min, respectively). Moreover, all the contours generated by the commercial multi-centric DL model were deemed clinically acceptable without or with only minor corrections, that were on average 1min06sec per patient. These remarkable results constitute key components for highly efficient workflows. Compared with the mono-centric DL solution no statistically significant differences were identified between geometrical accuracy indices (DICE and HD_{95%}). However, the manual correction time was larger (4min10sec vs 1min06sec, on average). Overall, the clinical evaluation from the 4 experts, led to similar conclusions for DL.1 and ABAS.2 contours. However, performing the manual corrections was more time-consuming on the ABAS.2 contours (6min31sec vs 4min10sec). Notably, a good agreement (DICE=0.85) among the reference expert and the other physician performing the manual corrections was observed, which was higher than expert IOV

previously reported in literature studies [218,337]. An IOV among experts was not conducted this study, however it constitutes future perspectives of the work. Evaluating contours from multiple observers, manually delineated from scratch and manually adjusted AS-based contours, would enable to assess the usefulness of AS in improving the consistency of delineation practices.

When analyzing the dose distributions, significant underdosage in the secondary PTV was observed regardless of the solution used. The loss in coverage was particularly detected on the CTVn4 which could be related with the discrepancies previously identified in the geometrical overlap. For both ABAS and DL methods, DICE results for CTVn4 were ≤ 0.72 , which was consistent with results from the literature [334,338]. For these contours, we recommend greater attention, before using them in treatment planning.

To best of our knowledge, the present study investigated for the first time 5 AS methods (3 multi-ABAS and 2 DL) for segmenting 3 separate CTVn levels. Additionally, auto-planning was used to assess dosimetric consequences which enabled decreased labor, no planner IOV, and an isolated effect coming from the CTVn contour only. Future perspectives of the work include to increase the DL model training database (perhaps also including operated patients), as well as the testing patient's cohort. Furthermore, a new network framework could be developed that combines anatomical landmarks from CTVn delineation guidelines in order to guide the node levels predictions.

Finally, the study demonstrates that AS methods for CTVn can be integrated in the RT workflows to reduce the time spent on manual delineation. At this moment, the proposed combination of AS for OARs and CTVn, together with auto-planning solutions could be used to improve efficiency of complex HN cases. Only the primary tumor volume would remain to be manually delineated by physicians. However, to enable ART, intra-fractional anatomical variations must be considered. In this regard, the next chapter will discuss different methods for synthetic CT image generation from daily CBCT images, that can be used to calculate plans on the anatomy of the day.

Chapter 5. Evaluation of different methods for synthetic CT image generation from daily CBCT images

5.1. Introduction

Currently, RT for complex HN cases is managed by combining IMRT and robust IGRT strategies. This ensures adequate target coverage and sufficient OARs sparing. Delivery of high accuracy radiation doses is however sensitive to patient anatomical variations that typically happen due to tumor shrinkage or weight loss. The use of kV-CBCT images acquired at the beginning of each treatment fraction, can enable to assess the delivered dose based on the anatomy of the day and to trigger re-planning. However, the CBCT image quality is rather “poor” (low soft-tissue contrast) and contains many image artifacts that result in inconsistent HU values. Moreover, CBCT image has a limited field-of-view (FOV) thus does not cover the whole patient contour as defined in the planning-CT (pCT) image. Finally, daily CBCT based dose calculation uncertainties are difficult to assess due to missing ground truth (doses on daily CT images).

Different approaches to overcome these issues exists [342]. A summary table is presented in the first chapter of this manuscript (Table 1.10-2). To assist ART, DL methods are promising, because they offer fast conversion of CBCT into synthetic-CT (sCT) images, and demonstrated image quality close to that of pCT [310,317].

The purpose of this study was to investigate different methods that enable dose calculations from daily CBCT images. A DL method for sCT generation was proposed by Elekta based on a cycleGAN architecture trained on unpaired CT and CBCT data. Additionally, three other methods from literature have been investigated and compared: CBCT-specific HU-ED curve, 3-class density assignment method (3C-DAM) and a deformable image registration algorithm (DIR).

5.2. Materials and methods

Patient data

Twenty-five patients following a specific adaptive IGRT protocol were selected for this study (Table 5.2-1). Each of them had CT scan and CBCT scan acquisition in the same hour following the established clinical protocol for HN patients. The CT images were acquired on a Siemens scanner (SOMATOM go.Sim) with an exposure of 120kV and a slice thickness of 2mm. The patients were immobilized with a personalized 5-points thermoplastic mask and they received contrast agent injection prior to CT image acquisition. The CBCT acquisition was performed on one of the 3 linear accelerators (Versa HD, Elekta AB) used for the treatments using the same imaging protocol (120 kV, M20, and 2mm slice thickness).

The CT and CBCT images, were registered based on bony anatomy using a 3D-rigid transformation (translation and rotation) with mutual information (MI) as similarity metric inside of the Monaco treatment planning system (TPS). On the CT images, OARs and target volumes were defined by the treating radiation oncologist. Using VMAT, a simultaneous integrated boost technique was used to deliver 70Gy to the PTV associated to the primary tumor and 54.25Gy to the PTV associated to prophylactic nodal target, in 35 fractions of 2Gy. The dose calculations were performed in Monaco TPS. The clinical plans were saved and the templates were applied to the CBCT images after correct positioning of the plan isocenter. With the assumption that not important anatomical variation was present between the two image scans, the set of contours from the CT with the correspondent approved clinical plan were used as reference for dose calculations.

Table 5.2-1 Test cohort

Patient Number	Tumor Localisation	TNM
Patient 1	Nasal cavity	T2 N3 M0
Patient 2	Nasal cavity	T3 N0 M0
Patient 3	Oral cavity	T4 Nx Mx
Patient 4	Oral cavity	T2 N1 M0
Patient 5	Oropharynx	T4 N2 M0
Patient 6	Oropharynx	T1 N3 M0
Patient 7	Oropharynx	T2 N1 M0
Patient 8	Hypopharynx	T3 N1 M0
Patient 9	Tongue	T1 N1 M0
Patient 10	Tongue	T2 N2 M0
Patient 11	Tongue	T3 N2 Mx
Patient 12	Tongue	T2 N2 M0
Patient 13	Tongue	T3 N2 M0
Patient 14	Tongue	T1 N2 M0
Patient 15	Tongue	T4 N3 M0
Patient 16	Tongue	T4 N1 M0
Patient 17	Tongue	T2 N2 M0
Patient 18	Soft palate	T3 N2 M0
Patient 19	Tonsils	T2 N1 M0
Patient 20	Larynx	T3 N2 M0
Patient 21	Larynx	T4 N2 M1
Patient 22	Larynx	T2 N0 M0
Patient 23	Larynx	T3 N0 M0
Patient 24	Larynx	T2 N0 M0
Patient 25	Larynx	T3 N0 M0

Synthetic-CT image generation methods

Four methods were investigated for dose calculation on CBCT images (Figure 5.2-1).

1. CBCT HU-ED curve from phantom measurements

Three HU-ED curves were established by acquiring images of the CIRS 062 phantom with the HN CBCT protocol acquisition on the 3 treatment machines used for treating the patients in the cohort (Figure 5.2-2). This phantom contained the same heterogeneous inserts with known ED, as used for the CT calibration, namely: lung inhale, lung exhale, adipose tissue, breast tissue, water equivalent, liver, muscle, trabecular bone and dense bone. When compared with other phantom configurations, one study concluded that a site-specific calibration curve yielded the best dose agreement [343]. Similarly, in our study, after trying several configurations of the phantom, the inner circle of the phantom was used for establishing the curve (Figure 5.2-2). The obtained relative HU-ED curves were thereafter introduced in the TPS and applied to the CBCT images for each patient accordingly. To account for the missing image information on the CBCT image, an additional structure was created by subtracting the CBCT image contour from the patient contour, and its ED was forced to that of water.

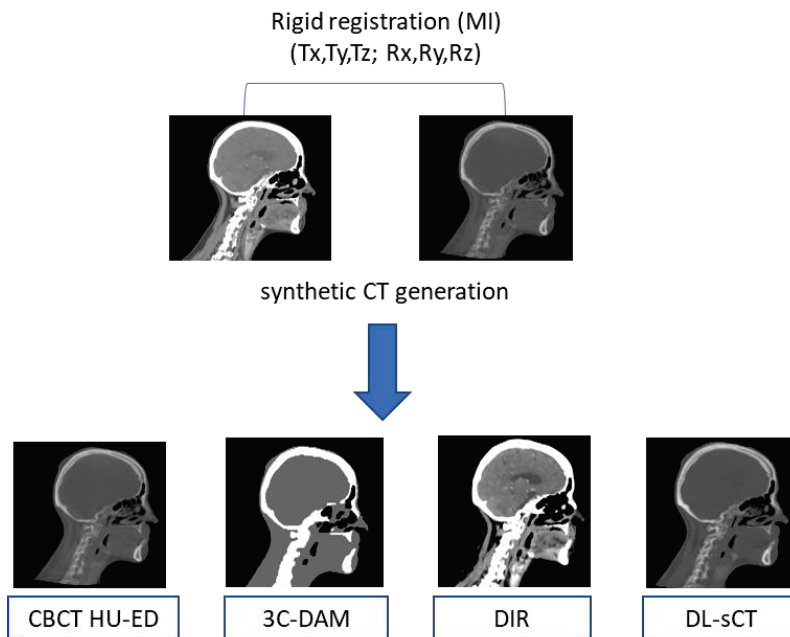


Figure 5.2-1 Methods used of creating synthetic CT from CBCT images; between the planning CT and the CBCT from the same day, a rigid registration was applied

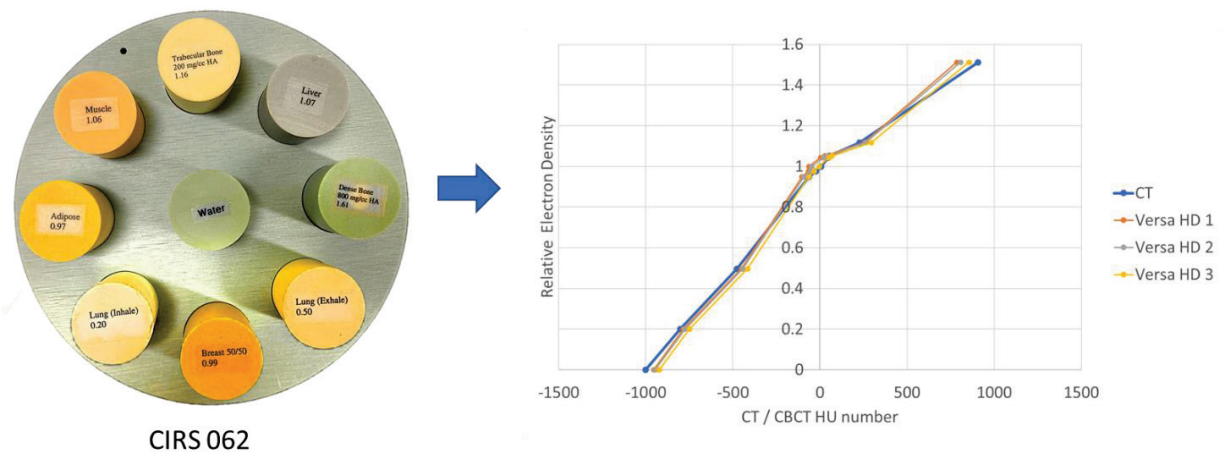


Figure 5.2-2 Phantom used for establishing the HU-ED curve for the CBCT systems on 3 Versa HD machines

2. Density assignment method

Similar to another study [344], on each patient reference CT, inside on the patient contour, 3 tissue classes were segmented based on HU thresholds. For air, soft tissue and bone, respectively, [-1024; -150], [-150; 150] and [150; 4096] ranges for HU values were used, and the meanHU values for the 3 tissue classes were documented for each patient. Similarly, CBCT images were segmented into 3 classes based on HU thresholding. Additional manual corrections of the class segmentations were necessary in case of image artifacts. The soft tissue segment was obtained by subtracting the bone and air segments from the patient

body contour on the CBCT image. Per class, meanHU values from the pCT segments were assigned to the corresponding tissue class on the CBCT images.

3. CT to CBCT deformable registration

Using the research software ADMIRE (v4.3, Elekta AB), a deformed image was created between the planning CT (moving image) and the CBCT (fixed image). Following a multi-resolution framework, first a rigid registration is performed based on local cross correlation metric and then the deformable registration is applied based on block-matching method and normalized-mean-of-squared-differences as similarity metric. The accuracy of the deformation fields obtained with ADMIRE was fully described elsewhere [310].

4. Deep learning method for sCT generation

A cycleGAN model was proposed by Elekta AB, that has been trained with unpaired CT and CBCT images. From each CBCT image, a deformed CT was obtained by applying DIR between the planning CT and the CBCT image. Two generators were trained, one to produce sCT images from the original CBCT images and another one to generate synthetic-CBCT (sCBCT) images from the original true CT images. Two discriminators were used, one to discriminate the sCT from real CT images and the other one to discriminate the sCBCT from real CBCT images. Additionally, a structural similarity index map (SSIM) weighted L1-loss term were adopted. Minimizing the L1-loss term, enforced the generated images to match the pixel values with the real images. Moreover, since in practice it is not realistic to have perfectly aligned corresponding CT for CBCT images even by use of advanced deformable registration methods, a SSIM-weighted term was introduced to eliminate potential distortions effects. The SSIM weights provide a mechanism of enforcing different levels of strengths on sCT images to match the targeted CT images in a pixel-by-pixel fashion. Adding a threshold on the SSIM weights allowed to select high similarity areas between the paired CBCT and CT images and to obtain more accurate sCT images. The final cycleGAN model was trained by Elekta on a multi-centric data base of CBCT images (>100 patients) including 150 CBCT scans (31 patients) from our department.

Image uncertainties evaluation

To assess the image quality, a voxel-wise comparison was performed between the pCT and the sCT images. The mean error (ME) and mean absolute error (MAE) between pCT and sCT images were calculated by the following formulas:

$$ME(X, Y) = \frac{1}{N} \sum_{i=1}^N HU_{pCT}(i) - HU_{sCT}(i)$$

$$MAE(X, Y) = \frac{1}{N} \sum_{i=1}^N |HU_{pCT}(i) - HU_{sCT}(i)|$$

where N represents the total number of voxels. The values inside the patient body contour of the sCT were considered for all the 4 methods.

Dosimetric accuracy evaluation

To evaluate the dosimetric accuracy, the clinical plan first saved as a template was calculated on the 4 sCT images obtained. From DVHs, dose differences were evaluated on the reference contours from the pCT. Clinically relevant dosimetric endpoints for target volumes ($V_{95\%}$, $D_{95\%}$, $D_{2\%}$) and OARs (D_{mean} , $D_{2\%}$, $D_{5\%}$) were considered upon the clinical protocol and recommendations [25]. Furthermore, a spatial dose evaluation between pCT and sCT dose distributions was performed by calculating 3D gamma analysis (3%/3mm, 2%/2mm criteria, no dose threshold).

Statistical analysis

Statistical differences in imaging and dosimetric points between the methods were assessed using Kruskal-Wallis test followed by post-hoc Dunn's test with Bonferroni correction in Python Notebook 3.8. P-values <0.05 were considered significant.

5.3. Results

Imaging points

The ME and MAE results from pixel-wise comparison between pCT and sCT images showed that DIR method provided the most similar image quality to the corresponding pCT with significant differences demonstrated ($p < 0.008$) compared with the other methods (Table 5.3-1). The largest discrepancies were observed for the adapted CBCT HU-ED curve, and the results were significantly different compared to all the other methods. At the same time, results from 3C-DAM and DL-sCT methods were similar and not statistically different. An illustration of MAE results can be visualized in Figure 5.3-1.

Table 5.3-1 Results from pixel-wise comparison between pCT and sCT images

	HU-ED curve	3C-DAM	DIR	DL-sCT
ME	139.5 ± 71.9 *	28.4 ± 27.1 *	-6.8 ± 12.1	27.9 ± 13.3 *
MAE	224.9 ± 32.0 *	145.6 ± 14.9 *	102.9 ± 25.3	137.9 ± 15.6 *

In bold are marked highlighted the best results among the methods and with * are marked the statistically significant differences when compared with the other methods.

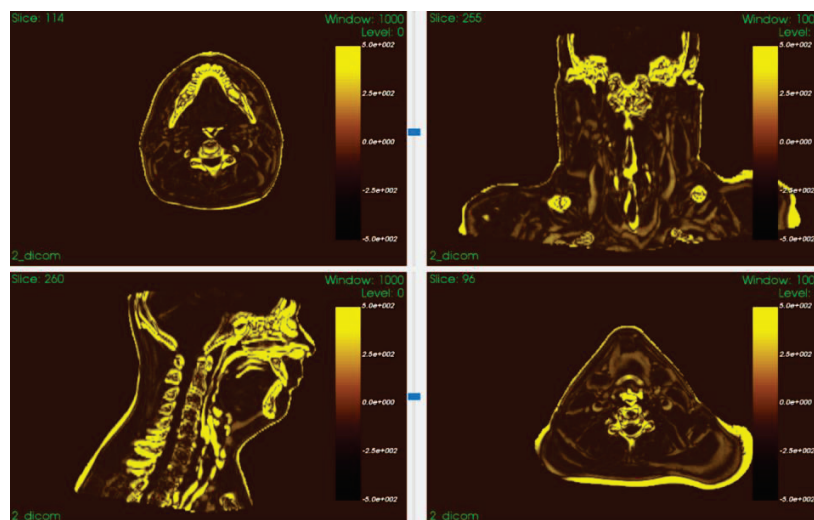


Figure 5.3-1 Illustration of mean absolute error results between CT and sCT image

Dosimetric analysis

In all the plans the dose objectives and constraints were achieved (Table 5.3-2). The mean differences were <1Gy in majority of the dosimetric endpoints and were significantly different compared with reference only in the targets DVH points (in D_{2%} to PTV_70Gy and in V_{95%}, D_{95%} and D_{2%} to PTV_54.25Gy). Larger dose deviations >1.2Gy were observed in V_{95%} of PTV_54.25Gy, that were significant for DIR and DL-sCT plans (p<0.001). Significant dose differences >1.2Gy were found also for HU-ED curve method in D_{2%} to both targets. Moreover, the largest dose deviations to OARs were seen for DIR method on the mean dose to the parotids (>2.6Gy), but the differences were not statistically significant. An illustration of DVH comparison can be visualized in Figure 5.3-2.

Table 5.3-2 Dose differences between the plans calculated on the pCT and on the sCT images

		Planning CT	HU-ED curve	3C-DAM	DIR	DL-sCT
		Reference values	Δ mean	Δ mean	Δ mean	Δ mean
PTV_70Gy	V _{95%} (%)	98.57 ± 1	0.21 ± 1.37	0.19 ± 1.19	0.39 ± 1.09	0.65 ± 1.39
PTV_70Gy	D _{95%} (Gy)	67.89 ± 0.67	-0.96 ± 1.28	-0.4 ± 0.72	0.04 ± 0.54	-0.02 ± 0.63
PTV_70Gy	D _{2%} (Gy)	71.79 ± 1.5	-1.83 ± 1.36	-2 ± 1.59	-0.96 ± 0.46	-0.68 ± 0.45
PTV_54.25Gy	V _{95%} (%)	97.86 ± 1.48	1.34 ± 2.07	1.22 ± 2.03	1.63 ± 1.86	2.08 ± 2.03
PTV_54.25Gy	D _{95%} (Gy)	52.93 ± 0.61	0.31 ± 1.2	0.5 ± 0.99	0.83 ± 0.86	0.99 ± 1.01
PTV_54.25Gy	D _{2%} (Gy)	54.09 ± 2.17	-1.28 ± 0.93	-0.21 ± 0.25	-0.67 ± 0.32	-0.48 ± 0.33
Parotid_R	D _{mean} (Gy)	21.3 ± 7.27	0.07 ± 0.75	0.06 ± 0.83	-2.69 ± 13.28	0.1 ± 0.75
Parotid_L	D _{mean} (Gy)	23.86 ± 11.03	-0.38 ± 0.86	-0.34 ± 1.38	-3.46 ± 15.24	-0.13 ± 0.95
Oral Cavity	D _{mean} (Gy)	40.69 ± 12.32	-0.78 ± 0.59	-0.67 ± 0.48	-0.22 ± 0.29	-0.44 ± 0.43
SpinalCord	D _{2%} (Gy)	31.78 ± 5.37	-0.5 ± 0.52	-0.21 ± 0.35	-0.07 ± 0.24	-0.06 ± 0.32
Brainstem	D _{2%} (Gy)	14.2 ± 9.25	-0.04 ± 0.67	0.16 ± 0.68	-0.03 ± 0.45	0.15 ± 0.58
Mandible	D _{5%} (Gy)	57.66 ± 8.91	-1.4 ± 0.83	-0.9 ± 0.59	-0.16 ± 0.63	-0.29 ± 0.61

Δ mean= mean dose differences between reference values on pCT and plans calculated on the sCT images; in bold are highlighted the significant dose differences compared to reference

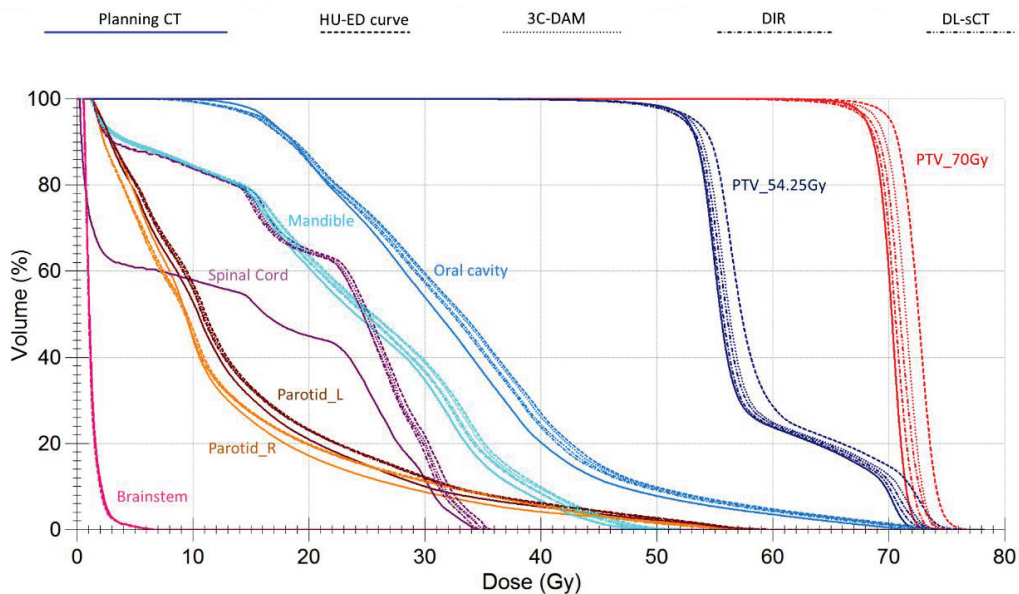


Figure 5.3-2 Illustration of a dose volume histogram comparison between plans calculated on pCT and sCT images

The results from 3D gamma analysis are summarized in Table 5.3-3, for 3%/3mm and 2%/2mm criteria. Overall DIR method provided the best dose agreement for majority of the structures considered. Based on the variance test, results from 3%/3mm gamma criteria were significantly different for all the structures considered ($p < 0.02$), except the parotids, spinal cord and brainstem. Similarly, significantly different results were found for the 2%/2mm criteria for the same structures and additionally for the spinal cord ($p < 0.02$). More precisely, with regard to gamma 3%/3mm, DIR and DLs-CT results were not significantly different ($p > 0.06$). However, DIR method results were significantly better compared to HU-ED method for all the structures except the parotids, spinal cord and brainstem ($p < 0.01$). Compared with 3C-DAM, DIR results were significantly better for PTV_70Gy, oral cavity and mandible ($p < 0.005$). Furthermore, with regard to 2%/2mm criteria, DIR results were significantly better compared with DL-sCT method for PTV_70Gy and for oral cavity ($p < 0.02$). When compared to HU-ED method, DIR results were significantly better for all listed structures except the parotids and the brainstem ($p < 0.02$) and, when compared with 3C-DAM, DIR results were significantly better for PTV_70Gy, oral cavity and the mandible ($p < 0.001$). An illustration of gamma analysis results can be visualized in Figure 5.3-3.

