PRE-CLINICAL ASSESSMENT OF HEAD AND NECK ORGANS AT RISK ATLAS-BASED AUTO-SEGMENTATION WITH ADVANCED METHODOLOGY FOR PAROTID GLANDS

by

Eman Hesham Khawandanh

B.Sc., Umm Al-Qura University, 2009

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

The Faculty of Graduate and Postdoctoral Studies

(Physics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

June 2016

ABSTRACT

Modern technology allows radiation therapy dose distributions to conform closely to targets, providing better sparing of adjacent anatomical structures (organs-at –risk, OARs) in the treatment of cancer. This has the potential to reduce radiation side effects. To take advantage of this technology, accurate delineation of both targets and organs at risk is essential. Routinely, organs at risk are delineated (segmented) on anatomical images using a manual process which can be both time consuming and error prone. Automated segmentation methodology is therefore an active area of research.

The objectives of this thesis were:

1) to assess the variability in manual segmentation of head and neck OARs and provide benchmark data for comparisons with auto-segmentation;

2) to compile and categorize a set of image data sets with expertly validated segmentations of head and neck OARs, forming an in-house constructed atlas library for use with an automated segmentation software tool;

3) to evaluate the performance of an atlas based auto-segmentation tool (MIM Maestro™) using the in-house constructed atlas library; and

4) to improve the auto-segmentation performance by studying the impact of the number and quality of atlas library cases and different user-defined settings.

Results of these studies indicate that the time required to segment a complete OAR set can be reduced to three minutes using the atlas-based auto-segmentation approach, versus 30 minutes for manual segmentation. With the exception of salivary

glands, the auto-segmentation performance was clinically acceptable for all organs. Atlas-based auto-segmentation performance for salivary glands was improved by increasing the quantity and quality of atlas cases in the library. The results provide novel insight into the behaviour of auto-segmentation algorithms. This performance evaluation of OAR segmentation in head and neck radiotherapy provides the basis for clinical implementation of the MIM Maestro ™ auto-segmentation software at the British Columbia Cancer Agency, Vancouver Centre.

PREFACE

This thesis is an original intellectual product of the author, E. Khawandanh. The fieldwork reported in Chapters 3 and 4 is based on work accomplished at BC Cancer Agency (BCCA)- Vancouver Center (VC). The data were collected under the REB approved protocol certificate number (AUTOSEG-HNC H-14-01538). I was responsible for

1- Selecting H&N cases from previously treated cases and checking the completeness of the organs at risk structure sets for each case individually.

2- Anonymizing each case using MIM Maestro™ anonymization tool.

3- Building and characterizing the distribution of cases in a head and neck ABAS tool using MIM™ atlas constructed from the anonymized cases.

4- Testing the performance of the designed in-house ABAS. This procedure was performed by determining geometric and dosimetric assessment indices.

Chapter 5 is based on segmentation data provided by Dr. John Wu. Dr. Wu validated the delineation of 206 parotid glands for 103 head and neck cancer cases. Haley Clark (PhD. candidate at UBC) was responsible for the anonymization process. This data was obtained under the REB approved protocol certificate number (H07-02073). My responsibilities included building the parotid in-house atlas, examining the user-defined settings and comparing the results with the built-in atlas using different

indices as explained in the thesis. A version of chapter 5 was presented at the annual scientific meeting of CARO in 2015 and the abstract has been published. Eman Khawandanh, Cheryl Duzenli, Haley Clark, John Wu, Steven Thomas, and Eric Brethelet: (Performance optimization of atlas-based parotid gland auto-segmentation using an in-house atlas library for head and neck radiotherapy planning). Radiotherapy and oncology. Vol.116 supplement 1, abstract number 86.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

3DCRT	Three-Dimensional Conformal Radiotherapy
ABAS	Atlas-Based Auto-Segmentation
BCCA	British Colombia Cancer Agency
CT	Computed Tomography
CTV	Clinical target volume
CV	Coefficient of variance
DEF	Image deformation
DICE	Similarity coefficient
D_{max}	Maximum dose
D_{mean}	Minimum dose
DoF	Degree of Freedom
DVH	Dose-Volume Histogram
GTV	Gross target volume
H&N	Head and neck
HD	Hausdorff distance
IMRT	Intensity Modulated Radiotherapy
LT	Left side
NMI	Normalized Mutual Information
NPC	Nasopharyngeal Carcinoma