Table 5.3-3 Gamma pass rate results between the plans calculated on the pCT and on the sCT images

	3D Global Gamma 3%/3mm			
	HU-ED	3C-DAM	DIR	DL-sCT
PTV_70Gy	80.29 ± 24.88 *	97.36 ± 5.76 *	99.68 ± 0.69	98.9 ± 3.11
PTV_54.25Gy	80.79 ± 17.47 *	96.25 ± 5.01	97.63 ± 2.88	96.59 ± 3.27
Parotid_R	99.43 ± 1.96	99.65 ± 1.37	99.52 ± 1.84	99.69 ± 1.41
Parotid_L	99.84 ± 0.36	99.52 ± 2.02	99.98 ± 0.06	99.93 ± 0.17
Oral Cavity	97.71 ± 2.76 *	98.75 ± 1.84	99.86 ± 0.34	99.3 ± 1.08
SpinalCord	98.68 ± 3.2	99.63 ± 1.06	99.73 ± 0.77	99.63 ± 1.1
Brainstem	99.95 ± 0.25	100 ± 0.01	100 ± 0	100 ± 0
Mandible	97.22 ± 5.34 *	98.83 ± 2.01 *	99.84 ± 0.34	99.44 ± 1.46
Patient contour	95 ± 4.47 *	97.2 ± 2.94	97.84 ± 2.67	97.25 ± 2.81

	3D Global Gamma 2%/2mm			
	HU-ED	3C-DAM	DIR	DL-sCT
PTV_70Gy	53.36 ± 31.16 *	81.88 ± 16.84 *	96.11 ± 5.45	89.93 ± 11.6 *
PTV_54.25Gy	61.33 ± 21.71 *	84.86 ± 13.06	92.63 ± 7.31	87.85 ± 9.49
Parotid_R	94.71 ± 9.36	94.55 ± 9.08	95.37 ± 10.49	95.36 ± 8.12
Parotid_L	95.99 ± 6.16	95.61 ± 12.66	98.13 ± 3.7	97.18 ± 4.85
Oral Cavity	82.62 ± 14.38 *	88.25 ± 10.15 *	97.4 ± 4.33	91.98 ± 7.34
SpinalCord	92.77 ± 9.92 *	97.76 ± 3.02	98.34 ± 2.53	98.13 ± 2.75
Brainstem	98.56 ± 4.7	99.34 ± 1.56	99.64 ± 1.49	99.66 ± 1.01
Mandible	85.01 ± 14.26 *	89.24 ± 8.6 *	96.88 ± 5.24	93.45 ± 7.04
Patient contour	86.04 ± 8.89 *	91.33 ± 6.31	93.47 ± 6.39	92.05 ± 5.56

In bold are highlighted the best gamma results and with * are marked the significant differences compared with the other methods

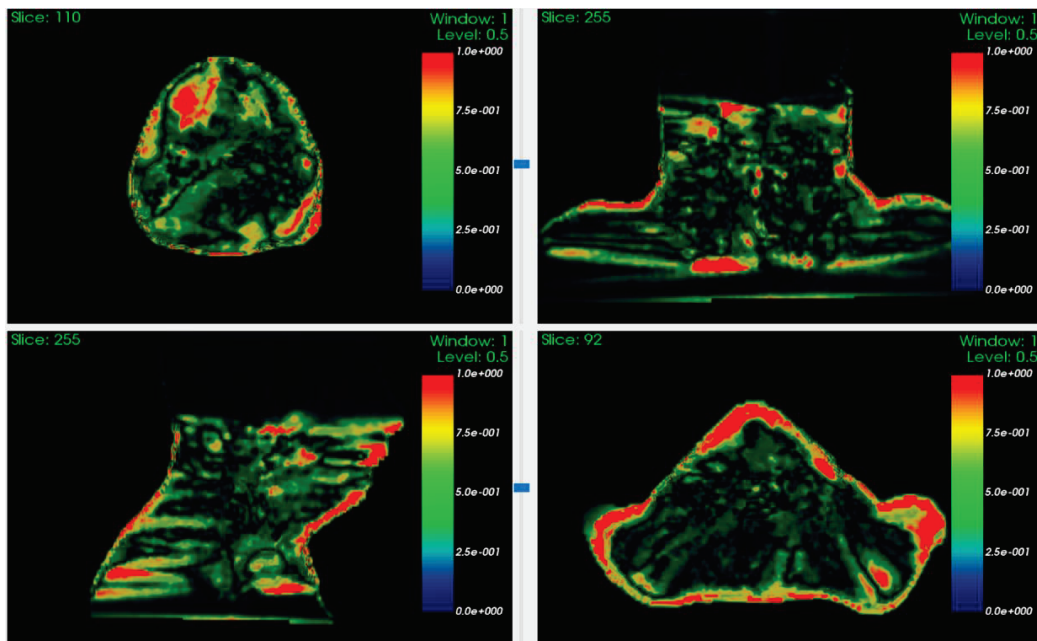


Figure 5.3-3 Illustration of one patient's gamma analysis results (3%/3mm) between reference and the plan calculated on sCT image

5.4. Discussion

This study provided results for dose calculations methods on CBCT images. Four methods were evaluated and compared. DIR and DL methods were the fastest, providing an sCT images in <1min. For 3C-DAM method, the sCT generation process involved several steps that were more time-consuming (approximately 5 min per patient to verify and correct the segmented tissue classes). Establishing the HU-ED curve was simple but the most laborious because the phantom irradiations were performed on the 3 treatment machines. The largest differences were in the dense bone insert: when compared to the pCT measured values, mean HU difference was 90 HU, and when compared among the three CBCT acquisitions, the mean difference was 46 HU. However once established, the curves were used accordingly for all the patients.

Overall the results showed that DIR method was the most accurate in both imaging and dosimetry points. However, the results were not always significantly different compared with DL method. Moreover, among all the methods, dose deviations on the parotids were the largest for DIR algorithm. Nevertheless, for majority of the structures, the gamma agreement was the highest for DIR solution for both gamma criteria. CBCT images were acquired shortly after performing the pCT, thus allowing to assume that the anatomical deformations were minimal. However, this configuration might have represented an advantage for the DIR algorithm over the other methods, for the calculation of the deformable vector fields. DL-sCT method yielded the best results after DIR, with significant dose deviations only for the elective PTV. Gamma rates were >96% and >87% for 3%/3mm and 2%/2mm criteria respectively, with the largest discrepancies observed on the PTVs. 3C-DAM method was similar in image accuracy when compared to DL-sCT method. The dose deviations were small <1.2Gy and the gamma rates were >96% and >81% for 3%/3mm and 2%/2mm criteria, respectively. Adapted HU-ED method had the worse image quality, dose deviations <1.4Gy and lowest gamma pass rates particularly on the target contours (<80%).

Another study from literature performed a similar comparison (using ADMIRE software with same DIR and DL method) based on 14 HN patients [310]. They reported ME/MAE results of 208.9/266.6 HU, 14.6/113.2 HU, -36.6/95.5 HU and 17.1/82.4 HU for a HU-ED, DAM, DIR and DL method, respectively. In

our study, the results were similar, with larger MAE results particularly for DAM (145.6 HU) and DL-sCT method (137.9 HU). Similar results were obtained for the dose deviations in targets and OARs (<1.2Gy). Gamma pass rates were also comparable between our study and Barateau et al, showing best agreement for DIR (98.8%) method, followed by DL method (98.1%). However, contrarily to our study results, in their comparison, slightly better image quality was obtained with the DL method (trained on 30 HN patients from the same center). Similarly, another study reported slightly lower ME/MAE (1.4/77.2 HU) for a similar DL solution (trained on a mono-centric database of 25 patients and a total of 120 unpaired CBCT images, using ADMIRE software) [317]. They demonstrated mean relative dose differences <1% to the gross tumor volumes and <5% for the OARs. Moreover, their reported gamma rates (on the patient volume) were 98.6% and 95% for 3%/3mm and 2%/2mm criteria, respectively. In our study, using the DL method, gamma rates of 97% and 92% were obtained for 3%/3mm and 2%/2mm criteria, respectively. Similarly, other related study using a cycleGAN-based method (trained on 90 HN CT and CBCT images), reported gamma values of 98.4% on HN localization and a MAE of 29.85 HU [82].

A known issue of GANs architectures is the model instability when networks are trained with only adversarial losses. The additional cycle consistency loss from cycleGAN aim to address this issue. Moreover, in the framework of the cycleGAN model proposed by Elekta in this study, a SSIM threshold with weighted L1-norm term was added to obtain a more robust model. However, although cycleGAN can remove most scatter artifacts on the CBCT images and correctly assess HU values to match those of the CT images, when both the CT and CBCT training datasets include metal artifacts, they will still be present in the sCT images. Moreover, CBCT images that have severe truncation problems will also yield sCT images that have truncations.

Another concern for GANs, is the application range. This is clearly present in our study where the model trained on a large multi-centric database, seemed to not provide a good generalizability to our test cohort. Better performances were demonstrated by similar network architectures that were trained and tested on uniform image acquisition cohorts [82,310].

As future perspectives of the study, we plan to evaluate the model performance when trained with uniform data and also to investigate solutions for the image truncation problem caused by the CBCT small FOV. Another direction will be to investigate the accuracy of the DIR method when using CBCT images from different treatment fractions. Furthermore, DIR accuracy can be evaluated with regard to dose accumulation, as proposed in another study [100]. Provided that the solutions are integrated in the TPS, the success of such work would enable accurate and efficient generation of sCT from CBCT images that can be further used for ART planning.

5.5. Conclusion

Finally, four methods for dose calculation on CBCT images were evaluated and compared in this study. DIR method demonstrated the best image quality and the best gamma pass rates. Next was the DL method. DAM provided comparable results with the DL solution but was more time-consuming. Using a HU-ED curve demonstrated the largest discrepancies in both image similarity metrics and dose calculation accuracy. Future investigations are needed to determine the accuracy of DIR in different fractions of the treatment, and the performance of DL based on a uniform training cohort. Ultimately, used with caution, DIR and DL methods can potentially be used for CBCT-based adaptive workflows for RT of HN patients.

5.6. Synthesis

The present study provided an evaluation and comparison of 4 CBCT-based dose calculation methods. Their performance was evaluated with regards to imaging dosimetric accuracy compared to the planning CT images. The workload and resources needed for each method were also considered.

The results demonstrated that DIR algorithm had the best performance, followed by the DL method. Moreover, both were fast, generating a sCT image in <1min. Class density assignment method had similar results with DL method, but was more time-consuming (~5min) due to the extra steps related to tissue class segmentation and HU numbers assignment. Finally, the adapted CBCT HU-ED curve had the worst results.

Several metrics to quantify the accuracy of sCT images exist [207], but no real consensus on the acceptance criteria. In our study we quantified ME and MAE for all the methods inside of the body contour of the CBCT image. Some other studies are documenting the results based on tissue segments (e.g bony structures, air cavities, soft tissue). This would be interesting to investigate further in our cohort of patients. Due to voxel-wise comparison, particular care should be taken to compensate for possible anatomical difference and/or potential errors in the initial registration step between the CT and the CBCT images. A solution from literature would be to apply negative margins (e.g. 2mm) to the segments where the HU number difference wants to be assessed [317].

With regards to the dose calculation accuracy, typically DVH-based dose differences and gamma analysis are performed to compare plans calculated on the sCT and the reference CT. In our study, target objectives and OARs constraints we considered upon clinical protocol and two gamma criteria were used (3%/3mm and 2%/2mm). We observed that generally dose deviations were small and not significant. Larger discrepancies were identified with the stricter criteria for gamma analysis, whereas only DIR method passed >90% the 2%/2mm gamma criteria inside all the structures considered. Perhaps a better understanding of the exact location of the inconsistent HU numbers would allow to find correlations with the dose deviations and gamma rates. Ultimately this would enable to define appropriate limits for acceptance criteria.

The study is original because a cycleGAN was evaluated on HN localization that was trained with unpaired CBCT images from a multi-centric database. The model trained on mono-centric data seem to provide better results in other studies [82,310]. Another issue that must be further investigated is the image truncation problem caused by the CBCT small FOV, which in our study was resolved by water density override where image information was missing. Furthermore, DIR accuracy must be evaluated with regard to different fractions of the treatment, and finally the accuracy DIR for dose accumulation must be assessed [100].

Provided that are integrated in the TPS, DIR and DL methods are the most promising solutions for enabling ART planning by generation of sCT from daily CBCT images.

Chapter 6. Conclusions and perspectives of the study

Finally, the work carried out during the 3-year thesis study, aimed to evaluate different methods to accelerate the RT workflow and ultimately enable to perform ART for HN cancer patients. The contributions were divided into four axes, and notably we have evaluated: 1) The performances of an automatic planning solution based on a priori MCO algorithm 2) Six solutions for automatic segmentation of OARs and CTVn on HN CT images; 3) Four methods that enable dose calculations on daily CBCT images. A seamlessly integration of these solutions would represent the success story of ART for patients with HN cancers.

Regarding speed and accuracy, DL solutions were very attractive. However, a great amount of work is needed for the training data base collection and curation. Other methods based on less input information (wish-list based for auto planning and atlas-based segmentation for automatic contouring) enabled to reduce the clinical load. More precisely, we demonstrated that manual optimization can be safely replaced by a wish-list based auto-planning, and manual contouring time can be shortened by performing corrections on AS contours. Additionally, both methods allow an improved consistency among operators. Their limitation remains the computational time. Lastly, we observed that the fastest and most accurate method to generate sCT from daily CBCT images was the use of DIR and DL methods. A limitation remains the limited FOV in the CBCT images, which in our study was resolved by water density override where the image information was missing.

Numerous future directions can be envisioned from this work. Short term perspectives with regards to auto-planning are testing and evaluating of faster solutions promised by Elekta company, and investigate them also on a larger cohort of HN cancer patients, where population can be divided in smaller groups upon the tumor's localization. Furthermore, several wish-lists can be established in function of the clinical protocol and cancer sites, and encourage their integration into the clinical routine. Perspectives of developing a DL-based solution for auto-planning can also be considered, in order to generate plans in only few seconds. Moreover, the methodology used for quantifying the plan quality can further be applied for evaluating other auto-planning solutions before clinical implementation in our department. Yet, an obstacle for the smooth integration of the wish-list based auto-planning algorithm in the clinical routine remain the nomenclature of the structures of interest, where AS solution may be the key facilitator.

With regards to the AS solutions, we have evaluated several commercially and non-commercially available solutions. Among them, some were investigated for the first time on HN OARs and/or CTVn volumes. From this work, short term perspectives can be proposed starting with a study to evaluate the relationship between the amount of training data and the DL model performance for each of the structures. In parallel, the open-source self-configuring nnUnet can be trained with the same input, and results can be compared in order to investigate the inter-play between the training data and the network architecture. This has already been initiated in the work of a master student. Moreover, it can be investigated the model performance when operated patients are included in training. We had already tested this option by including few cases of patients with larynx or trachea removed. The initial results however were not satisfying and we concluded that more operated patients needed to be included to potentially increase model's performance. Furthermore, a future perspective with regards to the patient database, is to include data also from other centers, that must nonetheless be consistent with the image acquisition protocol (2mm thick slice, contrast agent injection) and follow international delineation guidelines. By doing this, a multi-centric study can be conducted where the generalizability of the mono-centric DL model can be tested with respect to reference manual delineations from other physicians. Furthermore, by including manual delineation from multiple observers would allow to assess IOV among manual delineations and compare it to IOV from AS+manual corrections. Similar to another study [170], it would be interesting also to further split the observations into categories based on the experience of

physicians, namely residents, young and senior physicians. While conducting the proposed study, record of the manual delineation time should be performed so that the actual time reduction could be assessed. With regards to the dosimetric impact evaluation, we demonstrated in our study that OAR's distance to PTV was not consistently interrelated with the high dosimetric deviations. Perhaps a deeper investigation can be performed in order to determine in what situations editing the AS contours makes a clinical difference, and whether thresholds for DICE/HD_{95%} could be set. Furthermore, one study showed that surface DICE and added path length (APL) measures for contour accuracy were better indicators for the clinical delineation time saved when using AS contours [345]. It would be interesting to compare these measures also on the cohorts from our studies. Another future research perspective inspired from literature [334,346] would be to develop a DL network framework with contour constraints based on anatomical landmarks. Structures with well-defined borders on CT images, that constitute anatomical borders defined in the delineation guidelines (e.g. the hypoid bone, sternocleidomastoid muscle, vertebrae) can be segmented as auxiliary task output of a DL network. Then, distance maps can be applied for restricting the shape of the desired main segmentation output (predictions of OARs and CTVn volumes).

With regard to the last thesis contribution, on CBCT-based dose calculations, we can first envision as the next step, the optimization of the DL solution for sCT generation. A new model could be trained on uniform data from our department (mono-centric model), while trying also to optimize the network architecture (number of layers, image size, loss function, weights and thresholds etc.). Its performance should be then compared with that of the multi-centric model. Secondly, another function available in Monaco TPS (AdaptAnatomy) for sCT generation based on density overrides is planned to be investigated in the near future. For this, several strategies can be followed with respect to the regions of interest (ROIs): one can use the full set of structures present on the CT set or choose only few tissue classes. To make the process automated, one can use an atlas-based AS solution for segmenting a predefined list of bony and soft tissue structures on the CT. Since this is another TPS integrated sCT generation solution, it would be relevant to compare it with the formerly investigated methods. Another direction of research from this topic is the problematic of dose accumulation. DIR accuracy remain to be assessed so that dose calculation uncertainties can be estimated [100]. Furthermore, uncertainties in CBCT-based sCT generation could be propagated from errors in daily patient re-positioning. Therefore, dose deviations should also be investigated with regard to the patient re-positioning strategies. A study has already been initiated for evaluating the dosimetric uncertainties related to bone-matching (BM) registration alone, and BM+soft tissue guided registration between the daily CBCT and the CT images. Lastly, QA criteria for the clinical implementation of a sCT method should be investigated and proposed.

Long-term perspectives would be to evaluate a RT workflow that combines automated solutions for auto-contouring, auto-planning and sCT image generation. While complex IMRT techniques enabled to reach outstanding target coverage, the integration of modern RT solutions enables stronger focus towards minimizing toxicity to normal tissues. This requires however QA checks at each step in the treatment workflow. It would be clinically useful to evaluate margin strategies considering all the uncertainties related to the automated solutions used. Another goal is to enable selection of patients that will benefit from plan adaptation. This could be done based on intra-fraction anatomical variations, where we could compare volume overlap from OARs and CTVn used for planning, with volumes automatically segmented on the sCT generated from pre-treatment CBCT images. Critical thresholds can be investigated based on evaluating planned doses against doses created on the actual anatomy of the day.

Ultimately, the clinical validation of such RT workflow would decrease the manual workload, help in the harmonization of the clinical practices and enable fast decision making for RT of complex HN cancer cases. Results from clinical trials are awaited to determine the clinical benefits of ART.

Résumé étendu

Introduction

Le cancer est l'une des principales causes de décès dans le monde et peut se développer dans plusieurs régions de la tête et du cou (HN), notamment le larynx, l'oropharynx, le nasopharynx, l'hypopharynx, la thyroïde, les glandes salivaires, la cavité orale et les lèvres. D'après les statistiques de 2020, sur les 1 518 133 patients chez qui un cancer HN a été diagnostiqué, 34 % n'ont pas survécu [1]. La radiothérapie externe (RT) est l'un des traitements les plus efficaces pour ce type de tumeurs cancéreuses. Son principe est de délivrer des rayonnements ionisants de haute énergie (MV) à l'aide d'un accélérateur linéaire (LINAC), afin d'induire des dommages à l'ADN des cellules cancéreuses et de bloquer leur capacité à se multiplier. Les tissus normaux ont une plus grande capacité de réparation après une irradiation et la toxicité peut être limitée en divisant la dose sur plusieurs jours. Les techniques de radiothérapie à modulation d'intensité (IMRT) avec des distributions de dose hautement conformes et des gradients de dose abrupts constituent la norme pour la RT des tumeurs HN en assurant une couverture maximale de la zone cible et l'épargne des organes à risque (OAR). En général, une dose de radiothérapie cumulée de 70Gy est délivrée avec une intention curative sur plusieurs semaines en fractions quotidiennes de 1.8 – 2.0Gy. L'acquisition de l'image tomodensitométrique de planification (pCT) est un prérequis pour définir le positionnement de référence du patient, avoir accès à l'anatomie du patient et aux densités électroniques (ED) des tissus. Sur l'image pCT, le radiothérapeute définit le volume cible de planification (PTV) et les OAR, afin que les doses de rayonnement puissent être établies sur la base de contraintes dose-volume. Le contourage et la planification demandent une grande précision. Néanmoins, elles sont laborieuses et susceptibles de variations intra et inter-observateurs (IOV). De plus, des changements anatomiques (par exemple, perte de poids, réduction de la tumeur, déplacement des OAR) peuvent se produire entre le pCT et la première session de traitement, ainsi qu'entre les fractions de traitement, ce qui peut causer des différences entre les doses planifiées et les doses délivrées effectivement.

Des stratégies de RT adaptative (ART) ont été développées pour corriger les variations anatomiques intra-fractionnelles. Idéalement, les dispositifs d'imagerie tridimensionnelle (3D) installés dans la salle de traitement, principalement utilisés pour minimiser les erreurs de positionnement du patient, peuvent être utilisés pour évaluer les déformations anatomiques du patient et pour effectuer un nouveau calcul de la dose sur l'anatomie du jour. Les systèmes de tomographie à faisceau conique (CBCT) à faible énergie intégrés à la machine LINAC sont très répandus et utilisés pour vérifier la mise en place du patient. Par contre, la qualité de images n'est pas adaptée aux calculs de dose à cause de plusieurs inconvénients (artefacts d'image, incohérence des chiffres de l'unité Hounsfield (HU), et champ de vue limité (FOV)). Avec l'avènement des solutions d'intelligence artificielle (IA), plusieurs applications ont été proposées pour faciliter la mise en œuvre de l'ART. Parmi elles, la segmentation automatique des images (AS), la planification automatique du traitement (auto-planning) et la génération d'images CT synthétiques (sCT) à partir d'images CBCT sont abordées dans ce manuscrit de thèse.

L'objectif de cette thèse était d'étudier plusieurs solutions automatisées pour différentes étapes du flux de travail de la RT des patients atteints de cancer HN. En premier lieu, une solution de planification automatique a été évaluée. Ensuite, la performance de plusieurs solutions AS a été comparée pour les OARs HN et les niveaux de ganglions lymphatiques (CTVn) dans la région HN. Enfin, plusieurs méthodes qui permettent de calculer la dose sur des images CBCT du jour ont été étudiées. Le manuscrit représente le travail effectué au cours des trois dernières années et est organisé en six chapitres.

Le premier chapitre présente le contexte de l'étude. Dans la première partie est décrit le processus standard d'un traitement par RT, les modalités utilisées dans la RT guidée par l'image (IGRT), et le concept

d'ART avec un accent sur le traitement du cancer HN. En outre, l'émergence de solutions d'IA dans la RT est présentée en mettant l'accent sur les méthodes plus récentes en matière de AS, les solutions de planification automatique et les méthodes de génération d'images sCT. Dans la dernière partie du chapitre sont résumés les objectifs de la thèse.

Le deuxième chapitre décrit l'évaluation des performances d'une solution d'auto-planification par rapport à des plans de traitement optimisés manuellement. Ce travail représente une première contribution, en tant que second auteur, à l'article qui a été publié dans le *European Journal of Medical Physics* en 2021 [2].

Le troisième chapitre présente le travail de la deuxième contribution, l'évaluation des performances de six méthodes d'AS pour la segmentation des OARs sur des images CT. Ce travail a été publié dans le *Radiotherapy and Oncology Journal* [3]. L'évaluation des solutions AS était basée sur la demande en ressources, la précision géométrique, le temps nécessaire aux corrections manuelles et l'impact dosimétrique sur les distributions de dose RT calculées à l'aide de l'auto-planning.

De la même façon, dans le quatrième chapitre, les performances des six mêmes méthodes AS ont été évaluées sur les niveaux de ganglions lymphatiques (CTVn), qui sont généralement irradiés comme volumes cibles électifs dans la région HN. Les résultats de cette étude sont présentés sous la forme d'un article qui sera prochainement soumis également au *Radiotherapy and Oncology Journal*.

Le cinquième chapitre présente les résultats de l'évaluation de quatre méthodes de génération de sCT à partir d'images CBCT du jour. Une solution basée sur DL a été étudiée et comparée à d'autres méthodes proposées dans la littérature. Par rapport au pCT de référence, la précision de l'image et la précision du calcul de la dose ont été mesurées.

Enfin, le chapitre six résume les conclusions de la thèse et les perspectives futures de ce travail.

Les travaux réalisés au cours de ces trois années de thèse ont été financés par Elekta LTD, et ont été conduits au service de RT du Centre de Cancérologie Léon Bérard, dans l'équipe TOMORADIO des laboratoires CREATIS.

Chapitre 1 Contexte clinique

Le cancer est une maladie caractérisée par une croissance indésirable et incontrôlée de cellules qui se sont développées à partir de cellules normales du corps et qui présentent des mutations structurelles et fonctionnelles. Dans le monde, plus de 18 millions de patients se voient diagnostiquer un cancer chaque année. En 2020, les tumeurs de la tête et du cou (HN) ont contribué 8.2 % et 5 % respectivement à l'incidence et à la mortalité mondiales par cancer. D'un point de vue thérapeutique, le HN est une région difficile à traiter car de nombreux organes de cette région sont associés à des fonctions physiologiques telles que la respiration, la communication et la nutrition. Par conséquent, la prise en charge des patients atteints de cancer HN impose une approche thérapeutique multidisciplinaire impliquant la chirurgie, la radiothérapie externe (RT) ou interne et les traitements systémiques. Afin d'atteindre l'objectif thérapeutique, la prescription comprend souvent une combinaison de ces options thérapeutiques.

Le principe de la RT est d'induire des dommages à l'intérieur des cellules cancéreuses tout en limitant les effets sur les tissus normaux. En raison d'un cycle cellulaire atypique, avec une phase de division accélérée, les cellules cancéreuses sont plus sensibles aux rayonnements et peuvent donc être affectées de façon létale par les rayonnements, alors que l'effet sur les cellules normales n'est pas létal. L'accélérateur linéaire (LINAC) est l'équipement le plus répandu pour l'administration de RT. Il est capable de produire des rayons X de haute intensité, en accélérant des électrons vers une cible en tungstène. La plus grande partie de l'énergie cinétique de l'électron est transformée en chaleur et une petite fraction est émise sous forme de photons de rayons X. Une fois qu'un faisceau de rayonnement est produit dans un LINAC, il peut être modélisé à l'aide de plusieurs éléments à l'intérieur de la tête de traitement, notamment des filtres, des blocs et des collimateurs. L'introduction du collimateur multi-lames (MLC) dans

la conception du LINAC a permis de modéliser plus précisément le champ de rayonnement en fonction du contour des cibles tumorales. Cela a permis l'introduction de la RT modulée en intensité (IMRT) et la thérapie volumétrique par modulation d'arc (VMAT), qui sont devenues la modalité standard pour les traitements par RT des cas complexes tels que les cancers HN.

Une prescription de dose typique pour les patients HN au Centre de Cancérologie Léon Bérard consiste à délivrer 70Gy au volume cible planifié (PTV) associé à la tumeur primaire, et 54.25Gy au PTV associé à la cible ganglionnaire prophylactique, en 35 fractions de 2Gy. La radiothérapie guidée par l'image (IGRT) est le processus d'imagerie régulière, au cours d'une séance de RT, utilisé pour guider la position du patient, en comparant les images de simulation CT aux images de pré-traitement, acquises dans la salle de traitement avant la délivrance de la dose. L'objectif principal de l'IGRT est de réduire les erreurs d'installation et de positionnement du patient en corrigeant l'alignement de différentes images du même patient. Avec l'aide de l'IGRT, la radiothérapie adaptative (ART) cherche à prendre des mesures correctives, si nécessaire, en fonction des changements quotidiens de la tumeur et des tissus normaux. La principale limite de l'ART est le temps nécessaire pour adapter le plan à l'anatomie du jour. Pour faciliter la mise en œuvre de l'ART, des solutions d'intelligence artificielle ont émergé en RT pour plusieurs applications.

Dans ce contexte clinique, cette thèse propose d'évaluer plusieurs solutions automatisées pour différentes étapes du flux de traitement par RT des patients atteints de cancer HN, qui peuvent permettre la mise en œuvre de la ART pour cette localisation. Les contributions de ce travail ont été divisées en quatre axes :

1. Planification automatique du traitement
 - Nous avons évalué la qualité des plans de traitement HN en utilisant une solution de planification automatique par rapport aux plans de traitement manuels VMAT et TomoTherapy.
2. Segmentation automatique pour les OARs
 - Nous avons comparé 4 solutions de segmentation automatique basées sur une bibliothèque d'atlas et 2 solutions de segmentation automatique d'apprentissage profond (DL) pour la délimitation de 10 OARs typiquement délimités sur les images pCT des patients de cancers HN.
 - Nous avons évalué leurs performances en ce qui concerne la demande de ressources, la précision géométrique, le temps nécessaire aux corrections manuelles et l'impact dosimétrique sur les distributions de dose de RT calculées à l'aide de la planification automatique.
3. Segmentation automatique pour les niveaux ganglionnaires lymphatiques
 - Nous avons comparé les 6 mêmes solutions de segmentation automatique pour la délimitation de trois niveaux de ganglions lymphatiques (CTVn) sur des images CT qui sont habituellement irradiés comme volumes cibles secondaires.
 - Nous avons évalué leurs performances en termes de demande de ressources, de précision géométrique, de temps nécessaire aux corrections manuelles et d'impact dosimétrique sur les distributions de dose de RT calculées à l'aide de la planification automatique.
4. Calculs de dose basés sur le CBCT pour la ART
 - Nous avons comparé différentes méthodes pour générer des images CT synthétique (sCT) à partir d'images CBCT.
 - Nous avons évalué leur application potentielle pour la TAR en termes de précision du calcul de dose et de qualité d'image.

Chapitre 2. Validation d'une solution de planification de traitement automatisée

La première investigation des méthodes adaptatives pour le traitement du cancer HN, a été l'évaluation de la performance d'un algorithme d'optimisation de plan à base de multicritères a priori. L'objectif de l'étude était d'examiner la version de recherche de la solution de planification automatique mCycle (Elekta AB) par rapport à la planification manuelle conventionnelle utilisant la VMAT ou la tomothérapie hélicoïdale (HT) pour une cohorte de patients HN. Les résultats sont présentés sous la forme d'un article (Biston et al. [2]) qui a été publié dans le European Journal of Medical Physics en 2021. Dans ce travail, j'ai contribué à l'analyse des résultats.

Dans cette étude, les performances de la solution d'auto-planification mCycle ont été évaluées par rapport aux plans VMAT ou HT optimisés manuellement. La comparaison a été effectuée sur la base des calculs des indices de qualité des plans (PQI), de l'évaluation en aveugle par 2 médecins qualifiés, du nombre de points de contrôle (CP), du nombre d'unité de contrôle (MU), de scores de complexité de modulation (MCS) et des mesures d'assurance qualité. Sur la base d'une cohorte de 14 carcinomes du nasopharynx (HN supérieur) et de 14 "indications moyennes inférieures" (HN inférieur), la supériorité de la solution mCycle a été démontrée. De plus, les plans mCycle ont été considérés comme meilleurs que les plans manuels dans 75 % des cas. Ce résultat est cliniquement significatif car l'optimisation manuelle d'un cas complexe de HN nécessite au moins 3-4 optimisations et prend beaucoup plus de temps (>2h) par rapport à la solution automatique proposée (<1h). Par ailleurs, une solution sans utilisateur permet d'accroître la cohérence entre les planificateurs. Un autre grand avantage de l'utilisation de la solution mCycle est qu'elle permet de mieux épargner les OARs tout en maintenant la couverture souhaitée pour les PTV. C'est la principale raison pour laquelle elle a été préférée aux plans manuels. Par conséquent, les plans mCycle étaient plus complexes que les plans VMAT manuels, ce qui a considérablement augmenté les temps de traitement, sans toutefois avoir d'impact négatif sur les mesures d'assurance qualité.

À mon avis, cette preuve motive l'intégration clinique de solutions de planification automatique pour les cas complexes tels que HN, où une meilleure épargne des OAR peut être obtenue par des itérations automatiques que par un opérateur humain. Néanmoins, si d'autres améliorations du plan de traitement sont souhaitées, le plan proposé par mCycle peut être un bon point de départ pour des ajustements supplémentaires. Il convient de mentionner qu'un temps excessif est nécessaire pour obtenir une 'wish-liste' robuste et que le temps de calcul relativement long reste une limite pour l'ART. Pour obtenir la 'wish-liste', il faut mettre en place un processus de réglage itératif intensif qui implique un travail de collaboration entre l'équipe de cliniciens pour décider de l'ordre de priorité des différentes fonctions et des objectifs de dose. La complexité de la tâche augmente avec le nombre d'OAR critiques considérés, et plusieurs d'itérations sont nécessaires pour bien équilibrer les objectifs et les contraintes de dose. Néanmoins, une fois validé pour un protocole clinique et pour une localisation, il peut être rapidement adapté à un autre protocole de dose ayant des contraintes de dose similaires (par exemple une localisation HN avec 3 niveaux de dose PTV).

En ce qui concerne les perspectives d'avenir, la société Elekta s'efforce de fournir une solution de planification automatique plus rapide (<15min pour la planification HN) qui ouvrira les portes de l'ART pour les patients atteints de cancer HN. Cependant, la délimitation manuelle des organes sur l'anatomie du jour reste une limitation qui remet en cause les résultats prometteurs des solutions de AS. Ce sujet sera abordé plus en détail dans les deux prochains chapitres de la thèse.

Chapitre 3. Comparaison des méthodes basées sur l'atlas et l'apprentissage profond pour la délimitation des organes à risque sur les images de tomodensitométrie de la tête et du cou à l'aide d'un système automatisé de planification des traitements

Ce chapitre représente le travail d'un article qui a été publié dans le *Radiotherapy and Oncology Journal* en novembre 2022 [3]. L'objectif était d'évaluer et de comparer les performances de différentes méthodes de SA pour la segmentation des OARs sur des images CT HN.

Cette étude a fourni une comparaison détaillée entre 4 solutions basées sur une bibliothèque d'atlas (ABAS) et 2 solutions d'apprentissage profond (DL) pour la délimitation des OAR sur des images CT HN. Leurs performances ont été évaluées par rapport à plusieurs aspects qui sont pertinents lors de l'examen d'une solution AS, notamment : la demande de ressources en données du patient, le temps de calcul, la précision géométrique (recouvrement volumétrique et métrique de la distance de surface), le temps de correction manuelle et l'impact dosimétrique (en utilisant la planification automatique). Les résultats ont montré que les solutions DL avaient une précision globalement supérieure à celle des méthodes ABAS. Il a également été démontré que les contours hybrides ABAS présentaient un bon accord avec les contours de référence et étaient parfois meilleurs que les contours basés sur la méthode DL. Toutefois, si l'on considère le temps de calcul et le temps consacré aux corrections manuelles, les solutions DL se sont révélées plus efficaces.

En ce qui concerne les données nécessaires à l'entraînement d'une solution DL, comme d'autres études dans la littérature, notre étude a montré qu'avec un ensemble de données d'entraînement limité mais plus uniforme, un modèle peut obtenir des résultats homogènes pour la plupart des structures de HN. Cela peut être d'un grand intérêt pour les centres qui souhaitent adapter un modèle à leurs pratiques de contourage. En même temps, nous avons également démontré qu'un modèle entraîné avec une plus grande quantité de données multicentriques peut fournir une bonne généralisation à de nouveaux ensembles de données. L'exécution de corrections manuelles a été la plus efficace sur les contours du modèle DL monocentrique (en moyenne 18 minutes). Ceci est particulièrement significatif pour la charge de travail des dosimétristes. Ces considérations peuvent aider un service à choisir la solution de AS la mieux adaptée à ses besoins en fonction du temps et des ressources dont il dispose. En ce qui concerne l'impact dosimétrique, aucune différence statistique n'a été observée entre les plans créés avec AS sans ou avec des corrections manuelles. Ceci est cohérent avec les résultats de la littérature qui attestent que les corrections manuelles pourraient potentiellement être négligées pour les OAR. Par ailleurs, dans notre étude, nous avons examiné la corrélation entre la position de l'organe par rapport aux PTV et les différences de dose. L'objectif était de comprendre dans quelles situations la correction d'un contour AS est particulièrement importante. Malheureusement, nous n'avons pas réussi à identifier une tendance claire de cette relation car des différences de dose élevées n'ont pas été systématiquement observées à proximité de la cible. L'étude est originale car elle a évalué plusieurs solutions AS commerciales et non commerciales, parmi lesquelles 3 d'entre elles n'ont pas encore été étudiées sur la localisation des HN. Une autre nouveauté était l'utilisation d'une solution d'auto-planification dans l'étude dosimétrique pour éliminer le biais du planificateur.

Nous reconnaissons les limites de l'étude, en termes de petite cohorte de patients testés et de contours de référence provenant d'un seul expert. Cependant, l'ensemble de données hétérogènes pour les essais a permis de mettre à l'épreuve les différents algorithmes de AS. De surcroît, cette étude s'est délibérément concentrée sur une approche monocentrique, dans le but de déterminer quelle solution disponible dans le service était la plus précise et nécessitait moins de ressources en termes de données sur les patients et de main-d'œuvre. Cependant, les résultats statistiques pourraient bénéficier d'un plus grand nombre de patients inclus dans la cohorte de contrôle et d'un plus grand nombre d'observateurs impliqués dans la

tâche de correction manuelle. Les perspectives futures de ce travail incluent l'augmentation de la base de données pour la formation et l'incorporation de cas de patients opérés et non opérés.

Enfin, ces résultats sont d'un grand intérêt pour le développement des flux de travail de l'ART pour les patients HN car ils prouvent l'efficacité améliorée du flux de travail lors de l'utilisation de l'AS pour la délimitation de l'OAR combinée à des plans de traitement générés à l'aide d'une solution de planification automatique. La délimitation manuelle de la cible primaire reste une limite de temps, tandis que la AS des niveaux de ganglions lymphatiques qui sont habituellement irradiés comme cible secondaire est discutée dans le prochain chapitre du manuscrit.

Chapitre 4 Évaluation de différents algorithmes pour la segmentation automatique des ganglions lymphatiques de la tête et du cou sur des images CT

L'étude représente la troisième contribution de ce projet de doctorat et est une continuation du travail précédent sur l'AS des OAR. Nous avons comparé les 4 solutions ABAS et les 2 solutions DL pour la délimitation des niveaux CTVn 2,3 et 4, qui sont typiquement irradiés comme cibles secondaires dans la localisation HN. De plus, nous avons évalué la précision de leur union car ce volume est généralement utilisé pour la planification du traitement. Nous avons analysé les contours en termes de : temps de calcul, précision spatiale, acceptabilité clinique (évaluée par 4 médecins), temps nécessaire pour effectuer des corrections manuelles (effectuées par l'un des observateurs) et conséquences dosimétriques lors de l'utilisation des contours AS dans la planification du traitement.

Comme dans l'étude précédente sur les OAR, nous avons observé que les solutions DL offraient une meilleure précision par rapport aux méthodes ABAS pour la segmentation du CTVn sur les images CT. Cependant, les méthodes ABAS étaient également capables d'atteindre des DICE>0,80 en considérant l'union des volumes. Cela a révélé une observation importante, le fait que des imprécisions de contour significatives se produisent à la limite entre les niveaux, aux extrémités supérieures et inférieures. Le principal avantage d'une méthode ABAS était la petite quantité de ressources nécessaires. Cependant, l'inconvénient était le temps de calcul (6-10 minutes par patient, pour une bibliothèque de 10 atlas). En revanche, les méthodes DL ont été entraînées sur des bases de données plus importantes (49 patients pour la solution DL.1 et >100 patients pour la solution DL.2) mais les contours ont été générés plus rapidement (<1min et <2min, respectivement). De plus, tous les contours générés par le modèle commercial multicentrique DL.2 ont été considérés comme cliniquement acceptables sans ou avec seulement des corrections mineures, qui étaient en moyenne de 1min06sec par patient. Ces résultats remarquables constituent des éléments clés pour des flux de travail très efficaces. Par rapport à la solution DL monocentrique, aucune différence statistiquement significative n'a été identifiée entre les indices de précision géométrique (DICE et HD95%). Cependant, le temps de correction manuelle était plus important (4min10sec contre 1min06sec, en moyenne). Globalement, l'évaluation clinique des 4 experts a conduit à des conclusions similaires pour les contours DL.1 et ABAS.2. Cependant, la réalisation des corrections manuelles a pris plus de temps pour les contours ABAS.2 (6min31sec vs 4min10sec). Notamment, une bonne concordance (DICE=0.85) entre l'expert de référence et l'autre médecin effectuant les corrections manuelles a été observée, ce qui est supérieur à la variabilité entre des experts précédemment rapporté dans les études de la littérature [216,337]. Une mesure de l'IOV entre experts n'a pas été réalisée dans cette étude, mais elle constitue une perspective future de ce travail. L'évaluation de contours provenant de plusieurs observateurs, de contours délimités manuellement et de contours basés sur la AS ajustés manuellement, permettra d'évaluer l'utilité de la AS pour améliorer l'adhérence des pratiques de contourage.

Lors de l'analyse des distributions de dose, un sous-dosage significatif dans le PTV secondaire a été observé indépendamment de la solution utilisée. La perte de couverture a été particulièrement détectée

sur le CTVn4, ce qui pourrait être lié aux divergences précédemment identifiées dans la superposition géométrique de volumes de référence. Pour les méthodes ABAS et DL, les résultats de DICE pour CTVn4 étaient ≤ 0.72 , ce qui était cohérent avec les résultats de la littérature [334,338]. Pour ces contours, nous recommandons une plus grande attention, avant de les utiliser dans la planification du traitement.

À notre connaissance, la présente étude a examiné pour la première fois 5 méthodes d'AS (3 multi-ABAS et 2 DL) pour segmenter 3 niveaux CTVn distincts. De plus, la planification automatique a été utilisée pour évaluer les conséquences dosimétriques, ce qui a permis de réduire le travail, de ne pas avoir de facteur de planificateur et d'obtenir un effet isolé provenant uniquement du contour du CTVn. Les perspectives futures de ce travail incluent l'augmentation de la base de données de formation du modèle DL (incluant peut-être aussi des patients opérés), ainsi que la cohorte de patients de test. En outre, un nouveau cadre de réseau pourrait être développé qui combine les repères anatomiques des directives de contournage du CTVn afin de guider les prédictions des niveaux de CTVn.

Enfin, l'étude démontre que les méthodes AS pour le CTVn peuvent être intégrées dans les flux de travail de la RT afin de réduire le temps consacré à la délimitation manuelle. À l'heure actuelle, la combinaison proposée de la AS pour les OAR et le CTVn, associée à des solutions de planification automatique, pourrait être utilisée pour améliorer l'efficacité des cas complexes de HN. Seul le volume de la tumeur primaire resterait à délimiter manuellement par les médecins. Cependant, pour permettre l'ART, les variations anatomiques intra-fractionnelles doivent être prises en compte. À cet égard, le chapitre suivant abordera différentes méthodes de génération de CT synthétique à partir d'images CBCT du jour, qui peuvent être utilisées pour calculer des plans sur l'anatomie du jour.

Chapitre 5. Évaluation de différentes méthodes pour la génération d'images CT synthétiques à partir d'images CBCT du jour

Cette étude a fourni une évaluation et une comparaison de 4 méthodes de calcul de dose basées sur des images CBCT. Leurs performances ont été évaluées en termes de précision dosimétrique par rapport aux images CT de planification. La charge de travail et les ressources nécessaires pour chaque méthode ont également été prises en compte.