OAR	Organ at risk
PRV	Planning organ at risk volume
PTV	Planning target volume
RT	Right side
RTOG	Radiation Therapy Oncology Group
SD	Standard deviation
SPICE	Smart Probabilistic Image Contouring Engine
STAPLE	Simultaneous Truth and Performance Level Estimation
V_{ABAS}	Volume of auto-segmented structure
VCC	Vancouver Cancer Center
VMAT	Volumetric Modulated Arc Therapy
$V_{reference}$	Volume of reference structure
ΔD	Dose difference
ΔV	Volume difference

ACKNOWLEDGMENTS

Despite the fact that only my name appears on the cover of this thesis, many people have contributed to its accomplishment. I owe my acknowledgments to all those who have made this dream possible. Because of them, my Master's experience has been one that I will cherish forever.

My deepest gratitude is to my supervisor D**r. Cheryl Duzenli.** I have been amazingly lucky to have an advisor who taught me how to question thoughts and express ideas. The door to her office was always open whenever I had a question. Her patience and support encouraged me to overcome many situations and finish this thesis. Without her support, this thesis would have never been accomplished.

Also, I would like to thank **Dr. John Wu** who performed the segmentation of 206 parotid glands and **Tania Arora, Vivian Cheung and Kelly Heung** for the manual segmentation of the eight test cases.

Also, I appreciate the financial support from Saudi Arabian Cultural Bureau (SACB) and National Guard Health Affairs (NGHA-Jeddah) throughout the journey to the completion my degree.

Most importantly, none of this would have been possible without the love and the patience of my parents **Hesham Khawandanh and Suhaylah Shajaruldeen**. My

parents, whom this thesis is dedicated to, has been a continuous source of love, care, concern, support and strength among all these years. Their support and care helped me overcome difficulties and stay focused on my graduate studies.

Further, I would like to express my heartfelt gratitude to my **auntie Afaf and my siblings Khulood, Duaa', Afnan, Lenah, Amal, Muhammad, and Abdulwahab**. They have always been with me despite the geographical distance between us.

I am also grateful to my extended family who has aided and encouraged me throughout this effort. I warmly appreciate their support. Many friends have helped me stay strong during these difficult years. I greatly value their friendship and appreciate their belief in me.

DEDICATION

To my parents Hesham and Suhaylah

CHAPTER 1: INTRODUCTION

1.1 RADIOTHERAPY AS A TREATMENT CHOICE FOR HEAD AND NECK CANCER

Head and neck cancer includes tumours arising in the upper aerodigestive tract including the oral cavity, oropharynx, larynx, and hypo-pharynx. In Canada, for oral and laryngeal cancer alone, the estimated numbers of new cases in 2015 were 4400 and 1050 respectively. The estimated numbers of deaths were 1200 and 380, respectively [1]. Worldwide, an estimated 644,000 new cases are diagnosed each year, with two-thirds of these occurring in developing countries [2]. It has been published that men are affected significantly more than women with a ratio of approximately 3:1. Smoking and alcohol consumption are the most common etiological factors. Dietary factors are also important risk factors for oral and pharyngeal cancers. Human papillomavirus (HPV) is recognized to have a role in head and neck squamous cell carcinoma (HNSCC) [3]. Due to the anatomical location of head and neck cancer, the presence of bilateral lymph nodes metastases and the sensitivity of the lymph nodes to radiation, radiotherapy represents one of the most effective therapeutic approaches in the management of head and neck cancers. However, following head and neck radiotherapy, late treatment side effects are highly predominant and these have an impact on both organ function and patient quality of life. Various nervous system complications can result from radiation therapy such as cerebral radionecrosis and spinal cord radionecrosis. For salivary glands, radiation-induced xerostomia is the

most commonly reported late side effect. Thus, radiation therapy can result in difficulties with speech, swallowing and dental care. Permanent hearing loss may occur in 40-60% of patients who receive radiotherapy to areas such as the nasopharynx or para-nasal sinuses [4].