Les résultats ont montré que l'algorithme à base de recalage déformable (DIR) avait les meilleures performances, suivi par la méthode DL. De plus, les deux méthodes ont été rapides, générant une image sCT en moins d'une minute. La méthode d'affectation de la densité de classe (DAM) a donné des résultats similaires à ceux de la méthode DL, mais elle a demandé plus de temps (~5min) en raison des étapes supplémentaires liées à la segmentation des classes de tissus et à l'affectation des numéros HU. Enfin, la courbe HU-ED adaptée du CBCT a donné les pires résultats.

Il existe plusieurs métriques pour quantifier la précision des images sCT [207], mais aucun consensus réel sur les critères d'acceptation. Dans notre étude, nous avons quantifié l'erreur moyenne (ME) et l'erreur absolue moyenne (MAE) pour toutes les méthodes à l'intérieur du contour du corps de l'image CBCT. Certaines autres études documentent les résultats en fonction des segments de tissus (par exemple, les structures osseuses, les cavités aériennes, les tissus mous). Il serait intéressant d'approfondir cette étude dans notre cohorte de patients. En raison de la comparaison par voxel, une attention particulière doit être portée à la compensation d'éventuelles différences anatomiques et/ou d'erreurs potentielles dans l'étape initiale d'enregistrement entre les images CT et CBCT. Une solution proposée par la littérature consiste à appliquer des marges négatives (par exemple, 2 mm) aux segments où la différence d'indice HU doit être évaluée [317].

En ce qui concerne la précision du calcul de la dose, les différences de dose basées sur l'histogramme dose-volume (DVH) et l'analyse gamma sont généralement effectuées pour comparer les plans calculés sur des images sCT et des images CT de référence. Dans notre étude, les objectifs pour les volumes cibles

et les contraintes de l'OAR ont été pris en compte en fonction du protocole clinique et deux critères gamma ont été utilisés (3%/3mm et 2%/2mm). Nous avons observé que les écarts de dose étaient généralement faibles et non significatifs. Des écarts plus importants ont été identifiés avec les critères d'analyse gamma les plus stricts, alors que seule la méthode DIR a passé >90% le critère gamma de 2%/2mm à l'intérieur de toutes les structures considérées. Une meilleure compréhension de l'emplacement exact des chiffres HU incohérents permettrait peut-être de trouver des corrélations avec les écarts de dose et les débits gamma. A terme, cela permettrait de définir des limites appropriées pour les critères d'acceptation.

L'étude est originale car un cycleGAN qui a été entraîné avec des images CBCT non appariées provenant d'une base de données multicentrique, a été évalué sur la localisation HN. Le modèle entraîné sur des données monocentriques a semblé fournir de meilleurs résultats dans d'autres études [82,310]. Une autre question qui doit être étudiée plus en détail est le problème de la coupure de l'image causée par le champ de vue limité du CBCT, qui dans notre étude a été résolu en remplaçant la densité de l'eau lorsque les informations de l'image étaient manquantes. En outre, la précision du DIR doit être évaluée par rapport aux différentes fractions du traitement, ainsi que la précision du DIR pour l'accumulation de la dose doit être évaluée [100].

À condition d'être intégrées dans le système de planification des traitements, les méthodes DIR et DL sont les solutions les plus prometteuses pour permettre la planification des ART par la génération de sCT à partir d'images CBCT du jour.

Chapitre 6. Conclusions et perspectives de l'étude

Finalement, le travail effectué pendant les 3 années de thèse, visait à évaluer différentes méthodes pour accélérer le flux de travail de la RT et finalement permettre de réaliser l'ART pour les patients atteints de cancer HN. Les contributions ont été divisées en quatre axes, et nous avons notamment évalué : 1) Les performances d'une solution de planification automatique basée sur un algorithme MCO a priori ; 2) Six solutions de segmentation automatique des OAR et CTvN sur des images CT ; 3) Quatre méthodes permettant de calculer la dose sur des images CBCT du jour. Une intégration sans faille de ces solutions représenterait la réussite de l'ART pour les patients atteints de cancers HN.

En ce qui concerne la vitesse et la précision, les solutions DL étaient les plus intéressantes. Cependant, la collecte et la gestion de la base de données d'entraînement nécessitent un travail considérable. D'autres méthodes basées sur moins d'informations a priori ('wis-list' pour la planification automatique et segmentation basée sur une bibliothèque d'atlas pour le contourage automatique) ont permis de réduire la charge clinique. Plus précisément, nous avons démontré que l'optimisation manuelle peut être remplacée en toute sécurité par une planification automatique basée sur une 'wis-list', et que le temps de contourage manuel peut être raccourci en effectuant des corrections sur les contours générés automatiquement. De plus, les deux méthodes permettent une meilleure cohérence entre les opérateurs. Leur limite reste le temps de calcul. Enfin, nous avons observé que la méthode la plus rapide et la plus précise pour générer des sCT à partir d'images CBCT du jour était l'utilisation des méthodes DIR et DL. Une limitation reste le FOV limité dans les images CBCT, qui dans notre étude a été résolu par le remplacement de la densité de l'eau où les informations de l'image étaient manquantes.

De nombreuses perspectives d'avenir peuvent être envisagées à partir de ce travail. Les perspectives à court terme en ce qui concerne la planification automatique sont de tester et d'évaluer les solutions plus rapides promises par la société Elekta, et de les étudier également sur une plus grande cohorte de patients atteints de cancer HN, où la population peut être divisée en groupes plus petits en fonction de la localisation de la tumeur. En second lieu, plusieurs 'wis-lists' peuvent être établies en fonction du protocole clinique et des sites de cancer, et encourager leur intégration dans la routine clinique. Les perspectives de développement d'une solution basée sur la DL pour l'auto-planification peuvent également être envisagées, afin de générer des plans en quelques secondes seulement. Par ailleurs, la

méthodologie utilisée pour quantifier la qualité des plans peut être appliquée pour évaluer d'autres solutions de planification automatique avant leur mise en œuvre clinique dans notre service. Cependant, un obstacle à l'intégration harmonieuse de l'algorithme d'auto-planification basé sur la 'wish-list' dans la routine clinique reste la nomenclature des structures d'intérêt, où la solution AS peut être le facilitateur principal.

En ce qui concerne les solutions AS, nous avons évalué plusieurs solutions disponibles ou non dans le commerce. Parmi elles, certaines ont été étudiées pour la première fois sur des OARs HN et/ou des volumes CTVn. A partir de ce travail, des perspectives à court terme peuvent être proposées en commençant par une étude visant à évaluer la relation entre la quantité de données d'entraînement et la performance du modèle DL pour chacune des structures. En parallèle, l'open-source nnUnet network peut être entraîné avec les mêmes données d'entrée, et les résultats peuvent être comparés afin d'étudier l'interaction entre les données d'entraînement et l'architecture du réseau. Ceci a déjà été initié dans le travail d'un étudiant en master. En outre, il est possible d'étudier les performances du modèle lorsque des patients opérés sont inclus dans la formation. Nous avons déjà testé cette option en incluant quelques cas de patients ayant subi une ablation du larynx ou de la trachée. Cependant, les premiers résultats n'étaient pas satisfaisants et nous avons conclu qu'il fallait inclure encore plus de patients opérés pour augmenter potentiellement les performances du modèle. De plus, une perspective pour le futur en ce qui concerne la base de données de patients, est d'inclure également des données provenant d'autres centres, qui doivent néanmoins être cohérentes avec le protocole d'acquisition d'images (tranche de 2mm d'épaisseur, injection de produit de contraste) et suivre les directives internationales de contourage. Il est ainsi possible de réaliser une étude multicentrique permettant de tester la généralisation du modèle DL monocentrique par rapport aux contourages manuelles de référence d'autres médecins. De plus, l'inclusion de contours manuels provenant de plusieurs observateurs permettrait d'évaluer l'IOV parmi les contourages manuelles et de le comparer à l'IOV des AS avec des corrections manuelles. Comme dans une autre étude [170], il serait également intéressant de diviser les observations en catégories basées sur l'expérience des médecins, à savoir les résidents, les jeunes médecins et les médecins seniors. Lors de la réalisation de l'étude proposée, le temps de délimitation manuelle devrait être enregistré afin que la réduction réelle du temps puisse être évaluée. En ce qui concerne l'évaluation de l'impact dosimétrique, nous avons démontré dans notre étude que la distance de l'OAR au PTV n'était pas systématiquement liée aux écarts dosimétriques élevés. Une étude plus approfondie pourrait peut-être être réalisée afin de déterminer dans quelles situations l'édition des contours de l'OAR fait une différence clinique, et si des limites pour DICE/HD95% pourraient être fixés. Par ailleurs, une étude a montré que les mesures DICE de surface et APL (added path length) pour la précision des contours étaient de meilleurs indicateurs du temps de délimitation clinique gagné lors de l'utilisation des contours AS [345]. Il serait intéressant de comparer ces mesures également sur les cohortes de nos études. Une autre perspective de recherche future inspirée de la littérature [334,346] serait de développer un cadre de réseau DL avec des contraintes de contour basées sur des repères anatomiques. Les structures dont les limites sont bien définies sur les images CT et qui constituent des limites anatomiques définies dans les directives de délimitation (par exemple, l'os hypoïde, le muscle sternocléidomastoïdien, les vertèbres) peuvent être segmentées comme tâche auxiliaire d'un réseau DL. Ensuite, les modèles de distance peuvent être appliqués pour restreindre la forme des prédictions des volumes OAR et CTVn.

En ce qui concerne la dernière contribution de la thèse, sur les calculs de dose basés sur le CBCT, nous pouvons d'abord envisager comme prochaine étape, l'optimisation de la solution DL pour la génération de sCT. Un nouveau modèle pourrait être entraîné sur des données uniformes de notre département (modèle monocentrique), en essayant également d'optimiser l'architecture du réseau (nombre de couches, taille de l'image, fonction de perte, poids et seuils, etc.). Ses performances doivent ensuite être comparées à celles du modèle multicentrique. En second lieu, une autre fonction disponible dans Monaco TPS (AdaptAnatomy) pour la génération de sCT basée sur des densités modifiées devrait être étudiée. Pour

cela, plusieurs stratégies peuvent être suivies en ce qui concerne les régions d'intérêt (ROI): on peut utiliser l'ensemble des structures présentes sur le set CT ou choisir seulement quelques classes de tissus. Pour automatiser le processus, on peut utiliser une solution AS basée sur ABAS pour segmenter une liste prédéfinie de structures osseuses et de tissus mous sur le CT. Comme il s'agit d'une autre solution de génération de sCT intégrée au TPS, il serait pertinent de la comparer avec les méthodes étudiées précédemment. La problématique de l'accumulation des doses est une autre direction de recherche de ce sujet. La précision du DIR doit encore être évaluée afin que les incertitudes du calcul de la dose puissent être estimées [100]. En outre, les incertitudes liées à la génération de sCT par CBCT pourraient se propager à partir des erreurs de repositionnement du patient au jour le jour. Par conséquent, les écarts de dose doivent également être étudiés en fonction des stratégies de repositionnement du patient. Une étude a déjà été lancée pour évaluer les incertitudes dosimétriques liées à la registration par alignement osseux (BM) seulement, et au recalage guidé par BM+tissus mous entre le CBCT du jour et les images CT. Enfin, il est nécessaire d'étudier et de proposer des critères d'assurance qualité pour la mise en œuvre clinique d'une méthode sCT.

Les perspectives à long terme seraient d'évaluer un flux de travail de RT qui combine des solutions automatisées pour l'auto-contournement, l'auto-planification et la génération d'images sCT. Alors que les techniques complexes d'IMRT ont permis d'atteindre une couverture exceptionnelle du volume cible, l'intégration de solutions modernes de RT permet de se concentrer davantage sur la minimisation de la toxicité pour les tissus normaux. Cela nécessite toutefois des contrôles de qualité à chaque étape du flux de travail. Il serait cliniquement utile d'évaluer les stratégies de marge en tenant compte de toutes les incertitudes liées aux solutions automatisées utilisées. Un autre objectif est de permettre la sélection des patients qui bénéficieront d'une adaptation du plan. Cela pourrait être fait sur la base des variations anatomiques intra-fractionnelles, où nous pourrions comparer la superposition des volumes des OAR et CTVn utilisés pour la planification, avec les volumes automatiquement segmentés sur le sCT généré à partir des images CBCT de pré-traitement. Les limites critiques peuvent être étudiées en évaluant les doses planifiées par rapport aux doses créées sur l'anatomie du jour.

Finalement, la validation clinique de ce flux de travail de RT réduirait la charge de travail manuelle, aiderait à l'harmonisation des pratiques cliniques et permettrait une prise de décision rapide pour la RT des cas complexes de cancer HN. Les résultats des essais cliniques sont attendus pour déterminer les avantages cliniques de l'ART.

Bibliography

- [1] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin* 2021;71:209–49. <https://doi.org/10.3322/CAAC.21660>.
- [2] Biston M-C, Costea M, Gassa F, Serre A-A, Voet P, Larson R, et al. Evaluation of fully automated a priori MCO treatment planning in VMAT for head-and-neck cancer. *Phys Medica* 2021;87:31–8. <https://doi.org/10.1016/j.ejmp.2021.05.037>.
- [3] Costea M, Zlate A, Durand M, Baudier T, Grégoire V, Sarrut D, et al. Comparison of atlas-based and deep learning methods for organs at risk delineation on head-and-neck CT images using an automated treatment planning system. *Radiother Oncol* 2022;177:61–70. <https://doi.org/10.1016/J.RADONC.2022.10.029>.
- [4] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. 394 CA: A Cancer Journal for Clinicians Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA CANCER J CLIN* 2018;68:394–424. <https://doi.org/10.3322/caac.21492>.
- [5] Hashibe M, Brennan P, Chuang SC, Boccia S, Castellsague X, Chen C, et al. Interaction between tobacco and alcohol use and the risk of head and neck cancer: pooled analysis in the International Head and Neck Cancer Epidemiology Consortium. *Cancer Epidemiol Biomarkers Prev* 2009;18:541–50. <https://doi.org/10.1158/1055-9965.EPI-08-0347>.
- [6] Gillison ML, D’Souza G, Westra W, Sugar E, Xiao W, Begum S, et al. Distinct risk factor profiles for human papillomavirus type 16-positive and human papillomavirus type 16-negative head and neck cancers. *J Natl Cancer Inst* 2008;100:407–20. <https://doi.org/10.1093/JNCI/DJN025>.
- [7] Chaturvedi AK, Engels EA, Pfeiffer RM, Hernandez BY, Xiao W, Kim E, et al. Human papillomavirus and rising oropharyngeal cancer incidence in the United States. *J Clin Oncol* 2011;29:4294–301. <https://doi.org/10.1200/JCO.2011.36.4596>.
- [8] Amin MB, Edge SB, Greene FL, Schilsky RL, Brookland RK, Washington MK, et al. American Joint Committee on Cancer (AJCC). *AJCC Cancer Staging Manual*. 2017.
- [9] Shah JP. Staging for Head and Neck Cancer: Purpose, Process and Progress. *Indian J Surg Oncol* 2018;9:116. <https://doi.org/10.1007/S13193-018-0723-0>.
- [10] Lin A, Kim HM, Terrell JE, Dawson LA, Ship JA, Eisbruch A. Quality of life after parotid-sparing IMRT for head-and-neck cancer: A prospective longitudinal study. *Int J Radiat Oncol Biol Phys* 2003;57:61–70. [https://doi.org/10.1016/S0360-3016\(03\)00361-4](https://doi.org/10.1016/S0360-3016(03)00361-4).
- [11] Eisbruch A, Kim HM, Terrell JE, al. et. Xerostomia and its predictors following parotid-sparing irradiation of head-and-neck cancer. *Int J Radiat Oncol Biol Phys* 2001;50:695–704.
- [12] Peponi E, Glanzmann C, Willi B, Huber G, Studer G. Dysphagia in head and neck cancer patients following intensity modulated radiotherapy (IMRT) 2011;6:1–8.
- [13] PC Levendag DTPV. Dysphagia disorders in patients with cancer of the oropharynx are significantly affected by the radiation therapy dose to the superior and middle constrictor muscle: A dose-effect relationship. *Radiother Oncol* 2007;85:64–73. <https://doi.org/10.1016/j.radonc.2007.07.009>.
- [14] Steel GG, Mcmillan TJ, Peacock JH. The 5Rs of Radiobiology. *Int J Radiat Biol* 1989;56:1045–8. <https://doi.org/10.1080/09553008914552491>.

- [15] Song HQ, Zhou ZJ, Li LP. The linear quadratic model: usage, interpretation and challenges. *Phys Med Biol* 2018;64:01TR01. <https://doi.org/10.1088/1361-6560/AAF26A>.
- [16] Ezzell GA, Galvin JM, Low D, Palta JR, Rosen I, Sharpe MB, et al. Guidance document on delivery, treatment planning, and clinical implementation of IMRT: Report of the IMRT subcommittee of the AAPM radiation therapy committee. *Med Phys* 2003;30:2089–115. <https://doi.org/10.1118/1.1591194>.
- [17] Teoh M, Clark CH, Wood K, Whitaker S, Nisbet A. Volumetric modulated arc therapy: A review of current literature and clinical use in practice. *Br J Radiol* 2011;84:967–96. <https://doi.org/10.1259/BJR/22373346>.
- [18] Osborn J. Is VMAT beneficial for patients undergoing radiotherapy to the head and neck? *Radiography* 2017;23:73–6. <https://doi.org/10.1016/j.radi.2016.08.008>.
- [19] Nutting CM, Morden JP, Harrington KJ, Urbano TG, Bhide SA, Clark C, et al. Parotid-sparing intensity modulated versus conventional radiotherapy in head and neck cancer (PARSPORT): A phase 3 multicentre randomised controlled trial. *Lancet Oncol* 2011. [https://doi.org/10.1016/S1470-2045\(10\)70290-4](https://doi.org/10.1016/S1470-2045(10)70290-4).
- [20] Buciuman N, Marcu LG. Dosimetric justification for the use of volumetric modulated arc therapy in head and neck cancer-A systematic review of the literature. *Laryngoscope Investig Otolaryngol* 2021;6:999–1007. <https://doi.org/10.1002/lio2.642>.
- [21] Cilla S, Deodato F, Digesù C, Macchia G, Picardi V, Ferro M, et al. Assessing the feasibility of volumetric-modulated arc therapy using simultaneous integrated boost (SIB-VMAT): An analysis for complex head-neck, high-risk prostate and rectal cancer cases. *Med Dosim* 2014;39:108–16. <https://doi.org/10.1016/j.meddos.2013.11.001>.
- [22] Peters S, Schiefer H, Plasswilm L. A treatment planning study comparing Elekta VMAT and fixed field IMRT using the varian treatment planning system eclipse. *Radiat Oncol* 2014;9. <https://doi.org/10.1186/1748-717X-9-153>.
- [23] Mallick S, Benson R, Julka PK, Rath GK. Altered fractionation radiotherapy in head and neck squamous cell carcinoma. *J Egypt Natl Canc Inst* 2016;28:73–80. <https://doi.org/10.1016/J.JNCI.2016.02.004>.
- [24] Leech M, Coffey M, Mast M, Moura F, Osztavics A, Pasini D, et al. ESTRO ACROP guidelines for positioning, immobilisation and position verification of head and neck patients for radiation therapists. *Tech Innov Patient Support Radiat Oncol* 2017;1:1–7. <https://doi.org/10.1016/J.TIPSR0.2016.12.001>.
- [25] Grégoire V, Boisbouvier S, Giraud P, Maingon P, Pointreau Y, Vieilleuvigne L. Management and work-up procedures of patients with head and neck malignancies treated by radiation. *Cancer/Radiotherapie* 2022;26:147–55. <https://doi.org/10.1016/j.canrad.2021.10.005>.
- [26] Brouwer CL, Steenbakkers RJHM, Bourhis J, Budach W, Grau C, Grégoire V, et al. CT-based delineation of organs at risk in the head and neck region: DAHANCA, EORTC, GORTEC, HKNPCSG, NCIC CTG, NCRI, NRG Oncology and TROG consensus guidelines. *Radiother Oncol* 2015;117:83–90. <https://doi.org/10.1016/J.RADONC.2015.07.041>.
- [27] Grégoire V, Ang K, Budach W, Grau C, Hamoir M, Langendijk JA, et al. Delineation of the neck node levels for head and neck tumors: A 2013 update. DAHANCA, EORTC, HKNPCSG, NCIC CTG, NCRI, RTOG, TROG consensus guidelines. *Radiother Oncol* 2014;110:172–81. <https://doi.org/10.1016/j.radonc.2013.10.010>.

- [28] Grégoire V, Evans M, Le QT, Bourhis J, Budach V, Chen A, et al. Delineation of the primary tumour Clinical Target Volumes (CTV-P) in laryngeal, hypopharyngeal, oropharyngeal and oral cavity squamous cell carcinoma: AIRO, CACA, DAHANCA, EORTC, GEORCC, GORTEC, HKNPCSG, HNCIG, IAG-KHT, LPRHHT, NCIC CTG, NCRI, NRG Oncology, PHNS, SBRT, SOMERA, SRO, SSHNO, TROG consensus guidelines. *Radiother Oncol* 2018;126:3–24. <https://doi.org/10.1016/J.RADONC.2017.10.016>.
- [29] Grégoire V, Levendag P, Ang KK, Bernier J, Braaksma M, Budach V, et al. CT-based delineation of lymph node levels and related CTVs in the node-negative neck: DAHANCA, EORTC, GORTEC, NCIC, RTOG consensus guidelines. *Radiother Oncol* 2003;69:227–36. <https://doi.org/10.1016/J.RADONC.2003.09.011>.
- [30] Grégoire V, Coche E, Cosnard G, Hamoir M, Reychler H. Selection and delineation of lymph node target volumes in head and neck conformal radiotherapy. Proposal for standardizing terminology and procedure based on the surgical experience. *Radiother Oncol* 2000;56:135–50. [https://doi.org/10.1016/S0167-8140\(00\)00202-4](https://doi.org/10.1016/S0167-8140(00)00202-4).
- [31] News AW-I, 1999 undefined. ICRU report 62, prescribing, recording and reporting photon beam therapy (supplement to ICRU Report 50). CiNiiAcJp n.d.
- [32] Purdy JA. Current ICRU Definitions of Volumes: Limitations and Future Directions. *Semin Radiat Oncol* 2004;14:27–40. <https://doi.org/10.1053/j.semradonc.2003.12.002>.
- [33] Berthelsen AK, Dobbs J, Kjellén E, Landberg T, Möller TR, Nilsson P, et al. What's new in target volume definition for radiologists in ICRU Report 71? How can the ICRU volume definitions be integrated in clinical practice? *Cancer Imaging* 2007;7:104. <https://doi.org/10.1102/1470-7330.2007.0013>.
- [34] Van Herk M. Errors and Margins in Radiotherapy. *Semin Radiat Oncol* 2004;14:52–64. <https://doi.org/10.1053/j.semradonc.2003.10.003>.
- [35] Grégoire V, Mackie TR. State of the art on dose prescription, reporting and recording in Intensity-Modulated Radiation Therapy (ICRU report No. 83). *Cancer Radiother* 2011;15:555–9. <https://doi.org/10.1016/J.CANRAD.2011.04.003>.
- [36] Kearney M, Coffey M, Leong A. A review of Image Guided Radiation Therapy in head and neck cancer from 2009–2019 – Best Practice Recommendations for RTTs in the Clinic. *Tech Innov Patient Support Radiat Oncol* 2020;14:43–50. <https://doi.org/10.1016/J.TIPSR0.2020.02.002>.
- [37] Walter C, Boda-Heggemann J, Wertz H, Loeb I, Rahn A, Lohr F, et al. Phantom and in-vivo measurements of dose exposure by image-guided radiotherapy (IGRT): MV portal images vs. kV portal images vs. cone-beam CT. *Radiother Oncol* 2007;85:418–23. <https://doi.org/10.1016/J.RADONC.2007.10.014>.
- [38] Goyal S, Kataria T. *Image Guidance in Radiation Therapy: Techniques and Applications* 2014. <https://doi.org/10.1155/2014/705604>.
- [39] Jaffray DA. Kilovoltage volumetric imaging in the treatment room. *Front Radiat Ther Oncol* 2007;40:116–31. <https://doi.org/10.1159/000106031>.
- [40] Welsh JS, Patel RR, Ritter MA, Harari PM, Mackie TR, Mehta MP. Helical tomotherapy: an innovative technology and approach to radiation therapy. *Technol Cancer Res Treat* 2002;1:311–6. <https://doi.org/10.1177/153303460200100413>.
- [41] Ruchala KJ, Olivera GH, Schloesser EA, Mackie TR. Megavoltage CT on a tomotherapy system. *Phys*

Med Biol 1999;44:2597. <https://doi.org/10.1088/0031-9155/44/10/316>.

- [42] Van Gestel D, Van Vliet-Vroegindeweyj C, Van Den Heuvel F, Crijns W, Coelmont A, De Ost B, et al. RapidArc, SmartArc and TomoHD compared with classical step and shoot and sliding window intensity modulated radiotherapy in an oropharyngeal cancer treatment plan comparison 2013. <https://doi.org/10.1186/1748-717X-8-37>.
- [43] Van Gestel D, De Kerf G, Wouters K, Crijns W, Vermorcken JB, Gregoire V, et al. Fast Helical Tomotherapy in a head and neck cancer planning study: is time priceless? *Radiat Oncol* 2015;10. <https://doi.org/10.1186/S13014-015-0556-8>.
- [44] Pathmanathan AU, van As NJ, Kerkmeijer LGW, Christodouleas J, Lawton CAF, Vesprini D, et al. Magnetic Resonance Imaging-Guided Adaptive Radiation Therapy: A “Game Changer” for Prostate Treatment? *Int J Radiat Oncol Biol Phys* 2018;100:361–73. <https://doi.org/10.1016/J.IJROBP.2017.10.020>.
- [45] Murray J, Tree AC. Prostate cancer - Advantages and disadvantages of MR-guided RT. *Clin Transl Radiat Oncol* 2019;18:68–73. <https://doi.org/10.1016/J.CTRO.2019.03.006>.
- [46] Van Der Put RW, Kerkhof EM, Raaymakers BW, Jürgenliemk-Schulz IM, Lagendijk JJW. Contour propagation in MRI-guided radiotherapy treatment of cervical cancer: the accuracy of rigid, non-rigid and semi-automatic registrations. *Phys Med Biol* 2009;54:7135–50. <https://doi.org/10.1088/0031-9155/54/23/007>.
- [47] Maziero D, Straza MW, Ford JC, Bovi JA, Diwanji T, Stoyanova R, et al. MR-Guided Radiotherapy for Brain and Spine Tumors. *Front Oncol* 2021;11:626100. <https://doi.org/10.3389/FONC.2021.626100>.
- [48] Chuter RW, Pollitt A, Whitehurst P, Mackay RI, Van Herk M, McWilliam A. Assessing MR-linac radiotherapy robustness for anatomical changes in head and neck cancer. *Phys Med Biol* 2018;63. <https://doi.org/10.1088/1361-6560/AAC749>.
- [49] HIn BNC, Nutting C, Newbold K, Bhide S, McQuaid D, Dunlop A, et al. The impact of restricted length of treatment field and anthropometric factors on selection of head and neck cancer patients for treatment on the MR-L linac. *Br J Radiol* 2020;93. https://doi.org/10.1259/BJR.20200023/SUPPL_FILE/.
- [50] Steinmann A, Alvarez P, Lee H, Court L, Stafford R, Sawakuchi G, et al. MRIGRT head and neck anthropomorphic QA phantom: Design, development, reproducibility, and feasibility study. *Med Phys* 2020;47:604. <https://doi.org/10.1002/MP.13951>.
- [51] Boeke S, Mönnich D, Van Timmeren JE, Balermipas P. MR-Guided Radiotherapy for Head and Neck Cancer: Current Developments, Perspectives, and Challenges n.d. <https://doi.org/10.3389/fonc.2021.616156>.
- [52] Mutic S, Dempsey JF. The ViewRay system: magnetic resonance-guided and controlled radiotherapy. *Semin Radiat Oncol* 2014;24:196–9. <https://doi.org/10.1016/J.SEMRADONC.2014.02.008>.
- [53] Roberts DA, Sandin C, Vesanen PT, Lee H, Hanson IM, Nill S, et al. Machine QA for the Elekta Unity system: A Report from the Elekta MR-linac consortium. *Med Phys* 2021;48:e67–85. <https://doi.org/10.1002/MP.14764>.
- [54] Grégoire V, Guckenberger M, Haustermans K, Lagendijk JJW, Ménard C, Pötter R, et al. Image guidance in radiation therapy for better cure of cancer. *Mol Oncol* 2020;14:1470–91.

<https://doi.org/10.1002/1878-0261.12751>.

- [55] Lim-Reinders S, Keller BM, Al-Ward S, Sahgal A, Kim A. Online Adaptive Radiation Therapy. *Int J Radiat Oncol Biol Phys* 2017;99:994–1003. <https://doi.org/10.1016/J.IJROBP.2017.04.023>.
- [56] Morgan HE, Sher DJ. Adaptive radiotherapy for head and neck cancer. *Cancers Head Neck* 2020 51 2020;5:1–16. <https://doi.org/10.1186/S41199-019-0046-Z>.
- [57] Castadot P, Lee JA, Geets X, Grégoire V. Adaptive Radiotherapy of Head and Neck Cancer. *Semin Radiat Oncol* 2010;20:84–93. <https://doi.org/10.1016/J.SEMRADONC.2009.11.002>.
- [58] Castelli J, Simon A, Rigaud B, Chajon E, Thariat J, Benezery K, et al. Adaptive radiotherapy in head and neck cancer is required to avoid tumor underdose. *Acta Oncol* 2018;57:1267–70. <https://doi.org/10.1080/0284186X.2018.1468086>.
- [59] Zia SY, Smith W, Ozbek U, Mirza A, Sheu R, Ghafar R, et al. The impact of weight loss on setup accuracy for head and neck cancer patients in the era of image guided radiation therapy. *J Radiat Oncol* 2016 54 2016;5:359–62. <https://doi.org/10.1007/S13566-016-0274-3>.
- [60] Chen AM, Daly ME, Cui J, Mathai M, Benedict S, Purdy JA. Clinical outcomes among patients with head and neck cancer treated by intensity-modulated radiotherapy with and without adaptive replanning. *Head Neck* 2014;36:1541–6. <https://doi.org/10.1002/HED.23477>.
- [61] Liu Q, Liang J, Zhou D, Krauss DJ, Chen PY, Yan D. Dosimetric Evaluation of Incorporating Patient Geometric Variations Into Adaptive Plan Optimization Through Probabilistic Treatment Planning in Head and Neck Cancers. *Int J Radiat Oncol Biol Phys* 2018. <https://doi.org/10.1016/j.ijrobp.2018.03.062>.
- [62] Bujold A, Craig T, Jaffray D, Dawson LA. Image-guided radiotherapy: has it influenced patient outcomes? *Semin Radiat Oncol* 2012;22:50–61. <https://doi.org/10.1016/J.SEMRADONC.2011.09.001>.
- [63] Loo H, Fairfoul J, Chakrabarti A, Dean JC, Benson RJ, Jefferies SJ, et al. Tumour Shrinkage and Contour Change during Radiotherapy Increase the Dose to Organs at Risk but not the Target Volumes for Head and Neck Cancer Patients Treated on the TomoTherapy HiArt™ System. *Clin Oncol* 2011;23:40–7. <https://doi.org/10.1016/J.CLON.2010.09.003>.
- [64] Castelli J, Simon A, Louvel G, Henry O, Chajon E, Nassef M, et al. Impact of head and neck cancer adaptive radiotherapy to spare the parotid glands and decrease the risk of xerostomia. *Radiat Oncol* 2015;10. <https://doi.org/10.1186/S13014-014-0318-Z>.
- [65] Zhang P, Simon A, Rigaud B, Castelli J, Ospina Arango JD, Nassef M, et al. Optimal adaptive IMRT strategy to spare the parotid glands in oropharyngeal cancer. *Radiother Oncol* 2016;120:41–7. <https://doi.org/10.1016/J.RADONC.2016.05.028>.
- [66] Hansen EK, Bucci MK, Quivey JM, Weinberg V, Xia P. Repeat CT imaging and replanning during the course of IMRT for head-and-neck cancer. *Int J Radiat Oncol Biol Phys* 2006;64:355–62. <https://doi.org/10.1016/J.IJROBP.2005.07.957>.
- [67] Wang J, Bai S, Chen N, Xu F, Jiang X, Li Y, et al. The clinical feasibility and effect of online cone beam computer tomography-guided intensity-modulated radiotherapy for nasopharyngeal cancer. *Radiother Oncol* 2009;90:221–7. <https://doi.org/10.1016/J.RADONC.2008.08.017>.
- [68] Brouwer CL, Steenbakkers RJHM, van der Schaaf A, Sopacua CTC, van Dijk L V., Kierkels RGJ, et al. Selection of head and neck cancer patients for adaptive radiotherapy to decrease xerostomia. *Radiother Oncol* 2016;120:36–40. <https://doi.org/10.1016/J.RADONC.2016.05.025>.

- [69] Vickress JR, Battista J, Barnett R, Yartsev S. Online daily assessment of dose change in head and neck radiotherapy without dose-recalculation. *J Appl Clin Med Phys* 2018;19:659–65. <https://doi.org/10.1002/ACM2.12432>.
- [70] Liu C, Kim J, Kumarasiri A, Mayyas E, Brown SL, Wen N, et al. An automated dose tracking system for adaptive radiation therapy. *Comput Methods Programs Biomed* 2018;154:1–8. <https://doi.org/10.1016/J.CMPB.2017.11.001>.
- [71] Hu CC, Huang WT, Tsai CL, Wu JK, Chao HL, Huang GM, et al. Practically acquired and modified cone-beam computed tomography images for accurate dose calculation in head and neck cancer. *Strahlentherapie Und Onkol* 2011;187:633–44. <https://doi.org/10.1007/S00066-011-2247-1>.
- [72] Heukelom J, Fuller CD. Head and Neck Cancer Adaptive Radiation Therapy (ART): Conceptual Considerations for the Informed Clinician. *Semin Radiat Oncol* 2019;29:258–73. <https://doi.org/10.1016/J.SEMRADONC.2019.02.008>.
- [73] Ahunbay EE, Peng C, Godley A, Schultz C, Li XA. An on-line replanning method for head and neck adaptive radiotherapy. *Med Phys* 2009;36:4776–90. <https://doi.org/10.1118/1.3215532>.
- [74] Ahunbay EE, Peng C, Holmes S, Godley A, Lawton C, Li XA. Online adaptive replanning method for prostate radiotherapy. *Int J Radiat Oncol Biol Phys* 2010;77:1561–72. <https://doi.org/10.1016/J.IJROBP.2009.10.013>.
- [75] Li X, Ahunbay E, Godley A, Morrow N, Wilson JF, White J. An Online Replanning Technique for Breast Adaptive Radiation Therapy. *Int J Radiat Oncol* 2009;75:S71. <https://doi.org/10.1016/J.IJROBP.2009.07.180>.
- [76] Burrige N, Amer A, Marchant T, Sykes J, Stratford J, Henry A, et al. Online adaptive radiotherapy of the bladder: small bowel irradiated-volume reduction. *Int J Radiat Oncol Biol Phys* 2006;66:892–7. <https://doi.org/10.1016/J.IJROBP.2006.07.013>.
- [77] Foroudi F, Wong J, Kron T, Rolfo A, Haworth A, Roxby P, et al. Online adaptive radiotherapy for muscle-invasive bladder cancer: Results of a pilot study. *Int J Radiat Oncol Biol Phys* 2011;81:765–71. <https://doi.org/10.1016/J.IJROBP.2010.06.061>.
- [78] Richter A, Hu Q, Steglich D, Baier K, Wilbert J, Guckenberger M, et al. Investigation of the usability of conebeam CT data sets for dose calculation. *Radiat Oncol* 2008;3. <https://doi.org/10.1186/1748-717X-3-42>.
- [79] Dunlop A, McQuaid D, Nill S, Murray J, Poludniowski G, Hansen VN, et al. Comparison of CT number calibration techniques for CBCT-based dose calculation. *Strahlentherapie Und Onkol* 2015. <https://doi.org/10.1007/s00066-015-0890-7>.
- [80] Perichon ABN, Schick JCU, Chajon OHE, Lafond ASC. A density assignment method for dose monitoring in head-and-neck radiotherapy 2018. <https://doi.org/10.1007/s00066-018-1379-y>.
- [81] van Kranen S, Hamming-Vrieze O, Wolf A, Damen E, van Herk M, Sonke JJ. Head and Neck Margin Reduction With Adaptive Radiation Therapy: Robustness of Treatment Plans Against Anatomy Changes. *Int J Radiat Oncol Biol Phys* 2016. <https://doi.org/10.1016/j.ijrobp.2016.07.011>.
- [82] Liang X, Chen L, Nguyen D, Zhou Z, Gu X, Yang M, et al. Generating synthesized computed tomography (CT) from cone-beam computed tomography (CBCT) using CycleGAN for adaptive radiation therapy. *Phys Med Biol* 2019;64:125002. <https://doi.org/10.1088/1361-6560/ab22f9>.
- [83] Archambault Y, Boylan C, Bullock D, Morgas T, Peltola J, Ruokokoski E, et al. MAKING ON-LINE ADAPTIVE RADIOTHERAPY POSSIBLE USING ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

FOR EFFICIENT DAILY RE-PLANNING. *Med Phys Int J* 2020;8.

- [84] Byrne M, Archibald-Heeren B, Hu Y, Teh A, Beserminji R, Cai E, et al. Varian ethos online adaptive radiotherapy for prostate cancer: Early results of contouring accuracy, treatment plan quality, and treatment time. *J Appl Clin Med Phys* 2022;23. <https://doi.org/10.1002/ACM2.13479>.
- [85] ClinicalTrials.gov. A Prospective Study of Daily Adaptive Radiotherapy to Better Organ-at-Risk Doses in Head and Neck Cancer (DARTBOARD) n.d. <https://clinicaltrials.gov/ct2/show/NCT04883281>.
- [86] Otazo R, Lambin P, Pignol JP, Ladd ME, Schlemmer HP, Baumann M, et al. MRI-guided Radiation Therapy: An Emerging Paradigm in Adaptive Radiation Oncology. *Radiology* 2021;298:248–60. <https://doi.org/10.1148/RADIOL.2020202747/ASSET/IMAGES/LARGE/RADIOL.2020202747.FIG9.JPEG>.
- [87] Das IJ, Yadav P, Mittal BB. Emergence of MR-Linac in Radiation Oncology: Successes and Challenges of Riding on the MRgRT Bandwagon. *J Clin Med* 2022;11. <https://doi.org/10.3390/JCM11175136>.
- [88] Kim J, Garbarino K, Schultz L, Levin K, Movsas B, Siddiqui MS, et al. Dosimetric evaluation of synthetic CT relative to bulk density assignment-based magnetic resonance-only approaches for prostate radiotherapy. *Radiat Oncol* 2015;10:1–9. <https://doi.org/10.1186/S13014-015-0549-7/FIGURES/5>.
- [89] Ma X, Chen X, Wang Y, Qin S, Yan X, Cao Y, et al. Personalized Modeling to Improve Pseudo-Computed Tomography Images for Magnetic Resonance Imaging-Guided Adaptive Radiation Therapy. *Int J Radiat Oncol Biol Phys* 2022;113:885–92. <https://doi.org/10.1016/J.IJROBP.2022.03.032>.
- [90] Nierer L, Eze C, da Silva Mendes V, Braun J, Thum P, von Bestenbostel R, et al. Dosimetric benefit of MR-guided online adaptive radiotherapy in different tumor entities: liver, lung, abdominal lymph nodes, pancreas and prostate. *Radiat Oncol* 2022;17:1–14. <https://doi.org/10.1186/S13014-022-02021-6/FIGURES/4>.
- [91] Güngör G, Serbez İ, Temur B, Gür G, Kayalılar N, Mustafayev TZ, et al. Time Analysis of Online Adaptive Magnetic Resonance–Guided Radiation Therapy Workflow According to Anatomical Sites. *Pract Radiat Oncol* 2021;11:e11–21. <https://doi.org/10.1016/J.PRRO.2020.07.003>.
- [92] ClinicalTrials.gov. Adaptive Radiotherapy for Head and Neck Cancer n.d. <https://clinicaltrials.gov/ct2/show/NCT03096808>.
- [93] ClinicalTrials.gov. MRI - Guided Adaptive RadioTherapy for Reducing XerostomiA in Head and Neck Cancer (MARTHA-trial) (MARTHA) n.d. <https://clinicaltrials.gov/ct2/show/NCT03972072>.
- [94] Clinicaltrials.gov. PEARL PET-based Adaptive Radiotherapy Clinical Trial (PEARL) n.d. <https://clinicaltrials.gov/ct2/show/NCT03935672>.
- [95] ClinicalTrials.gov. Comparison of Adaptive Dose Painting by Numbers With Standard Radiotherapy for Head and Neck Cancer. (C-ART-2) n.d. <https://clinicaltrials.gov/ct2/show/NCT01341535>.
- [96] ClinicalTrials.gov. Adaptive, Image-guided, Intensity-modulated Radiotherapy for Head and Neck Cancer in the Reduced Volumes of Elective Neck n.d. <https://clinicaltrials.gov/ct2/show/NCT01287390>.
- [97] ClinicalTrials.gov. Adaptive Radiation Treatment for Head and Neck Cancer (ARTFORCE) n.d. <https://clinicaltrials.gov/ct2/show/NCT01504815>.
- [98] Heukelom J, Hamming O, Bartelink H, Hoebbers F, Giralt J, Herlestam T, et al. Adaptive and

- innovative Radiation Treatment FOR improving Cancer treatment outcome (ARTFORCE); a randomized controlled phase II trial for individualized treatment of head and neck cancer. *BMC Cancer* 2013;13. <https://doi.org/10.1186/1471-2407-13-84>.
- [99] Bahig H, Yuan Y, Mohamed ASR, Brock KK, Ng SP, Wang J, et al. Magnetic Resonance-based Response Assessment and Dose Adaptation in Human Papilloma Virus Positive Tumors of the Oropharynx treated with Radiotherapy (MR-ADAPTOR): An R-IDEAL stage 2a-2b/Bayesian phase II trial. *Clin Transl Radiat Oncol* 2018;13:19–23. <https://doi.org/10.1016/J.CTRO.2018.08.003>.
- [100] Qin A, Liang J, Han X, O’Connell N, Yan D. The impact of deformable image registration methods on dose warping. *Med Phys* 2018;45:1287–94. <https://doi.org/10.1002/mp.12741>.
- [101] Rigaud B, Simon A, Castelli J, Gobeli M, Ospina Arango JD, Cazoulat G, et al. Evaluation of deformable image registration methods for dose monitoring in head and neck radiotherapy. *Biomed Res Int* 2015;2015. <https://doi.org/10.1155/2015/726268>.
- [102] Glide-Hurst C, Lee P, Yock A, Olsen J, Cao M, Siddiqui F, et al. Adaptive Radiation Therapy (ART) Strategies and Technical Considerations: A State of the ART Review From NRG Oncology. *Int J Radiat Oncol Biol Phys* 2021;109.
- [103] Yan D, Jaffray DA, Wong JW. A model to accumulate fractionated dose in a deforming organ. *Int J Radiat Oncol Biol Phys* 1999;44:665–75. [https://doi.org/10.1016/S0360-3016\(99\)00007-3](https://doi.org/10.1016/S0360-3016(99)00007-3).
- [104] García-Alvarez JA, Zhong H, Schultz CJ, Li XA, Kainz K. Incorporating uncertainty bounds in daily deformable dose accumulation for adaptive radiation therapy of head-and-neck cancer. *Med Phys* 2022. <https://doi.org/10.1002/MP.16085>.
- [105] Lowther NJ, Marsh SH, Louwe RJW. Quantifying the dose accumulation uncertainty after deformable image registration in head-and-neck radiotherapy. *Radiother Oncol* 2020;143:117–25. <https://doi.org/10.1016/J.RADONC.2019.12.009>.
- [106] Brock KK, Ohrt AN, Gryshkevych S, McCulloch MM, Cazoulat G, Mohamed AS, et al. Clinical Implementation of Daily Dose Accumulation and Adaptive Radiotherapy. *Int J Radiat Oncol* 2020;108:e371–2. <https://doi.org/10.1016/j.ijrobp.2020.07.2381>.
- [107] Kong VC, Marshall A, Chan HB. Cone Beam Computed Tomography: The Challenges and Strategies in Its Application for Dose Accumulation. *J Med Imaging Radiat Sci* 2016. <https://doi.org/10.1016/j.jmir.2015.09.012>.
- [108] Brink A, Stewart K, Hargrave C. Evaluation of dose accumulation methods and workflows utilising cone beam computed tomography images. *J Med Radiat Sci* 2022. <https://doi.org/10.1002/JMRS.622>.
- [109] Veiga C, Lourenço AM, Mouinuddin S, Van Herk M, Modat M, Ourselin S, et al. Toward adaptive radiotherapy for head and neck patients: Uncertainties in dose warping due to the choice of deformable registration algorithm. *Med Phys* 2015;42:760–9. <https://doi.org/10.1118/1.4905050>.
- [110] Pukala J, Johnson PB, Shah AP, Langen KM, Bova FJ, Staton RJ, et al. Benchmarking of five commercial deformable image registration algorithms for head and neck patients. *J Appl Clin Med Phys* 2016;17:25–40. <https://doi.org/10.1120/JACMP.V17I3.5735>.
- [111] Hamet P, Tremblay J. Artificial intelligence in medicine. *Metabolism* 2017;69:S36–40. <https://doi.org/10.1016/J.METABOL.2017.01.011>.
- [112] Introduction to deep learning n.d. <https://medium.com/analytics-vidhya/introduction-to-deep-learning-e3a7899a04a3>.