In 1991, Emami et al. [5] presented dose tolerance limits for organs at risk (OAR), for a three volume categories (one-third, two-thirds, and whole organ) in terms of 5% and 50% chance of a specific adverse side effect occurring within five years. These dose limits were primarily relevant for dose prescriptions of 1.8 and 2 Gy per day for 5 days per week. This publication aimed to address the clinical demand for organ tolerances for the purpose of treatment planning based on available information up to that time. Nevertheless, there were obvious limitations in the publication, as it was a literature review up to 1991. This completely pre-dated the Three-dimensional conformal radiotherapy (3DCRT), Intensity Modulated Radiotherapy (IMRT), Image-guided radiotherapy (IGRT), and Volumetric Modulated Arc Therapy (VMAT) era. Three arbitrary volumes with only one severe complication endpoint chosen and the results were only valid for conventional fractionation. Over the last two decades, the practice of radiation oncology has completely changed. The choice of complication endpoints has been significantly updated. Also, there have been radical changes in radiotherapy technologies such as the utilization of 3DCRT, IMRT, IGRT, and VMAT, which have become standard treatment techniques with sophisticated evaluation tools. As a result, dose distributions have become more complex. In contrast, for the new treatment paradigms, such as hypo-fractionated radiotherapy, SBRT, and SRS,

prescription schemes have changed to doses > 2 Gy per fraction for hypo-fractionated radiotherapy [6], and sometimes up to 20 Gy per fraction for SRS techniques. It is known that normal tissue dose tolerance and tumour dose-response highly depend on the dose per fraction and the number of fractions [7]. All of these factors point to the need for more reliable data on biological response to radiation therapy. Throughout the last two decades, a large volume of published data has become available to establish a relationship between dosimetric parameters and normal tissue clinical outcomes. Because of variations in methodologies, across publications, it is a difficult task for practicing radiation oncologists to apply these data. Recognizing these challenges and the apparent need for a simple format, the Quantitative Analysis of Normal Tissue Effects in the Clinic (QUANTEC) was initiated [8]. QUANTEC was sponsored by ASTRO in association with the AAPM. The goals of the QUANTEC group were to review the available literature on volumetric/dosimetric normal tissue complications and provide a simple data set to the radiation oncologist, physicists, and dosimetrists to be utilized in treatment planning.

1.2 FROM CONVENTIONAL TREATMENT TO VOLUMETRIC RADIOTHERAPY

In the 1980s, Three-dimensional Conformal Radiotherapy (3DCRT) became available for cancer treatment[6]. However, the side effects of radiation, such as skin reaction and xerostomia, were quite severe. The prescribed dose in radiation therapy takes into account the tumour control probability (TCP) and normal tissue

complication probability (NTCP). Accordingly, for 3DCRT, the therapeutic dose was limited by the dose tolerance of the adjacent organs. To manage this problem, the dose was often delivered in different phases. Each phase had a different dose prescription and beam configuration in order to limit dose to organs at risk (OAR). For example, for T2b N2 M0 stage Nasopharyngeal Carcinoma (NPC), the treatment could be split into two phases.

In Phase I, 40 Gy was prescribed to be delivered in 20 fractions to a broad area. Then, in Phase II, 26 Gy was delivered in 13 fractions to a much smaller region. Each phase consisted of different numbers of beams and different beam directions [9]. Moreover, due to the large extent of the target in some cases, a single field was insufficient to cover the entire region and, thus, separate fields covered the superior and inferior regions. This approach created another concern, that being how to ensure the dose at the junction of the two fields was neither too hot nor too cold. This was addressed by moving the junction several times at intervals of a certain number of fractions. Additionally, the dose homogeneity and its conformity to the shape of targets could not be optimized by 3DCRT techniques and significant amounts of normal tissue were irradiated.

Modern radiation therapy planning is highly customized to the three-dimensional (3D) shape and volume of the structures. The implementation of Intensity Modulated Radiotherapy (IMRT) has improved radiotherapy outcomes by reducing patient complications [10]. IMRT has the ability to deliver the dose using treatment fields with variable beam shape and intensity throughout each field. As a result, a more

conformal dose distribution to the target can be achieved. Also, the dose to the critical organs may be reduced significantly compared to 3DCRT plans. Parotid-sparing trials indicated that the incidence of grade ≥2 xerostomia one year after treatment was significantly reduced with the utilization of IMRT compared to 3DCRT (38% versus 74% respectively) [11]. However, to achieve this level of dose conformity, longer treatment time was required due to need for a large number of beams and the speed of the beam fluence modulation by the movement of multi-leaf collimator (MLC) leaves. Subsequently, a large number of beam monitor units (MUs) were used.