- [113] Bzdok D, Krzywinski M, Altman N. Points of significance: Machine learning: Supervised methods. *Nat Methods* 2018;15:5–6. <https://doi.org/10.1038/NMETH.4551>.
- [114] Lopez C, Tucker S, Salameh T, Tucker C. An unsupervised machine learning method for discovering patient clusters based on genetic signatures 2018;85:30–9. <https://doi.org/10.1016/J.JBI.2018.07.004>.
- [115] Hoi SCH, Jin R, Zhu J, Lyu MR. Semisupervised SVM batch mode active learning with applications to image retrieval. *ACM Trans Inf Syst* 2009;27. <https://doi.org/10.1145/1508850.1508854>.
- [116] Schmidhuber J. Deep Learning in neural networks: An overview. *Neural Networks* 2015;61:85–117. <https://doi.org/10.1016/J.NEUNET.2014.09.003>.
- [117] Lecun Y, Bengio Y, Hinton G. Deep learning. *Nat* 2015 5217553 2015;521:436–44. <https://doi.org/10.1038/nature14539>.
- [118] Meyer P, Noblet V, Mazzara C, Lallement A. Survey on deep learning for radiotherapy. vol. 98. Elsevier Ltd; 2018. <https://doi.org/10.1016/j.compbimed.2018.05.018>.
- [119] Convolutional neural network tutorial n.d. <https://www.simplilearn.com/tutorials/deep-learning-tutorial/convolutional-neural-network>.
- [120] Khan A, Sohail A, Zahoora U, Qureshi AS. A survey of the recent architectures of deep convolutional neural networks. *Artif Intell Rev* 2020 538 2020;53:5455–516. <https://doi.org/10.1007/S10462-020-09825-6>.
- [121] Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. *Adv Neural Inf Process Syst* 2012;25.
- [122] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit* 2015;2016-December:770–8. <https://doi.org/10.48550/arxiv.1512.03385>.
- [123] J van der V, S W, S D, D R, W C, F M, et al. Benefits of deep learning for delineation of organs at risk in head and neck cancer. *Radiother Oncol* 2019;138:68–74. <https://doi.org/10.1016/J.RADONC.2019.05.010>.
- [124] Amjad A, Xu J, Thill D, Lawton C, Hall W, Awan MJ, et al. General and custom deep learning autosegmentation models for organs in head and neck, abdomen, and male pelvis. *Med Phys* 2022;49:1686–700. <https://doi.org/10.1002/mp.15507>.
- [125] Hu K, Lin A, Young A, Kubicek G, Piper JW, Nelson AS, et al. Timesavings for Contour Generation in Head and Neck IMRT: Multi-institutional Experience with an Atlas-based Segmentation Method. *Int J Radiat Oncol Biol Phys* 2008;72:S391. <https://doi.org/10.1016/J.IJROBP.2008.06.1261>.
- [126] HONG T, TOME W, CHAPPELL R, HARARI P. Variations in target delineation for head and neck IMRT: An international multi-institutional study. *Int J Radiat Oncol* 2004. [https://doi.org/10.1016/s0360-3016\(04\)01130-7](https://doi.org/10.1016/s0360-3016(04)01130-7).
- [127] Nelms BE, Tomé WA, Robinson G, Wheeler J. Variations in the contouring of organs at risk: Test case from a patient with oropharyngeal cancer. *Int J Radiat Oncol Biol Phys* 2012;82:368–78. <https://doi.org/10.1016/j.ijrobp.2010.10.019>.
- [128] Brouwer CL, Steenbakkens RJHM, van den Heuvel E, Duppen JC, Navran A, Bijl HP, et al. 3D Variation in delineation of head and neck organs at risk. *Radiat Oncol* 2012;7:1–10. <https://doi.org/10.1186/1748-717X-7-32/FIGURES/4>.

- [129] Mukesh M, Bs MB, Benson R, Jena R, Hoole A, Roques T, et al. Interobserver variation in clinical target volume and organs at risk segmentation in post-parotidectomy radiotherapy: can segmentation protocols help? n.d. <https://doi.org/10.1259/bjr/66693547>.
- [130] Vinod SK, Jameson MG, Min M, Holloway LC. Uncertainties in volume delineation in radiation oncology: A systematic review and recommendations for future studies. *Radiother Oncol* 2016;121:169–79. <https://doi.org/10.1016/j.radonc.2016.09.009>.
- [131] Lim TY, Gillespie E, Murphy J, Moore KL. Clinically Oriented Contour Evaluation Using Dosimetric Indices Generated From Automated Knowledge-Based Planning. *Int J Radiat Oncol* 2019;103:1251–60. <https://doi.org/10.1016/j.IJROBP.2018.11.048>.
- [132] Sharp G, Fritscher KD, Pekar V, Peroni M, Shusharina N, Veeraraghavan H, et al. Vision 20/20: Perspectives on automated image segmentation for radiotherapy. *Med Phys* 2014;41. <https://doi.org/10.1118/1.4871620>.
- [133] Vrtovec T, Močnik D, Strojjan P, Pernuš F, Ibragimov B. Auto-segmentation of organs at risk for head and neck radiotherapy planning: From atlas-based to deep learning methods. vol. 47. John Wiley and Sons Ltd; 2020. <https://doi.org/10.1002/mp.14320>.
- [134] Kosmin M, Ledsam J, Romera-Paredes B, Mendes R, Moinuddin S, de Souza D, et al. Rapid advances in auto-segmentation of organs at risk and target volumes in head and neck cancer. *Radiother Oncol* 2019;135:130–40. <https://doi.org/10.1016/j.radonc.2019.03.004>.
- [135] Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. *BMC Med Imaging* 2015. <https://doi.org/10.1186/s12880-015-0068-x>.
- [136] Maier-Hein L, Reinke A, Christodoulou E, Glocker B, Godau P, Isensee F, et al. Metrics reloaded: Pitfalls and recommendations for image analysis validation 2022. <https://doi.org/10.48550/arxiv.2206.01653>.
- [137] Voet PWJ, Dirkx MLP, Teguh DN, Hoogeman MS, Levendag PC, Heijmen BJM. Does atlas-based autosegmentation of neck levels require subsequent manual contour editing to avoid risk of severe target underdosage? A dosimetric analysis. *Radiother Oncol* 2011;98:373–7. <https://doi.org/10.1016/j.radonc.2010.11.017>.
- [138] Jameson MG, Holloway LC, Vial PJ, Vinod SK, Metcalfe PE. A review of methods of analysis in contouring studies for radiation oncology. *J Med Imaging Radiat Oncol* 2010;54:401–10. <https://doi.org/10.1111/j.1754-9485.2010.02192.x>.
- [139] Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The cancer imaging archive (TCIA): Maintaining and operating a public information repository. *J Digit Imaging* 2013. <https://doi.org/10.1007/s10278-013-9622-7>.
- [140] Raudaschl PF, Zaffino P, Sharp GC, Spadea MF, Chen A, Dawant BM, et al. Evaluation of segmentation methods on head and neck CT: Auto-segmentation challenge 2015. *Med Phys* 2017. <https://doi.org/10.1002/mp.12197>.
- [141] Guo H, Wang J, Xia X, Zhong Y, Peng J, Zhang Z, et al. The dosimetric impact of deep learning-based auto-segmentation of organs at risk on nasopharyngeal and rectal cancer. *Radiat Oncol* 2021;16:1–14. <https://doi.org/10.1186/S13014-021-01837-Y/TABLES/4>.
- [142] Tao CJ, Yi JL, Chen NY, Ren W, Cheng J, Tung S, et al. Multi-subject atlas-based auto-segmentation reduces interobserver variation and improves dosimetric parameter consistency for organs at risk in nasopharyngeal carcinoma: A multi-institution clinical study. *Radiother Oncol* 2015;115:407–11.

<https://doi.org/10.1016/j.radonc.2015.05.012>.

- [143] van Dijk L V., Van den Bosch L, Aljabar P, Peressutti D, Both S, Steenbakkers Roel JHM, et al. Improving automatic delineation for head and neck organs at risk by Deep Learning Contouring. *Radiother Oncol* 2020;142:115–23. <https://doi.org/10.1016/j.radonc.2019.09.022>.
- [144] van Rooij W, Dahele M, Ribeiro Brandao H, Delaney AR, Slotman BJ, Verbakel WF. Deep Learning-Based Delineation of Head and Neck Organs at Risk: Geometric and Dosimetric Evaluation. *Int J Radiat Oncol Biol Phys* 2019;104:677–84. <https://doi.org/10.1016/J.IJROBP.2019.02.040>.
- [145] Tsuji SY, Hwang A, Weinberg V, Yom SS, Quivey JM, Xia P. Dosimetric Evaluation of Automatic Segmentation for Adaptive IMRT for Head-and-Neck Cancer. vol. 77. 2010. <https://doi.org/10.1016/j.ijrobp.2009.06.012>.
- [146] Ma Z, Tavares JMRS, Jorge RMN. A review on the current segmentation algorithms for medical images. *IMAGAPP 2009 - Proc 1st Int Conf Comput Imaging Theory Appl* 2009:135–40. <https://doi.org/10.5220/0001793501350140>.
- [147] Ferreira A, Gentil F, Tavares JMRS. Segmentation algorithms for ear image data towards biomechanical studies. *Comput Methods Biomech Biomed Engin* 2014;17:888–904. <https://doi.org/10.1080/10255842.2012.723700>.
- [148] Ma Z, Tavares JMRS, Jorge RN, Mascarenhas T. A review of algorithms for medical image segmentation and their applications to the female pelvic cavity. <https://doi.org/10.1080/10255840903131878> 2009;13:235–46. <https://doi.org/10.1080/10255840903131878>.
- [149] Xu A, Wang L, Feng S, Qu Y. Threshold-based level set method of image segmentation. *Proc - 3rd Int Conf Intell Networks Intell Syst ICINIS 2010* 2010:703–6. <https://doi.org/10.1109/ICINIS.2010.181>.
- [150] Çiğla C, Alatan AA. Region-based image segmentation via graph cuts. *2008 IEEE 16th Signal Process Commun Appl Conf SIU 2008*:1–4. <https://doi.org/10.1109/SIU.2008.4632542>.
- [151] Z Y-Q, G W-H, C Z-C, T J-T, L L-Y. Medical images edge detection based on mathematical morphology. *Conf Proc . Annu Int Conf IEEE Eng Med Biol Soc IEEE Eng Med Biol Soc Annu Conf 2005*;2005:6492–5. <https://doi.org/10.1109/IEMBS.2005.1615986>.
- [152] X H, MS H, PC L, LS H, DN T, P V, et al. Atlas-based auto-segmentation of head and neck CT images. *Med Image Comput Comput Assist Interv* 2008;11:434–41. https://doi.org/10.1007/978-3-540-85990-1_52.
- [153] Larrue A, Gujral D, Nutting C, Gooding M. The impact of the number of atlases on the performance of automatic multi-atlas contouring. *Phys Medica* 2015. <https://doi.org/10.1016/j.ejmp.2015.10.020>.
- [154] La Macchia M, Fellin F, Amichetti M, Cianchetti M, Gianolini S, Paola V, et al. Systematic evaluation of three different commercial software solutions for automatic segmentation for adaptive therapy in head-and-neck, prostate and pleural cancer. *Radiat Oncol* 2012. <https://doi.org/10.1186/1748-717X-7-160>.
- [155] Gresswell S, Renz P, Werts D, Arshoun Y. (P059) Impact of Increasing Atlas Size on Accuracy of an Atlas-Based Auto-Segmentation Program (ABAS) for Organs-at-Risk (OARS) in Head and Neck (H&N) Cancer Patients. *Int J Radiat Oncol* 2017. <https://doi.org/10.1016/j.ijrobp.2017.02.155>.
- [156] Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (STAPLE):

An algorithm for the validation of image segmentation. *IEEE Trans Med Imaging* 2004;23:903–21. <https://doi.org/10.1109/TMI.2004.828354>.

- [157] Teguh DN, Levendag PC, Voet PWJ, Al-Mamgani A, Han X, Wolf TK, et al. Clinical validation of atlas-based auto-segmentation of multiple target volumes and normal tissue (swallowing/mastication) structures in the head and neck. *Int J Radiat Oncol Biol Phys* 2011. <https://doi.org/10.1016/j.ijrobp.2010.07.009>.
- [158] Yang J, Beadle BM, Garden AS, Gunn B, Rosenthal D, Ang K, et al. Auto-segmentation of low-risk clinical target volume for head and neck radiation therapy. *Pract Radiat Oncol* 2014;4. <https://doi.org/10.1016/j.prro.2013.03.003>.
- [159] Hoogeman MS, Han X, Teguh D, Voet P, Nowak P, Wolf T, et al. Atlas-based Auto-segmentation of CT Images in Head and Neck Cancer: What is the Best Approach? *Int J Radiat Oncol* 2008. <https://doi.org/10.1016/j.ijrobp.2008.06.196>.
- [160] Stapleford LJ, Lawson JD, Perkins C, Edelman S, Davis L, McDonald MW, et al. Evaluation of Automatic Atlas-Based Lymph Node Segmentation for Head-and-Neck Cancer. *Int J Radiat Oncol Biol Phys* 2010;77:959–66. <https://doi.org/10.1016/j.ijrobp.2009.09.023>.
- [161] Hoang Duc AK, Eminowicz G, Mendes R, Wong S-L, McClelland J, Modat M, et al. Validation of clinical acceptability of an atlas-based segmentation algorithm for the delineation of organs at risk in head and neck cancer. *Med Phys* 2015;42:5027–34. <https://doi.org/10.1118/1.4927567>.
- [162] Lee H, Lee E, Kim N, Kim J ho, Park K, Lee H, et al. Clinical evaluation of commercial atlas-based auto-segmentation in the head and neck region. *Front Oncol* 2019;9:1–9. <https://doi.org/10.3389/fonc.2019.00239>.
- [163] Urago Y, Okamoto H, Kaneda T, Murakami N, Kashihara T, Takemori M, et al. Evaluation of auto-segmentation accuracy of cloud-based artificial intelligence and atlas-based models. *Radiat Oncol* 2021;16:175. <https://doi.org/10.1186/s13014-021-01896-1>.
- [164] Chen A, Niermann KJ, Deeley MA, Dawant BM. Evaluation of multiple-atlas-based strategies for segmentation of the thyroid gland in head and neck CT images for IMRT. *Phys Med Biol* 2012;57:93–111. <https://doi.org/10.1088/0031-9155/57/1/93>.
- [165] Coupé P, Manjón J V., Fonov V, Pruessner J, Robles M, Collins DL. Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation. *Neuroimage* 2011. <https://doi.org/10.1016/j.neuroimage.2010.09.018>.
- [166] Van de Velde J, Wouters J, Vercauteren T, De Gerssem W, Achten E, De Neve W, et al. Optimal number of atlases and label fusion for automatic multi-atlas-based brachial plexus contouring in radiotherapy treatment planning. *Radiat Oncol* 2016;11:1–9. <https://doi.org/10.1186/S13014-015-0579-1/FIGURES/5>.
- [167] Han X. WE-E-213CD-06: A Locally Adaptive, Intensity-Based Label Fusion Method for Multi- Atlas Auto-Segmentation. *Med. Phys.*, vol. 39, 2012, p. 3960. <https://doi.org/10.1118/1.4736162>.
- [168] Liu Q, Qin A, Liang J, Yan D. Evaluation of Atlas-Based Auto-Segmentation and Deformable Propagation of Organs-at-Risk for Head-and-Neck Adaptive Radiotherapy. *Recent Patents Top Imaging* 2016;5:79–87. <https://doi.org/10.2174/2451827105999160415123925>.
- [169] Qazi AA, Pekar V, Kim J, Xie J, Breen SL, Jaffray DA. Auto-segmentation of normal and target structures in head and neck CT images: A feature-driven model-based approach. *Med Phys* 2011. <https://doi.org/10.1118/1.3654160>.

- [170] Walker G V., Awan M, Tao R, Koay EJ, Boehling NS, Grant JD, et al. Prospective randomized double-blind study of atlas-based organ-at-risk autosegmentation-assisted radiation planning in head and neck cancer. *Radiother Oncol* 2014. <https://doi.org/10.1016/j.radonc.2014.08.028>.
- [171] Thomson D, Boylan C, Liptrot T, Aitkenhead A, Lee L, Yap B, et al. Evaluation of an automatic segmentation algorithm for definition of head and neck organs at risk. *Radiat Oncol* 2014. <https://doi.org/10.1186/1748-717X-9-173>.
- [172] Fritscher KD, Peroni M, Zaffino P, Spadea MF, Schubert R, Sharp G. Automatic segmentation of head and neck CT images for radiotherapy treatment planning using multiple atlases, statistical appearance models, and geodesic active contours. *Med Phys* 2014;41. <https://doi.org/10.1118/1.4871623>.
- [173] Fortunati V, Verhaart RF, Van Der Lijn F, Niessen WJ, Veenland JF, Paulides MM, et al. Tissue segmentation of head and neck CT images for treatment planning: A multiatlas approach combined with intensity modeling. *Med Phys* 2013. <https://doi.org/10.1118/1.4810971>.
- [174] Han X. Learning-Boosted Label Fusion for Multi-atlas Auto-Segmentation. *Lect Notes Comput Sci (Including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)* 2013;8184 LNCS:17–24. https://doi.org/10.1007/978-3-319-02267-3_3.
- [175] Minaee S, Boykov Y, Porikli F, Plaza A, Kehtarnavaz N, Terzopoulos D. *Image Segmentation Using Deep Learning: A Survey* 2020.
- [176] Fu Y, Lei Y, Wang T, Curran WJ, Liu T, Yang X. A review of deep learning based methods for medical image multi-organ segmentation. *Phys Medica* 2021;85:107–22. <https://doi.org/10.1016/j.ejmp.2021.05.003>.
- [177] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *3rd Int Conf Learn Represent ICLR 2015 - Conf Track Proc* 2014. <https://doi.org/10.48550/arxiv.1409.1556>.
- [178] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit* 2015;07-12-June-2015:1–9. <https://doi.org/10.1109/CVPR.2015.7298594>.
- [179] Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications 2017. <https://doi.org/10.48550/arxiv.1704.04861>.
- [180] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. *Proc - 30th IEEE Conf Comput Vis Pattern Recognition, CVPR 2017* 2017;2017-January:2261–9. <https://doi.org/10.1109/CVPR.2017.243>.
- [181] Shelhamer E, Long J, Darrell T. Fully Convolutional Networks for Semantic Segmentation. n.d.
- [182] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation n.d.
- [183] Siddique N, Paheding S, Elkin CP, Devabhaktuni V. U-net and its variants for medical image segmentation: A review of theory and applications. *IEEE Access* 2021. <https://doi.org/10.1109/ACCESS.2021.3086020>.
- [184] Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. *Lect Notes Comput Sci (Including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)* 2016;9901 LNCS:424–32.

<https://doi.org/10.48550/arxiv.1606.06650>.

- [185] Milletari F, Navab N, Ahmadi SA. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. *Proc - 2016 4th Int Conf 3D Vision, 3DV 2016* 2016;565–71. <https://doi.org/10.1109/3DV.2016.79>.
- [186] Zhu W, Huang Y, Zeng L, Chen X, Liu Y, Qian Z, et al. AnatomyNet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy. *Med Phys* 2019;46:576–89. <https://doi.org/10.1002/MP.13300>.
- [187] Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods* 2020 18:203–11. <https://doi.org/10.1038/s41592-020-01008-z>.
- [188] Yi X, Walia E, Babyn P. Generative adversarial network in medical imaging: A review. *Med Image Anal* 2019;58:101552. <https://doi.org/10.1016/J.MEDIA.2019.101552>.
- [189] Dong X, Lei Y, Wang T, Thomas M, Tang L, Curran WJ, et al. Automatic multiorgan segmentation in thorax CT images using U-net-GAN. *Med Phys* 2019;46:2157–68. <https://doi.org/10.1002/MP.13458>.
- [190] Yang J, Veeraraghavan H, Armato SG, Farahani K, Kirby JS, Kalpathy-Kramer J, et al. Auto-segmentation for thoracic radiation treatment planning: A grand challenge at AAPM 2017. *Med Phys* 2018;45:4568–81. <https://doi.org/10.1002/mp.13141>.
- [191] Chen X, Sun S, Bai N, Han K, Liu Q, Yao S, et al. A deep learning-based auto-segmentation system for organs-at-risk on whole-body computed tomography images for radiation therapy. *Radiother Oncol* 2021;160:175–84. <https://doi.org/10.1016/j.radonc.2021.04.019>.
- [192] Tang H, Chen X, Liu Y, Lu Z, You J, Yang M, et al. Clinically applicable deep learning framework for organs at risk delineation in CT images. *Nat Mach Intell* 2019 110 2019;1:480–91. <https://doi.org/10.1038/s42256-019-0099-z>.
- [193] Ibragimov B, Xing L. Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks: *Med Phys* 2017. <https://doi.org/10.1002/mp.12045>.
- [194] Willems S, Crijs W, La Greca Saint-Estevan A, Van Der Veen J, Robben D, Depuydt T, et al. Clinical implementation of deepvoxnet for auto-delineation of organs at risk in head and neck cancer patients in radiotherapy. *Lect Notes Comput Sci (Including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)* 2018;11041 LNCS:223–32. https://doi.org/10.1007/978-3-030-01201-4_24.
- [195] Nikolov S, Blackwell S, Zverovitch A, Mendes R, Livne M, de Fauw J, et al. Clinically applicable segmentation of head and neck anatomy for radiotherapy: Deep learning algorithm development and validation study. *J Med Internet Res* 2021;23.
- [196] Feng X, Qing K, Tustison NJ, Meyer CH, Chen Q. Deep convolutional neural network for segmentation of thoracic organs-at-risk using cropped 3D images 2019. <https://doi.org/10.1002/mp.13466>.
- [197] Zhou X, Takayama R, Wang S, Hara T, Fujita H. Deep learning of the sectional appearances of 3D CT images for anatomical structure segmentation based on an FCN voting method. *Med Phys* 2017;44:5221–33. <https://doi.org/10.1002/MP.12480>.
- [198] Gibson E, Giganti F, Hu Y, Bonmati E, Bandula S, Gurusamy K, et al. Automatic Multi-organ Segmentation on Abdominal CT with Dense V-networks. *IEEE Trans Med Imaging* 2018;37:1822. <https://doi.org/10.1109/TMI.2018.2806309>.

- [199] Kim H, Jung J, Kim J, Cho B, Kwak J, Jang JY, et al. Abdominal multi-organ auto-segmentation using 3D-patch-based deep convolutional neural network. *Sci Reports* 2020 101 2020;10:1–9. <https://doi.org/10.1038/s41598-020-63285-0>.
- [200] Men K, Dai J, Li Y. Automatic segmentation of the clinical target volume and organs at risk in the planning CT for rectal cancer using deep dilated convolutional neural networks. *Med Phys* 2017;44:6377–89. <https://doi.org/10.1002/MP.12602>.
- [201] Liu Z, Liu X, Xiao B, Wang S, Miao Z, Sun Y, et al. Segmentation of organs-at-risk in cervical cancer CT images with a convolutional neural network. *Phys Medica* 2020;69:184–91. <https://doi.org/10.1016/J.EJMP.2019.12.008>.
- [202] Wang Y, Zhao L, Wang M, Song Z. Organ at Risk Segmentation in Head and Neck CT Images Using a Two-Stage Segmentation Framework Based on 3D U-Net. *IEEE Access* 2019;7:144591–602. <https://doi.org/10.1109/ACCESS.2019.2944958>.
- [203] Liu Y, Lei Y, Fu Y, Wang T, Zhou J, Jiang X, et al. Head and neck multi-organ auto-segmentation on CT images aided by synthetic MRI. *Med Phys* 2020;47:4294–302. <https://doi.org/10.1002/MP.14378>.
- [204] X R, L X, D N, Y S, H Z, D S, et al. Interleaved 3D-CNNs for joint segmentation of small-volume structures in head and neck CT images. *Med Phys* 2018;45:2063–75. <https://doi.org/10.1002/MP.12837>.
- [205] Zabel WJ, Conway JL, Gladwish A, Skliarenko J, Didiodato G, Goorts-Matthews L, et al. Clinical Evaluation of Deep Learning and Atlas-Based Auto-Contouring of Bladder and Rectum for Prostate Radiation Therapy. *Pract Radiat Oncol* 2021;11:e80–9. <https://doi.org/10.1016/j.ppro.2020.05.013>.
- [206] Robert C, Munoz A, Moreau D, Mazurier J, Sidorski G, Gasnier A, et al. Clinical implementation of deep-learning based auto-contouring tools—Experience of three French radiotherapy centers. *Cancer/Radiotherapie* 2021;25:607–16. <https://doi.org/10.1016/j.canrad.2021.06.023>.
- [207] Vandewinckele L, Claessens M, Dinkla A, Brouwer C, Crijs W, Verellen D, et al. Overview of artificial intelligence-based applications in radiotherapy: Recommendations for implementation and quality assurance. *Radiother Oncol* 2020;153:55–66. <https://doi.org/10.1016/j.radonc.2020.09.008>.
- [208] Johnson JM, Khoshgofaar TM. Survey on deep learning with class imbalance n.d. <https://doi.org/10.1186/s40537-019-0192-5>.
- [209] Tajbakhsh N, Jeyaseelan L, Li Q, Chiang JN, Wu Z, Ding X. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Med Image Anal* 2020;63:101693. <https://doi.org/10.1016/J.MEDIA.2020.101693>.
- [210] Chaney EL, Pizer S, Joshi S, Broadhurst R, Fletcher T, Gash G, et al. Automatic male pelvis segmentation from CT images via statistically trained multi-object deformable m-rep models. *Int J Radiat Oncol Biol Phys* 2004;60:S153–4. <https://doi.org/10.1016/J.IJROBP.2004.06.067>.
- [211] AL G, R L, N W, S S, R T, P K, et al. Validation of a method for automatic image fusion (BrainLAB System) of CT data and 11C-methionine-PET data for stereotactic radiotherapy using a LINAC: first clinical experience. *Int J Radiat Oncol Biol Phys* 2003;56:1450–63. [https://doi.org/10.1016/S0360-3016\(03\)00279-7](https://doi.org/10.1016/S0360-3016(03)00279-7).
- [212] Commowick O, Grégoire V, Malandain G. Atlas-based delineation of lymph node levels in head and neck computed tomography images. *Radiother Oncol* 2008;87:281–9. <https://doi.org/10.1016/j.radonc.2008.01.018>.

- [213] McDonald BA, Cardenas C, O'Connell N, Ahmed S, Naser MA, Wahid KA, et al. Investigation of Autosegmentation Techniques on T2-Weighted MRI for Off-line Dose Reconstruction in MR-Linac Adapt to Position Workflow for Head and Neck Cancers. *MedRxiv* 2021:2021.09.30.21264327. <https://doi.org/10.1101/2021.09.30.21264327>.
- [214] Gooding MJ, Chu K, Conibear J, Dilling T, Durrant L, Fuss M, et al. Multicenter Clinical Assessment of DIR Atlas-Based Autocontouring. *Int J Radiat Oncol Biol Phys* 2013;87:S714–5. <https://doi.org/10.1016/j.IJROBP.2013.06.1892>.
- [215] JS, K L, V K, J X, KK B, J M, et al. Automated weekly replanning for intensity-modulated radiotherapy of cervix cancer. *Int J Radiat Oncol Biol Phys* 2010;78:350–8. <https://doi.org/10.1016/j.IJROBP.2009.07.1699>.
- [216] B H, T C, M S, P K, M H, U O, et al. Automatic segmentation of thoracic and pelvic CT images for radiotherapy planning using implicit anatomic knowledge and organ-specific segmentation strategies. *Phys Med Biol* 2008;53:1751–71. <https://doi.org/10.1088/0031-9155/53/6/017>.
- [217] Ung M, Rouyar-Nicolas A, Limkin E, Petit C, Sarrade T, Carre A, et al. Improving Radiotherapy Workflow Through Implementation of Delineation Guidelines & AI-Based Annotation. *Int J Radiat Oncol* 2020;108:e315. <https://doi.org/10.1016/j.IJROBP.2020.07.753>.
- [218] Wong J, Fong A, McVicar N, Smith S, Giambattista JJ, Wells D, et al. Comparing deep learning-based auto-segmentation of organs at risk and clinical target volumes to expert inter-observer variability in radiotherapy planning. *Radiother Oncol* 2020;144:152–8. <https://doi.org/10.1016/j.radonc.2019.10.019>.
- [219] Hussein M, Heijmen BJM, Verellen D, Nisbet A. Automation in intensity modulated radiotherapy treatment planning-a review of recent innovations. *Br J Radiol* 2018;91. <https://doi.org/10.1259/BJR.20180270/ASSET/IMAGES/LARGE/BJR.20180270.G004.JPEG>.
- [220] Meyer P, Biston M-C, Khamphan C, Marghani T, Mazurier J, Bodez V, et al. Automation in radiotherapy treatment planning: Examples of use in clinical practice and future trends for a complete automated workflow Automatisation de la planification du traitement par irradiation : exemples d'implémentations cliniques et futures tend. *Cancer/Radiothérapie* 2021;25:617–22. <https://doi.org/10.1016/j.canrad.2021.06.006>.
- [221] McIntosh C, Purdie TG. Contextual Atlas Regression Forests: Multiple-Atlas-Based Automated Dose Prediction in Radiation Therapy. *IEEE Trans Med Imaging* 2016;35:1000–12. <https://doi.org/10.1109/TMI.2015.2505188>.
- [222] Petrovic S, Khussainova G, Jagannathan R. Knowledge-light adaptation approaches in case-based reasoning for radiotherapy treatment planning. *Artif Intell Med* 2016;68:17–28. <https://doi.org/10.1016/j.ARTMED.2016.01.006>.
- [223] Sheng Y, Li T, Zhang Y, Lee WR, Yin FF, Ge Y, et al. Atlas-guided prostate intensity modulated radiation therapy (IMRT) planning. *Phys Med Biol* 2015;60:7277–91. <https://doi.org/10.1088/0031-9155/60/18/7277>.
- [224] Chanyavanich V, Das SK, Lee WR, Lo JY. Knowledge-based IMRT treatment planning for prostate cancer. *Med Phys* 2011;38:2515–22. <https://doi.org/10.1118/1.3574874>.
- [225] Skarpmann Munter J, Sjölund J. Dose-volume histogram prediction using density estimation. *Phys Med Biol* 2015;60:6923. <https://doi.org/10.1088/0031-9155/60/17/6923>.
- [226] Yang Y, Xing L. Clinical knowledge-based inverse treatment planning. *Phys Med Biol* 2004;49:5101.

<https://doi.org/10.1088/0031-9155/49/22/006>.

- [227] Tol JP, Delaney AR, Dahele M, Slotman BJ, Verbakel WFAR. Evaluation of a knowledge-based planning solution for head and neck cancer. *Int J Radiat Oncol Biol Phys* 2015;91:612–20. <https://doi.org/10.1016/J.IJROBP.2014.11.014>.
- [228] Yang Y, Li T, Yuan L, Ge Y, Yin FF, Lee WR, et al. Quantitative comparison of automatic and manual IMRT optimization for prostate cancer: The benefits of DVH prediction. *J Appl Clin Med Phys* 2015;16:241–50. <https://doi.org/10.1120/JACMP.V16I2.5204>.
- [229] Schubert C, Waletzko O, Weiss C, Voelzke D, Toperim S, Roeser A, et al. Intercenter validation of a knowledge based model for automated planning of volumetric modulated arc therapy for prostate cancer. The experience of the German RapidPlan Consortium. *PLoS One* 2017;12. <https://doi.org/10.1371/JOURNAL.PONE.0178034>.
- [230] Hussein M, South CP, Barry MA, Adams EJ, Jordan TJ, Stewart AJ, et al. Clinical validation and benchmarking of knowledge-based IMRT and VMAT treatment planning in pelvic anatomy. *Radiother Oncol* 2016;120:473–9. <https://doi.org/10.1016/J.RADONC.2016.06.022>.
- [231] Fogliata A, Belosi F, Clivio A, Navarria P, Nicolini G, Scorsetti M, et al. On the pre-clinical validation of a commercial model-based optimisation engine: Application to volumetric modulated arc therapy for patients with lung or prostate cancer. *Radiother Oncol* 2014;113:385–91. <https://doi.org/10.1016/J.RADONC.2014.11.009>.
- [232] Powis R, Bird A, Brennan M, Hinks S, Newman H, Reed K, et al. Clinical implementation of a knowledge based planning tool for prostate VMAT. *Radiat Oncol* 2017;12:1–8. <https://doi.org/10.1186/S13014-017-0814-Z/TABLES/1>.
- [233] Wang J, Hu W, Yang Z, Chen X, Wu Z, Yu X, et al. Is it possible for knowledge-based planning to improve intensity modulated radiation therapy plan quality for planners with different planning experiences in left-sided breast cancer patients? *Radiat Oncol* 2017;12. <https://doi.org/10.1186/S13014-017-0822-Z>.
- [234] Foy JJ, Marsh R, Ten Haken RK, Younge KC, Schipper M, Sun Y, et al. An analysis of knowledge-based planning for stereotactic body radiation therapy of the spine. *Pract Radiat Oncol* 2017;7:e355–60. <https://doi.org/10.1016/J.PPRO.2017.02.007>.
- [235] Delaney AR, Dahele M, Tol JP, Slotman BJ, Verbakel WFAR. Knowledge-based planning for stereotactic radiotherapy of peripheral early-stage lung cancer. *Acta Oncol (Madr)* 2017;56:490–5. <https://doi.org/10.1080/0284186X.2016.1273544>.
- [236] Snyder KC, Kim J, Reding A, Fraser C, Gordon J, Ajlouni M, et al. Development and evaluation of a clinical model for lung cancer patients using stereotactic body radiotherapy (SBRT) within a knowledge-based algorithm for treatment planning. *J Appl Clin Med Phys* 2016;17:263–75. <https://doi.org/10.1120/JACMP.V17I6.6429>.
- [237] Ahmed S, Nelms B, Gintz D, Caudell J, Zhang G, Moros EG, et al. A method for a priori estimation of best feasible DVH for organs-at-risk: Validation for head and neck VMAT planning. *Med Phys* 2017;44:5486–97. <https://doi.org/10.1002/MP.12500>.
- [238] Habraken SJM, Sharfo AWM, Buijsen J, Verbakel WFAR, Haasbeek CJA, Öllers MC, et al. The TRENDY multi-center randomized trial on hepatocellular carcinoma – Trial QA including automated treatment planning and benchmark-case results. *Radiother Oncol* 2017;125:507–13. <https://doi.org/10.1016/J.RADONC.2017.09.007>.

- [239] Fogliata A, Nicolini G, Clivio A, Vanetti E, Laksar S, Tozzi A, et al. A broad scope knowledge based model for optimization of VMAT in esophageal cancer: Validation and assessment of plan quality among different treatment centers. *Radiat Oncol* 2015;10. <https://doi.org/10.1186/S13014-015-0530-5>.
- [240] Fogliata A, Wang PM, Belosi F, Clivio A, Nicolini G, Vanetti E, et al. Assessment of a model based optimization engine for volumetric modulated arc therapy for patients with advanced hepatocellular cancer. *Radiat Oncol* 2014;9:236. <https://doi.org/10.1186/S13014-014-0236-0>.
- [241] Sheng Y, Ge Y, Yuan L, Li T, Yin FF, Wu QJ. Outlier identification in radiation therapy knowledge-based planning: A study of pelvic cases. *Med Phys* 2017;44:5617–26. <https://doi.org/10.1002/MP.12556>.
- [242] Schreiber E, Fox T. Prior-knowledge treatment planning for volumetric arc therapy using feature-based database mining. *J Appl Clin Med Phys* 2014;15:19–27. <https://doi.org/10.1120/JACMP.V15I2.4596>.
- [243] Mayo CS, Yao J, Eisbruch A, Balter JM, Litzenberg DW, Matuszak MM, et al. Incorporating big data into treatment plan evaluation: Development of statistical DVH metrics and visualization dashboards. *Adv Radiat Oncol* 2017;2:503–14. <https://doi.org/10.1016/J.ADRO.2017.04.005>.
- [244] Zarepisheh M, Long T, Li N, Tian Z, Romeijn HE, Jia X, et al. A DVH-guided IMRT optimization algorithm for automatic treatment planning and adaptive radiotherapy replanning. *Med Phys* 2014;41. <https://doi.org/10.1118/1.4875700>.
- [245] Good D, Lo J, Lee WR, Wu QJ, Yin FF, Das SK. A knowledge-based approach to improving and homogenizing intensity modulated radiation therapy planning quality among treatment centers: An example application to prostate cancer planning. *Int J Radiat Oncol Biol Phys* 2013;87:176–81. <https://doi.org/10.1016/J.IJROBP.2013.03.015>.
- [246] Fogliata A, Reggiori G, Stravato A, Lobefalo F, Franzese C, Franceschini D, et al. RapidPlan head and neck model: The objectives and possible clinical benefit. *Radiat Oncol* 2017;12. <https://doi.org/10.1186/S13014-017-0808-X>.
- [247] Krayenbuehl J, Norton I, Studer G, Guckenberger M. Evaluation of an automated knowledge based treatment planning system for head and neck. *Radiat Oncol* 2015;10. <https://doi.org/10.1186/s13014-015-0533-2>.
- [248] Lu R, Radke RJ, Hong L, Chui CS, Xiong J, Yorke E, et al. Learning the relationship between patient geometry and beam intensity in breast intensity-modulated radiotherapy. *IEEE Trans Biomed Eng* 2006;53:908–20. <https://doi.org/10.1109/TBME.2005.863987>.
- [249] Shiraishi S, Moore KL. Knowledge-based prediction of three-dimensional dose distributions for external beam radiotherapy. *Med Phys* 2016;43:378–87. <https://doi.org/10.1118/1.4938583>.
- [250] Liu J, Wu QJ, Kirkpatrick JP, Yin FF, Yuan L, Ge Y. From active shape model to active optical flow model: A shape-based approach to predicting voxel-level dose distributions in spine SBRT. *Phys Med Biol* 2015;60:N83–92. <https://doi.org/10.1088/0031-9155/60/5/N83>.
- [251] McIntosh C, Purdie TG. Voxel-based dose prediction with multi-patient atlas selection for automated radiotherapy treatment planning. *Phys Med Biol* 2017;62:415–31. <https://doi.org/10.1088/1361-6560/62/2/415>.
- [252] McIntosh C, Welch M, McNiven A, Jaffray DA, Purdie TG. Fully automated treatment planning for head and neck radiotherapy using a voxel-based dose prediction and dose mimicking method. *Phys*

- Med Biol 2017;62:5926–44. <https://doi.org/10.1088/1361-6560/AA71F8>.
- [253] Delaney AR, Tol JP, Dahele M, Cuijpers J, Slotman BJ, Verbakel WFAR. Effect of Dosimetric Outliers on the Performance of a Commercial Knowledge-Based Planning Solution. *Int J Radiat Oncol Biol Phys* 2016;94:469–77. <https://doi.org/10.1016/j.ijrobp.2015.11.011>.
- [254] Boylan C, Rowbottom C. A bias-free, automated planning tool for technique comparison in radiotherapy - application to nasopharyngeal carcinoma treatments. *J Appl Clin Med Phys* 2014;15:213–25. <https://doi.org/10.1120/JACMP.V15I1.4530>.
- [255] Kraysenbuehl J, Di Martino M, Guckenberger M, Andratschke N. Improved plan quality with automated radiotherapy planning for whole brain with hippocampus sparing: A comparison to the RTOG 0933 trial. *Radiat Oncol* 2017;12:1–7. <https://doi.org/10.1186/s13014-017-0896-7/FIGURES/4>.
- [256] Kraysenbuehl J, Zamburlini M, Ghandour S, Pachoud M, Tol J, Guckenberger M. Planning comparison of five automated treatment planning solutions for locally advanced head and neck cancer 2018:1–8.
- [257] Nawa K, Haga A, Nomoto A, Sarmiento RA, Shiraishi K, Yamashita H, et al. Evaluation of a commercial automatic treatment planning system for prostate cancers. *Med Dosim* 2017;42:203–9. <https://doi.org/10.1016/j.meddos.2017.03.004>.
- [258] Li X, Wang L, Wang J, Han X, Xia B, Wu S, et al. Dosimetric benefits of automation in the treatment of lower thoracic esophageal cancer: Is manual planning still an alternative option? *Med Dosim* 2017;42:289–95. <https://doi.org/10.1016/j.meddos.2017.06.004>.
- [259] Hansen CR, Nielsen M, Bertelsen AS, Hazell I, Holtved E, Zukauskaitė R, et al. Automatic treatment planning facilitates fast generation of high-quality treatment plans for esophageal cancer. *Acta Oncol (Madr)* 2017;56:1495–500. https://doi.org/10.1080/0284186X.2017.1349928/SUPPL_FILE/IONC_A_1349928_SM4685.ZIP.
- [260] Wang S, Zheng D, Zhang C, Ma R, Bennion NR, Lei Y, et al. Automatic planning on hippocampal avoidance whole-brain radiotherapy. *Med Dosim* 2017;42:63–8. <https://doi.org/10.1016/j.meddos.2016.12.002>.
- [261] Song Y, Wang Q, Jiang X, Liu S, Zhang Y, Bai S. Fully automatic volumetric modulated arc therapy plan generation for rectal cancer Automatic rectal VMAT planning. *Radiat Oncol* 2016;119:531–6. <https://doi.org/10.1016/j.radonc.2016.04.010>.
- [262] Zhang X, Li X, Quan EM, Pan X, Li Y. A methodology for automatic intensity-modulated radiation treatment planning for lung cancer. *Phys Med Biol* 2011;56:3873–93. <https://doi.org/10.1088/0031-9155/56/13/009>.
- [263] Yan H, Yin FF, Willett C. Evaluation of an artificial intelligence guided inverse planning system: Clinical case study. *Radiat Oncol* 2007;83:76–85. <https://doi.org/10.1016/j.radonc.2007.02.013>.
- [264] Tol JP, Dahele M, Peltola J, Nord J, Slotman BJ, Verbakel WFAR. Automatic interactive optimization for volumetric modulated arc therapy planning 2015:1–12. <https://doi.org/10.1186/s13014-015-0388-6>.
- [265] Hazell I, Bzdusek K, Kumar P, Hansen CR, Bertelsen A, Eriksen JG, et al. Automatic planning of head and neck treatment plans. *J Appl Clin Med Phys* 2016;17:272–82. <https://doi.org/10.1120/JACMP.V17I1.5901>.