In 2008, Karl Otto [12] introduced an advanced form of IMRT that could be performed efficiently on a linear accelerator with a single rotation of the gantry. This became known as Volumetric Modulated Arc Therapy (VMAT). This technique was based on an aperture-based algorithm for treatment plan optimization where the dose can be delivered during a 360-degree rotation of the beam using variable dose rate, field shapes and gantry speed to achieve the desired results. The dose optimization process depends strongly on the shape of the target and the proximity of the organs at risk. Because VMAT utilizes the full dynamic range of gantry motion and dose rate, treatments can be delivered in significantly shorter times than afforded by IMRT [12-15]. This significant advantage has made VMAT an attractive choice for many radiation therapy treatment sites today.

Because of the complex geometry of head and neck cancer targets, and the proximity of the OARs VMAT has become the preferred technique for head and neck cancer treatment. VMAT has also facilitated the introduction of the simultaneous

integrated boost technique, allowing the delivery of different dose prescriptions to various target volumes within the same treatment fraction. It has been stated in many different studies that VMAT achieves suitable conformal dose distributions with enhanced target coverage and normal tissue sparing compared to conventional techniques [16]. As a result, a high dose to the target can be achieved while maintaining the dose to normal tissue within the recommended tolerances. However, as this technique is customized to the shape, volume and location, to accomplish high level of conformal dose distribution, accurate delineation of the target and OAR volumes is required.

1.3 DELINEATION INACCURACY IN ORGANS AT RISK AND ITS IMPACT ON TREATMENT PLANS

Delineation, contouring and segmentation are synonyms that used alternatively in this thesis explaining the definition of the target volumes and the organs at risk (OARs) by drawing lines on the CT image planes. These lines are combined together to create a 3D volume for each delineated structure. These contours are used for dose optimization to design treatment plans and also used to create dose volume histograms (DVHs), which help to identify the amount of dose delivered to each structure at the end of the treatment. However, geometric uncertainty in the delineation of target volumes and OARs is a potential source of uncertainties in the radiotherapy process [17]. It has been suggested that, for different tumour sites, the greatest systematic error

in radiotherapy planning is related to the ability of the delineator to localize and contour the target volume in a consistent way [18]. The magnitude of this variation can be defined by various factors such as the images used for contouring (image modality, resolution, contrast), human factors (individual experience, professional background), and the utilization of the delineation guidelines [17,19,20]. This variation has implications for plan evaluations and meaningful comparisons between institutions [21]. Target and OAR delineation variability in head and neck cancer treatment has been investigated in different studies and a summary of these studies follows below.

A study presented by Brouwer et al. [22] aimed to identify the location where inter-observer variation in OAR segmentation was considered significant. The author investigated the magnitude and 3D localization of this variation between five expert observers for head and neck OAR delineation. For the purpose of this study, each organ was divided into sub regions to identify the amount of variation in cranial, caudal, medial, lateral, anterior and posterior directions. Due to the larger image resolution in the cranial-caudal direction, indistinctness of the delineation guidelines, the limitation of segmentation only performed on transverse CT slices, and poor differentiation from adjacent tissues, the most significant variations were indicated in the cranial, caudal and medial regions for almost all OARs. It has been also mentioned that these errors in contouring may affect treatment results by either increasing the dose to normal tissue, which might subsequently increase treatment morbidity or by inadequate target delineation, which increases the risk of tumour recurrence.

Several other groups have investigated the dosimetric impact of inter-observer

variation. A study was done by Loo et al. [23] aimed at evaluating the inter-clinician variation in parotid gland segmentation and its impact on IMRT plans. Four radiation oncologists and three radiologists participated in this study. The tolerance dose for parotid gland sparing was set to 24 Gy. This study showed that upon superimposing the actual IMRT treatment plans on 70 sets of study contours, although the mean parotid dose achieved in the original plan was within 10% of 24 Gy for all cases, only 53% of the oncologist volumes and 55% of radiologist volumes met the same target. Mean dose was within 20% of 24Gy for 80 and 90% of the oncologist and radiologist volumes, respectively. It is also stated that parotid DVHs of 46% of the study contours were significantly different from those used clinically, such that different treatment plans would have been produced.