- [266] Hansen CR, Bertelsen A, Hazell I, Zukauskaitė R, Gyldenkerne N, Johansen J, et al. Automatic treatment planning improves the clinical quality of head and neck cancer treatment plans. *Clin Transl Radiat Oncol* 2016;1:2–8. <https://doi.org/10.1016/J.CTRO.2016.08.001>.
- [267] Speer S, Klein A, Kober L, Weiss A, Yohannes I, Bert C. Automation of radiation treatment planning : Evaluation of head and neck cancer patient plans created by the Pinnacle3 scripting and Auto-Planning functions. *Strahlenther Onkol* 2017;193:656–65. <https://doi.org/10.1007/S00066-017-1150-9>.
- [268] Kusters JMAM, Bzdusek K, Kumar P, van Kollenburg PGM, Kunze-Busch MC, Wendling M, et al. Automated IMRT planning in Pinnacle: A study in head-and-neck cancer. *Strahlentherapie Und Onkol* 2017;193:1031–8. <https://doi.org/10.1007/s00066-017-1187-9>.
- [269] Gintz D, Latifi K, Caudell J, Nelms B, Zhang G, Moros E, et al. Initial evaluation of automated treatment planning software. *J Appl Clin Med Phys* 2016;17:331–46. <https://doi.org/10.1120/JACMP.V17I3.6167>.
- [270] Teichert K, Süß P, Serna JI, Monz M, Küfer KH, Thieke C. Comparative analysis of Pareto surfaces in multi-criteria IMRT planning. *Phys Med Biol* 2011;56:3669. <https://doi.org/10.1088/0031-9155/56/12/014>.
- [271] Ghandour S, Matzinger O, Pachoud M. Volumetric-modulated arc therapy planning using multicriteria optimization for localized prostate cancer. *J Appl Clin Med Phys* 2015;16:258–69. <https://doi.org/10.1120/JACMP.V16I3.5410>.
- [272] McGarry CK, Bokrantz R, O’Sullivan JM, Hounsell AR. Advantages and limitations of navigation-based multicriteria optimization (MCO) for localized prostate cancer IMRT planning. *Med Dosim* 2014;39:205–11. <https://doi.org/10.1016/J.MEDDOS.2014.02.002>.
- [273] Kamran SC, Mueller BS, Paetzold P, Dunlap J, Niemierko A, Bortfeld T, et al. Multi-criteria optimization achieves superior normal tissue sparing in a planning study of intensity-modulated radiation therapy for RTOG 1308-eligible non-small cell lung cancer patients. *Radiother Oncol* 2016;118:515–20. <https://doi.org/10.1016/J.RADONC.2015.12.028>.
- [274] Rønne HS, Wee L, Pløen J, Appelt AL. Feasibility of preference-driven radiotherapy dose treatment planning to support shared decision making in anal cancer. *Acta Oncol (Madr)* 2017;56:1277–85. https://doi.org/10.1080/0284186X.2017.1315174/SUPPL_FILE/IONC_A_1315174_SM3884.PDF.
- [275] Kyroudi A, Petersson K, Ghandour S, Pachoud M, Matzinger O, Ozsahin M, et al. Discrepancies between selected Pareto optimal plans and final deliverable plans in radiotherapy multi-criteria optimization. *Radiother Oncol* 2016;120:346–8. <https://doi.org/10.1016/J.RADONC.2016.05.018>.
- [276] Monz M, Küfer KH, Bortfeld TR, Thieke C. Pareto navigation—algorithmic foundation of interactive multi-criteria IMRT planning. *Phys Med Biol* 2008;53:985. <https://doi.org/10.1088/0031-9155/53/4/011>.
- [277] Thieke C, Küfer KH, Monz M, Scherrer A, Alonso F, Oelfke U, et al. A new concept for interactive radiotherapy planning with multicriteria optimization: First clinical evaluation. *Radiother Oncol* 2007;85:292–8. <https://doi.org/10.1016/J.RADONC.2007.06.020>.
- [278] Serna JI, Monz M, Küfer KH, Thieke C. Trade-off bounds for the Pareto surface approximation in multi-criteria IMRT planning. *Phys Med Biol* 2009;54:6299. <https://doi.org/10.1088/0031-9155/54/20/018>.
- [279] Craft DL, Hong TS, Shih HA, Bortfeld TR. Improved planning time and plan quality through

- multicriteria optimization for intensity-modulated radiotherapy. *Int J Radiat Oncol Biol Phys* 2012;82:83–90. <https://doi.org/10.1016/j.ijrobp.2010.12.007>.
- [280] Chen H, Craft DL, Gierga DP. Multicriteria optimization informed VMAT planning. *Med Dosim* 2014;39:64–73. <https://doi.org/10.1016/J.MEDDOS.2013.10.001>.
- [281] Kierkels RGJ, Visser R, Bijl HP, Langendijk JA, van 't Veld AA, Steenbakkers RJHM, et al. Multicriteria optimization enables less experienced planners to efficiently produce high quality treatment plans in head and neck cancer radiotherapy. *Radiat Oncol* 2015;10:1–9. <https://doi.org/10.1186/S13014-015-0385-9/FIGURES/2>.
- [282] Müller BS, Shih HA, Efsthathiou JA, Bortfeld T, Craft D. Multicriteria plan optimization in the hands of physicians: A pilot study in prostate cancer and brain tumors. *Radiat Oncol* 2017;12:1–11. <https://doi.org/10.1186/S13014-017-0903-Z/TABLES/2>.
- [283] Wala J, Craft D, Paly J, Zietman A, Efsthathiou J. Maximizing dosimetric benefits of IMRT in the treatment of localized prostate cancer through multicriteria optimization planning. *Med Dosim* 2013;38:298–303. <https://doi.org/10.1016/J.MEDDOS.2013.02.012>.
- [284] Breedveld S, Storchi PRM, Voet PWJ, Heijmen BJM. ICycle: Integrated, multicriterial beam angle, and profile optimization for generation of coplanar and noncoplanar IMRT plans. *Med Phys* 2012. <https://doi.org/10.1118/1.3676689>.
- [285] Della Gala G, Dirx MLP, Hoekstra N, Fransen D, Lanconelli N, van de Pol M, et al. Fully automated VMAT treatment planning for advanced-stage NSCLC patients. *Strahlentherapie Und Onkol* 2017. <https://doi.org/10.1007/s00066-017-1121-1>.
- [286] Sharfo AWM, Breedveld S, Voet PWJ, Heijkoop ST, Mens JWM, Hoogeman MS, et al. Validation of Fully Automated VMAT Plan Generation for Library-Based Plan-of-the-Day Cervical Cancer Radiotherapy. *PLoS One* 2016;11:e0169202. <https://doi.org/10.1371/JOURNAL.PONE.0169202>.
- [287] Voet PWJJ, Dirx MLPP, Breedveld S, Fransen D, Levendag PC, Heijmen BJMM. Toward Fully Automated Multicriterial Plan Generation : A Prospective Clinical Study. *Radiat Oncol Biol* 2013;85:866–72. <https://doi.org/10.1016/j.ijrobp.2012.04.015>.
- [288] Breedveld S, Storchi PRM, Keijzer M, Heemink AW, Heijmen BJM. A novel approach to multi-criteria inverse planning for IMRT. *Phys Med Biol* 2007;52:6339. <https://doi.org/10.1088/0031-9155/52/20/016>.
- [289] Naccarato S, Rigo M, Pellegrini R, Voet P, Akhiat H, Gurrera D, et al. Automated Planning for Prostate Stereotactic Body Radiation Therapy on the 1.5 T MR-Linac. *Adv Radiat Oncol* 2022;7:100865. <https://doi.org/10.1016/j.adro.2021.100865>.
- [290] Jagt TZ, Janssen TM, Betgen A, Wiersema L, Verhage R, Garritsen S, et al. Benchmarking daily adaptation using fully automated radiotherapy treatment plan optimization for rectal cancer. *Phys Imaging Radiat Oncol* 2022;24:7–13. <https://doi.org/10.1016/j.phro.2022.08.006>.
- [291] Buschmann M, Sharfo AWM, Penninkhof J, Seppenwoolde Y, Goldner G, Georg D, et al. Automated volumetric modulated arc therapy planning for whole pelvic prostate radiotherapy. *Strahlentherapie Und Onkol* 2018;194:333–42. <https://doi.org/10.1007/S00066-017-1246-2/FIGURES/2>.
- [292] Voet PWJ, Dirx MLP, Breedveld S, Al-Mamgani A, Incrocci L, Heijmen BJM. Fully automated volumetric modulated arc therapy plan generation for prostate cancer patients. *Int J Radiat Oncol Biol Phys* 2014. <https://doi.org/10.1016/j.ijrobp.2013.12.046>.

- [293] Buergy D, Sharfo AWM, Heijmen BJM, Voet PWJ, Breedveld S, Wenz F, et al. Fully automated treatment planning of spinal metastases - A comparison to manual planning of Volumetric Modulated Arc Therapy for conventionally fractionated irradiation. *Radiat Oncol* 2017;12:1–7. <https://doi.org/10.1186/S13014-017-0767-2/FIGURES/4>.
- [294] Sharfo AWM, Stieler F, Kupfer O, Heijmen BJM, Dirx MLP, Breedveld S, et al. Automated VMAT planning for postoperative adjuvant treatment of advanced gastric cancer. *Radiat Oncol* 2018;13:1–8. <https://doi.org/10.1186/S13014-018-1032-Z/FIGURES/3>.
- [295] Sharfo AWM, Voet PWJ, Breedveld S, Mens JWM, Hoogeman MS, Heijmen BJM. Comparison of VMAT and IMRT strategies for cervical cancer patients using automated planning. *Radiother Oncol* 2015. <https://doi.org/10.1016/j.radonc.2015.02.006>.
- [296] Babier A, Mahmood R, McNiven AL, Diamant A, Chan TCY. Knowledge-based automated planning with three-dimensional generative adversarial networks. *Med Phys* 2018;47:297–306. <https://doi.org/10.48550/arxiv.1812.09309>.
- [297] Miki K, Kusters M, Nakashima T, Saito A, Kawahara D, Nishibuchi I, et al. Evaluation of optimization workflow using custom-made planning through predicted dose distribution for head and neck tumor treatment. *Phys Med* 2020;80:167–74. <https://doi.org/10.1016/J.EJMP.2020.10.028>.
- [298] Li X, Wang C, Sheng Y, Zhang J, Wang W, Yin FF, et al. An artificial intelligence-driven agent for real-time head-and-neck IMRT plan generation using conditional generative adversarial network (cGAN). *Med Phys* 2021;48:2714–23. <https://doi.org/10.1002/MP.14770>.
- [299] Ma L, Chen M, Gu X, Lu W. Deep learning-based inverse mapping for fluence map prediction. *Phys Med Biol* 2020;65. <https://doi.org/10.1088/1361-6560/ABC12C>.
- [300] Nilsson V, Gruselius H, Zhang T, DeKerf G, Claessens M. Probabilistic dose prediction using mixture density networks for automated radiation therapy treatment planning. *Phys Med Biol* 2021;66. <https://doi.org/10.1088/1361-6560/ABDD8A>.
- [301] Bakx N, Bluemink H, Hagelaar E, van der Sangen M, Theuws J, Hurkmans C. Development and evaluation of radiotherapy deep learning dose prediction models for breast cancer. *Phys Imaging Radiat Oncol* 2021;17:65–70. <https://doi.org/10.1016/J.PHRO.2021.01.006>.
- [302] Johnstone E, Wyatt JJ, Henry AM, Short SC, Sebag-Montefiore D, Murray L, et al. Systematic Review of Synthetic Computed Tomography Generation Methodologies for Use in Magnetic Resonance Imaging–Only Radiation Therapy. *Int J Radiat Oncol* 2018;100:199–217. <https://doi.org/10.1016/J.IJROBP.2017.08.043>.
- [303] Zhu L, Xie Y, Wang J, Xing L. Scatter correction for cone-beam CT in radiation therapy. *Med Phys* 2009;36:2258–68. <https://doi.org/10.1118/1.3130047>.
- [304] Xu Y, Bai T, Yan H, Ouyang L, Pompos A, Wang J, et al. A practical cone-beam CT scatter correction method with optimized Monte Carlo simulations for image-guided radiation therapy. *Phys Med Biol* 2015;60:3567–87. <https://doi.org/10.1088/0031-9155/60/9/3567>.
- [305] Wu M, Keil A, Constantin D, Star-Lack J, Zhu L, Fahrig R. Metal artifact correction for x-ray computed tomography using kV and selective MV imaging. *Med Phys* 2014;41. <https://doi.org/10.1118/1.4901551>.
- [306] Zhao W, Fu G-T, Sun C-L, Wang Y-F, Wei C-F, Cao D-Q, et al. Beam hardening correction for a cone-beam CT system and its effect on spatial resolution. *Chinese Phys C* 2011;35:978. <https://doi.org/10.1088/1674-1137/35/10/018>.

- [307] Zhang Y, Yu H. Convolutional Neural Network Based Metal Artifact Reduction in X-Ray Computed Tomography. *IEEE Trans Med Imaging* 2018;37:1370–81. <https://doi.org/10.1109/TMI.2018.2823083>.
- [308] Xu S, Prinsen P, Wiegert J, Manjeshwar R. Deep residual learning in CT physics: Scatter correction for spectral CT. 2017 IEEE Nucl Sci Symp Med Imaging Conf NSS/MIC 2017 - Conf Proc 2018. <https://doi.org/10.1109/NSSMIC.2017.8532979>.
- [309] Barateau A, Céleste M, Lafond C, Henry O, Couespel S, Simon A, et al. Calcul de dose de radiothérapie à partir de tomographies coniques : état de l’art. *Cancer/Radiothérapie* 2018;22:85–100. <https://doi.org/10.1016/J.CANRAD.2017.07.050>.
- [310] Barateau A, De Crevoisier R, Largent A, Mylona E, Perichon N, Castelli J, et al. Comparison of CBCT-based dose calculation methods in head and neck cancer radiotherapy: from Hounsfield unit to density calibration curve to deep learning. *Med Phys* 2020;47:4683–93. <https://doi.org/10.1002/MP.14387>.
- [311] Giacometti V, King RB, Agnew CE, Irvine DM, Jain S, Hounsell AR, et al. An evaluation of techniques for dose calculation on cone beam computed tomography. *Br J Radiol* 2019;92. <https://doi.org/10.1259/BJR.20180383>.
- [312] Macfarlane M, Wong D, Hoover DA, Wong E, Johnson C, Battista JJ, et al. Patient-specific calibration of cone-beam computed tomography data sets for radiotherapy dose calculations and treatment plan assessment. *J Appl Clin Med Phys* 2018;19:249–57. <https://doi.org/10.1002/ACM2.12293>.
- [313] Marchant TE, Joshi KD, Moore CJ. Accuracy of radiotherapy dose calculations based on cone-beam CT: comparison of deformable registration and image correction based methods. *Phys Med Biol* 2018;63. <https://doi.org/10.1088/1361-6560/AAB0F0>.
- [314] Li Y, Zhu J, Liu Z, Teng J, Xie Q, Zhang L, et al. A preliminary study of using a deep convolution neural network to generate synthesized CT images based on CBCT for adaptive radiotherapy of nasopharyngeal carcinoma. *Phys Med Biol* 2019;64:145010. <https://doi.org/10.1088/1361-6560/AB2770>.
- [315] Chen L, Liang X, Shen C, Jiang S, Wang J. Synthetic CT Generation from CBCT images via Deep Learning. *Med Phys* 2020;47:1115. <https://doi.org/10.1002/MP.13978>.
- [316] Xue X, Ding Y, Shi J, Hao X, Li X, Li D, et al. Cone Beam CT (CBCT) Based Synthetic CT Generation Using Deep Learning Methods for Dose Calculation of Nasopharyngeal Carcinoma Radiotherapy n.d. <https://doi.org/10.1177/15330338211062415>.
- [317] Eckl M, Hoppen L, Sarria GR, Boda-Heggemann J, Simeonova-Chergou A, Steil V, et al. Evaluation of a cycle-generative adversarial network-based cone-beam CT to synthetic CT conversion algorithm for adaptive radiation therapy. *Phys Medica* 2020;80:308–16. <https://doi.org/10.1016/J.EJMP.2020.11.007>.
- [318] Awan M, Kalpathy-Cramer J, Gunn GB, Beadle BM, Garden AS, Phan J, et al. Prospective assessment of an atlas-based intervention combined with real-time software feedback in contouring lymph node levels and organs-at-risk in the head and neck: Quantitative assessment of conformance to expert delineation. *Pract Radiat Oncol* 2013;3:186–93. <https://doi.org/10.1016/J.PPRO.2012.11.002>.
- [319] Levendag PC, Hoogeman M, Teguh D, Wolf T, Hibbard L, Wijers O, et al. Atlas Based Auto-segmentation of CT Images: Clinical Evaluation of using Auto-contouring in High-dose, High-precision Radiotherapy of Cancer in the Head and Neck. *Int J Radiat Oncol* 2008;72:S401.

<https://doi.org/10.1016/j.ijrobp.2008.06.1285>.

- [320] Field M, Hardcastle N, Jameson M, Aherne N, Holloway L. Machine learning applications in radiation oncology. *Phys Imaging Radiat Oncol* 2021;19:13–24.
- [321] Nelms BE, Robinson G, Markham J, Velasco K, Boyd S, Narayan S, et al. Variation in external beam treatment plan quality: An inter-institutional study of planners and planning systems. *Pract Radiat Oncol* 2012;2:296–305. <https://doi.org/10.1016/J.PRRO.2011.11.012>.
- [322] Batumalai V, Jameson MG, Forstner DF, Vial P, Holloway LC. How important is dosimetrist experience for intensity modulated radiation therapy? A comparative analysis of a head and neck case. *Pract Radiat Oncol* 2013;3:e99–106. <https://doi.org/10.1016/J.PRRO.2012.06.009>.
- [323] Clements M, Schupp N, Tattersall M, Brown A, Larson R. Monaco treatment planning system tools and optimization processes. *Med Dosim* 2018;43:106–17. <https://doi.org/10.1016/j.meddos.2018.02.005>.
- [324] Chen W, Li Y, Dyer BA, Feng X, Rao S, Benedict SH, et al. Deep learning vs. atlas-based models for fast auto-segmentation of the masticatory muscles on head and neck CT images. *Radiat Oncol* 2020;15. <https://doi.org/10.1186/s13014-020-01617-0>.
- [325] Nemoto T, Futakami N, Yagi M, Kumabe A, Takeda A, Kunieda E, et al. Efficacy evaluation of 2D, 3D U-Net semantic segmentation and atlas-based segmentation of normal lungs excluding the trachea and main bronchi. *J Radiat Res* 2020. <https://doi.org/10.1093/jrr/rrz086>.
- [326] Fang Y, Wang J, Chen S, Guo Y, Zhang Z, Hu W. The Impact Of Training Sample Size On Deep Learning Based Organ Auto Segmentation For Head Neck. *Int J Radiat Oncol* 2020. <https://doi.org/10.1016/j.ijrobp.2020.07.228>.
- [327] Hänsch A, Gass T, Morgas T, Haas B, Meine H, Klein J, et al. PV-0530: Parotid gland segmentation with deep learning using clinical vs. curated training data. *Radiother Oncol* 2018. [https://doi.org/10.1016/s0167-8140\(18\)30840-5](https://doi.org/10.1016/s0167-8140(18)30840-5).
- [328] Brouwer CL, Steenbakkers RJHM, Bourhis J, Budach W, Grau C, Grégoire V, et al. Head and neck guidelines CT-based delineation of organs at risk in the head and neck region: DAHANCA, EORTC, GORTEC, HKNPCSG, NCIC CTG, NCRI, NRG Oncology and TROG consensus guidelines 2015. <https://doi.org/10.1016/j.radonc.2015.07.041>.
- [329] Grégoire V, Eisbruch A, Hamoir M, Levendag P. Proposal for the delineation of the nodal CTV in the node-positive and the post-operative neck. *Radiother Oncol* 2006;79:15–20. <https://doi.org/10.1016/J.RADONC.2006.03.009>.
- [330] van der Veen J, Gulyban A, Nuyts S. Interobserver variability in delineation of target volumes in head and neck cancer. *Radiother Oncol* 2019;137:9–15. <https://doi.org/10.1016/J.RADONC.2019.04.006>.
- [331] Lim JY, Leech M. Use of auto-segmentation in the delineation of target volumes and organs at risk in head and neck. vol. 55. Taylor and Francis Ltd; 2016. <https://doi.org/10.3109/0284186X.2016.1173723>.
- [332] Sjöberg C, Lundmark M, Granberg C, Johansson S, Ahnesjö A, Montelius A. Clinical evaluation of multi-atlas based segmentation of lymph node regions in head and neck and prostate cancer patients. *Radiat Oncol* 2013;8. <https://doi.org/10.1186/1748-717X-8-229>.
- [333] Teguh DN, Levendag PC, Voet PWJ, Al-Mamgani A, Han X, Wolf TK, et al. Clinical validation of atlas-based auto-segmentation of multiple target volumes and normal tissue (swallowing/mastication)

structures in the head and neck. *Int J Radiat Oncol Biol Phys* 2011;81:950–7. <https://doi.org/10.1016/j.ijrobp.2010.07.009>.

- [334] Daisne JF, Blumhofer A. Atlas-based automatic segmentation of head and neck organs at risk and nodal target volumes: A clinical validation. *Radiat Oncol* 2013;8. <https://doi.org/10.1186/1748-717X-8-154>.
- [335] Chen A, Deeley MA, Niermann KJ, Moretti L, Dawant BM. Combining registration and active shape models for the automatic segmentation of the lymph node regions in head and neck CT images. *Med Phys* 2010;37:6338–46. <https://doi.org/10.1118/1.3515459>.
- [336] Gorthi S, Duay V, Houhou N, Bach Cuadra M, Schick U, Becker M, et al. Segmentation of head and neck lymph node regions for radiotherapy planning using active contour-based atlas registration. *IEEE J Sel Top Signal Process* 2009;3:135–47. <https://doi.org/10.1109/JSTSP.2008.2011104>.
- [337] J van der V, S W, H B, F M, S N. Deep learning for elective neck delineation: More consistent and time efficient. *Radiother Oncol* 2020;153:180–8. <https://doi.org/10.1016/J.RADONC.2020.10.007>.
- [338] Strijbis VIJ, Dahele M, Gurney-Champion OJ, Blom GJ, Vergeer MR, Slotman BJ, et al. Deep Learning for Automated Elective Lymph Node Level Segmentation for Head and Neck Cancer Radiotherapy. *Cancers (Basel)* 2022;14:5501. <https://doi.org/10.3390/CANCERS14225501/S1>.
- [339] Yang X, Jani AB, Rossi PJ, Mao H, Curran WJ, Liu T. Patch-Based Label Fusion for Automatic Multi-Atlas-Based Prostate Segmentation in MR Images n.d. <https://doi.org/10.1117/12.2216424>.
- [340] Wong J, Huang V, Wells D, Giambattista J, Giambattista J, Kolbeck C, et al. Implementation of deep learning-based auto-segmentation for radiotherapy planning structures: a workflow study at two cancer centers. *Radiat Oncol* 2021;16:1–10. <https://doi.org/10.1186/S13014-021-01831-4/FIGURES/4>.
- [341] Voet PWJ, Dirks MLP, Teguh DN, Hoogeman MS, Levendag PC, Heijmen BJM. Does atlas-based autosegmentation of neck levels require subsequent manual contour editing to avoid risk of severe target underdosage? A dosimetric analysis. *Radiother Oncol* 2011;98:373–7. <https://doi.org/10.1016/j.radonc.2010.11.017>.
- [342] Giacometti V, Hounsell AH, McGarry CK. A review of dose calculation approaches with cone beam CT in photon and proton therapy. *Phys Med* 2020;76:243–76. <https://doi.org/10.1016/J.EJMP.2020.06.017>.
- [343] Rong Y, Smilowitz J, Tewatia D, Tomé WA, Paliwal B. Dose Calculation on KV Cone Beam CT Images: An Investigation of the Hu-Density Conversion Stability and Dose Accuracy Using the Site-Specific Calibration. *Med Dosim* 2010;35:195–207. <https://doi.org/10.1016/j.meddos.2009.06.001>.
- [344] Barateau A, Perichon N, Castelli J, Schick U, Henry O, Chajon E, et al. A density assignment method for dose monitoring in head-and-neck radiotherapy. *Strahlentherapie Und Onkol* 2019. <https://doi.org/10.1007/s00066-018-1379-y>.
- [345] Vaassen F, Hazelaar C, Vaniqui A, Gooding M, van der Heyden B, Canters R, et al. Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. *Phys Imaging Radiat Oncol* 2020;13:1–6. <https://doi.org/10.1016/j.phro.2019.12.001>.
- [346] Balagopal A, Nguyen D, Morgan H, Weng Y, Dohopolski M, Lin MH, et al. A deep learning-based framework for segmenting invisible clinical target volumes with estimated uncertainties for post-operative prostate cancer radiotherapy. *Med Image Anal* 2021;72:102101. <https://doi.org/10.1016/J.MEDIA.2021.102101>.