Nelms et al. [24] investigated the geometric and dosimetric variation between clinicians' OAR delineations. The scope of this study was international. Clinical sites from six different countries participated in this study. The study examined OAR variability over a population of 32 independent treatment planners for a single oropharynx case. Only six OARs were evaluated. The clinical team in the center where the patient was treated created the reference OAR set. They found a significant organ-specific variation in the contouring over a population of clinicians. Of these six OARs, the most variable organs were the brainstem and the two parotid glands and the least variable was the brain. The dosimetric impact of the variation was estimated by overlaying the reference OARs onto the dose grid optimized according to each planner's OARs, and then the differences in the dose were quantified. Depending on the

level of the contour variations and the dose gradients in the plan, they found substantial dose differences resulting from contouring variation ranging from (–289% to 56%) for mean OAR dose and (–22% to 35%) for maximum dose. These variances underscore the importance of accuracy and consistency in OAR contouring.

Feng et al. [25] performed a study designed to evaluate the variability of organ at risk (OAR) delineation and the resulting impact on IMRT treatment plan optimization. Three radiation oncologists delineated OARs of ten oropharyngeal cancer patients. OARs included the parotids, submandibular glands, pharyngeal constrictors, larynx, and glottis. A total of 30 IMRT treatment plans were created to find the effect of this variation. The mean difference in total volume for each OAR across all the contours was 1 cm^3 (σ 0.5 cm^3, range 0.1-5 cm^3) and the pharyngeal constrictors had the largest mean difference in volume. In the dosimetric analysis, the mean difference in OAR dose was only 0.9 Gy (range 0.6-1.1 Gy, σ 0.1 Gy), with the smallest difference observed for glottis and the largest for both submandibular glands and the larynx.

Due to the reported geometric and resulting dosimetric variation in manual segmentation, various published studies have suggested solutions aimed at reducing inter- and intra-observer variation and maintaining delineation consistency. It is strongly recommended to use segmentation protocols, as they significantly improve the segmentation consistency of targets and OARs [19,22]. Moreover, the implementation of various imaging modalities, such as MRI, which provides superior soft-tissue contrast of surrounding soft tissues, and PET, which provides a good

localization of the target, enhanced the reproducibility of contour segmentations [27,26]. Furthermore, the introduction of protocol based one-on-one training has a positive impact on inter and intra clinician variation [23]. Recently, different studies demonstrated that the reproducibility of structure contours improved and the inter-observer variation reduced by using atlas based auto segmentation [36].

1.4 CLINICAL IMPLEMENTATION OF ATLAS BASED AUTO-SEGMENTATION

Manual segmentation is the standard routine for organ delineation in many clinics, despite being a time-consuming task subject to inter- and intra-observer variation. Additionally, the emerging adaptive radiotherapy paradigm will be limited to a very small numbers of cases because of the necessity for volume re-segmentation, which takes substantially increase the physicians' workload. To widely implement adaptive radiotherapy, an automatic segmentation is required [29,30]. Atlas Based Auto Segmentation (ABAS) is a potential solution designed to automatically generate contours in less time than manual segmentation, without manual intervention. The word "atlas" is referred to a particular model includes a complete description of the relationships between sets of images, which will be discussed in more detail in section 2.3. ABAS aims to reduce the manual workload and improve the quality of the result. It is hoped that ABAS will become a readily available tool for different treatment sites [28,31,32].

ABAS incorporates prior information into the segmentation process. It uses a CT data set or multiple data sets that include previously validated contours forming an atlas. By using an atlas selection method, new contours are generated automatically after deformable image registration process between atlas and test image. Commercially available treatment planning software generally provides an auto segmentation tool with a built-in library of atlas cases. The user may also have the option to build an in-house atlas library with institution specific CT sets and contours.

As ABAS becomes state-of-art in the contouring field, more and more studies are being published on the validation of its performance for clinical use. A literature review was done based on a search of PubMed, ScienceDirect and Google scholar databases. Publications that related to atlas based auto segmentation specific to head and neck radiation therapy treatment planning were included in this review. A summary of findings follows below.

All the studies confirmed significant reduction in contouring time. One of these publications (Walker et al. [33]) aimed to determine the feasibility of real-time workflow and assess contouring time-reduction using auto-segmentation software for head and neck OARs in a representative clinical population. Also, it evaluated individual OAR acceptability of auto-segmentation definition of head and neck OARs using resident and expert physician comparators. A total of 40 cases were used in this study, and each case was delineated by one of the eight residents and approved by one of the seven head and neck physicians. The result showed 31% time reduction between manual contouring process and ABAS followed by manual editing. They confirmed that

atlas-based auto-segmentation provided a detectable time saving in the generation of OAR over manual contouring. However, attending the oncologist approval for all OARs remains vital, as many structures, including the optic chiasm and optic nerves, were poorly defined by the auto-segmentation algorithm.

A number of publications validate the performance of ABAS for lymph-nodes CTVs and OARs segmentations and investigated its influence on reducing the variation between observers. Commowick et al. [34] constructed an average symmetric single atlas subject from 45 node-negative pharyngo-laryngeal squamous cell carcinoma CT images databases. They used it to auto-contour lymph-node levels (levels II, III and IV), the parotids, the brainstem, the spinal cord, the mandible and the sub-mandibular glands. They evaluated its performance using twelve test cases. For the qualitative and quantitative evaluation, a leave-one-out test evaluation method was accomplished. They found over-segmentation in most of the structures especially for lymph nodes level II. They attributed this variation to the differences in the neck fat amount as well as position differences between the patient and the atlas. They mentioned the need to quantify an inter- and intra-observer variability of the manual segmentations and compare it to the over-segmented structures that were created by the atlas. They expect poor results for lymph-nodes auto-segmentation if this atlas is used to contour patients with node-positive tumours. Another study was done by Jinzhong et al. [38], who investigated the auto-delineation of low-risk CTV for patients with unilateral tonsil cancer using sixteen patients previously treated cases. They identified six cases of median size in terms of body mass index as atlas cases and used the other ten cases

as test cases. The authors evaluated the performance of multi-atlas subject versus single-atlas subject for low risk CTV. In addition to that, they studied the effect of individual oncologists' practice on auto-segmentation. They stated that the difficulty of defining CTV is due to the lack of clear anatomical boundaries. Their results demonstrated that using ABAS for low-risk CTV is feasible for unilateral tonsil cancers. They showed that multi-atlas segmentation is more effective than single-atlas segmentation in delineating the target volume. Speight et al. [29] validated the usage of ABAS in the adaptive radiotherapy field by assessing the accuracy of automatically segmented CTV volumes produced by ABAS on re-CT and comparing them to volumes manually contoured by three observers at multiple time points, thus estimating inter- and intra-observer variation. Also, they estimated the time that ABAS could save in the adaptive H&N radiotherapy. The results showed that the accuracy of CTV volumes created by ABAS were close to inter-observer variation. However, small regions with significant discrepancies required editing before clinical use. They stated that using ABAS reduces contouring time by a factor of three, which helps facilitate adaptive radiotherapy. Moreover, Stapleford et al. [35] focused on the lymph node regions for HNC patients. The goal of this study was to assess if auto-segmentation could decrease inter-physician variability while maintaining accuracy. They analyzed how physicians modify auto-segmented CTV in terms of size, shape, and position. Five patients were segmented using Velocity Medical Systems' built in atlas and the results were modified by five radiation oncologists. They referred the discrepancies between manual and automatic contours to the presence of positive lymph nodes in the test cases, whereas

the auto-contours were based on guidelines for N0 neck nodes. They found that most of the time, physicians thought that the auto-contours require modification because the volumes were too large; however, on average they removed only 7% of the auto-segmented volumes and generated modified contours that were larger than their initial manual contours. In a recent study that was published early in 2015, Tao et al. [36] performed a multi-institution study to evaluate whether using multi-subject atlas-based auto-contouring with manual modification can reduce inter-observer variation and improve the consistency of dosimetric parameters for OARs. For the purpose of this study, eight oncologists from eight independent institutions manually contoured twenty OARs for sixteen NPC patients using either Focal (Elekta), Pinnacle (Philips) or Eclipse (Varian) treatment planning systems for manual delineation. ABAS (Version 2.01.00, Elekta AB) was used to create auto-contoured structures. Within a month, the eight radiation oncologists reviewed and edited the final multi-subject auto-segmented OARs using consensus guidelines. Accordingly, the results showed that edited ABAS reduced the variation of volume and the center of mass shift in 3D for most of OARs compared to the initial manual contours. Moreover, edited ABAS reduced inter-observer variation of the D max and mean dose to these organs especially for small volumes and improved the consistency of these dosimetric parameters.

Another approach of different studies is to compare the performance of a multi-subject atlas with a single-subjects atlas. Teguh et al. [37] quantified the accuracy of auto-segmentation using ABAS and assessed its applicability for clinical use. They compared the result of using multiple-subject atlas with a single-subject atlas. Also,

they quantified the differences among the clinically used contours, auto-contours, and edited auto-contours. The quality of all contours were quantified by the similarity coefficient and scored by an expert observer. They found consistent superior performance by using a multiple-subject atlas compared to the single-subject atlas. However, the auto-contoured structures still require editing to be used for planning optimization. Pirozzi et al. [39] also evaluated the auto-segmentation results from a single best-matched atlas compared to the results from multiple atlas using a 20 subject head and neck atlases containing targets and normal structures. They found that a multi-atlas segmentation approach achieves the closest similarity to manually defined contours. Using four best matches multi-atlas segmentations was the best trade-off between accuracy and segmentation time.

In addition to that, the performance of different auto-segmentation approaches was also investigated. Sims et al. [40] aimed to evaluate the accuracy of an atlas-based automatic segmentation (ABAS) tool (Version 3.1 ISOgray^{TM} ABAS system by DOSIsoft) for the H&N using a mean image atlas approach. The mean image was constructed from 45 node-negative pharyngeal-laryngeal squamous cell carcinoma CT datasets. In this work, only the brainstem, mandible and parotid automatic contours were studied. They compared the manual segmentation of two expert oncologists with the modified ABAS contours. They noted systematic over-segmentation of parotids that was significant in the anterior/posterior and med/lateral directions. Under-segmentation of the brainstem was also noted with maximum discrepancy in the superior direction. Careful review and editing of ABAS structures were recommended.

Thomson et al. [35] evaluated the accuracy of Smart Probabilistic Image Contouring Engine (SPICE) automatic segmentation to define salivary glands, swallowing structures and cochlea. It is a commercially available algorithm that combines an atlas-based and model-based approach to delineated head and neck lymph node levels and OARs. The gold standard contours were created using the Simultaneous Truth and Performance Level Estimation (STAPLE) algorithm. They found that SPICE auto-contoured structures were less accurate than manual structures, but acceptable for the definition of salivary glands. However, modified SPICE contours remained inferior to manual contours and the utilization of SPICE compared with manual delineation did not result in time-saving or efficiency improvement.

La Macchia et al. [30] evaluated various commercially available auto-segmentation tools to determine their clinical viability in delineating the head and neck, pelvis and thorax OARs for adaptive radiotherapy. All volumes of interest were delineated on treatment planning CT manually by three in-field specialized oncologists for five locally advanced head and neck cases that represented their atlas cases. Using three commercially available programs—ABAS 2.0 (Elekta), MIM 5.1.1, (MIMVista), and VelocityAI 2.6.2(Velocity Medical Systems)—to auto-segment a re-plan CT for each patient. A single-subject atlas approach was used for MIM and VelocityAI. They concluded that there was significant time-savings with auto-contouring of approximately 60 min for head and neck cases. They found a higher degree of agreement between contours using MIM 5.1.1 compared to ABAS 2.0 and VelocityAI. 2.6.2.

Furthermore, the impact of geometrical variation on the dose distribution was also discussed in different publications. The primary aim of the study by Voet et al. [41] was to investigate the impact of adjusting the auto-segmented neck nodes levels and OARs on the final dose distribution. They compared the dose distribution using auto-contours with the dose distribution using manually edited auto-contours. The quality of auto-contours was evaluated by the similarity coefficients and mean distances between edited and non-edited contours in the two structure sets. The second aim was to investigate whether observed DSC coefficients or mean contour distances predict for PTV under-dosage. The dosimetric impact of editing auto-contours was investigated using 18 IMRT plans generated based on edited and non-edited auto-contours PTV and salivary glands. For each plan, the mean doses of salivary glands were compared and the dose coverage of the PTVs were evaluated by quantifying V95% and D99%. They stated that editing of target contours generated by ABAS is essential to avoid PTV under-dosages. Furthermore, for highly conformal IMRT plans with steep dose gradients towards OARs, small geometrical differences in the target volume led to a large dosimetric impact on the dose coverage. For OARs, editing salivary glands has a very small impact on the mean dose. DSC coefficients and mean contour distances are useful for geometric quantification but have no significant dosimetrical impact on the plan quality.

To summarize, all the studies quantified and qualified the efficiency of ABAS performance for head and neck cancer targets and OARs for different approaches. Although most of the generated contours required editing, substantial reduction in the

contouring time was achieved using ABAS. In addition, all the studies confirmed that using a multi-subject atlas was superior in performance to a single-subject atlas. However, oncologist approval for all CTVs and OARs remains essential.

Although ABAS helps reduce delineation time, manually modifying the generated contours is highly recommended. . This process remains a tedious task that might negate the time-saving associated with auto-segmentation [21.35]. Conventionally, manual segmentation is considered to set the standard that auto-segmentation system should approach. However, inter-observer variation makes assessing the accuracy of auto-segmentation methods difficult because of the absence of ground truth [42]. In addition, uncertainties in patient positioning and anatomical changes during the course of radiation may outweigh contouring discrepancies when assessing actual delivered dose distributions. It has been stated in some studies that despite volume and overlap differences, OAR dose differences were small. These arguments would suggest that if the auto-generated contours fall within the manual observers' variation, no significant impact on the treatment quality should be observed by using auto-segmentation versus manual segmentation.

To the best of our knowledge, there is no published study looking at parameters that might influence ABAS performance such as the effect of increasing the number of subjects in the atlas library with different characteristics (i.e., gender, ages, head shapes and sizes) to cover a larger patient population. Likewise, no publication systematically compared the geometrical and dosimetrical auto-segmentation result with intra- and inter-observer variation.

1.5 THE PURPOSE OF THIS THESIS

This thesis was aimed at evaluating and improving auto-segmentation of OARs in head and neck radiation therapy treatment planning. Chapter 2 describes the methodology common to auto-segmentation algorithms and performance assessment of segmentation quality in general.

Chapter 3 describes the first of three studies performed. The objective of this study was to quantify the inter- and intra-observer geometric variation in manually defined OAR contours for head and neck treatment planning and the dosimetric variation resulting from this geometric variation. Timing data for the manual segmentation process was also acquired. The results of this study provided benchmark data that were used for the subsequent two studies.

Chapter 4 describes the second study. The objective of this study was to build and evaluate an in-house head and neck atlas for ABAS auto-segmentation of organs at risk.. Organ by organ assessment of contouring quality and overall timing was performed and compared with results for the manual processes reported on in Chapter 3 and other published studies.

Chapter 5 describes the third study, which was an investigation into strategies to improve ABAS segmentation of parotid glands. An in-house atlas designed specifically for parotid gland auto-contour consisting of 103 subjects was constructed. A single expert oncologist validated these contours to overcome the influence of inter-observer variation. Different user-defined settings were evaluated to define the

settings associated with the best performance. The ABAS results were compared geometrically and dosimetrically with those of the first two studies.

Chapter 6 summarizes the conclusions reached in this body of work and describes potential future work.

CHAPTER 2: ATLAS BASED AUTO-SEGMENTATION

2.1 INTRODUCTION

Current medical imaging modalities, such as Computed Tomography (CT) and Magnetic Resonance Imaging (MRI), allow three-dimensional (3-D) image reconstruction of internal organs. Also, the implementation of Single Positron Emission Computed Tomography (SPECT) and Positron Emission Tomography (PET) has helped to provide specific information about the anatomy and function of organs. Consequently, these technologies have expanded the knowledge of normal and abnormal structures for medical decision making. In the radiotherapy field, imaging is an important part of the treatment planning routine since it is used to define the treatment target and the surrounding normal structures.

As mentioned in the first chapter, the targets and organs at risk (OARs) are routinely manually delineated on CT images. Contoured images are then transferred to a treatment planning system for plan optimization and dose calculation. In the intensity modulated radiotherapy (IMRT) scheme, plan optimization strongly depends on the segmentation accuracy and the distance between the target and the other organs. Therefore, accurate delineation is necessary for treatment planning. However, manual contouring is a time-consuming task that is subject to intra- and inter-observer variation. Various studies have attempted to solve this problem by creating contouring guidelines that seek to reduce these variations. Nevertheless, there is no published

evidence that shows using guidelines reduces workload or time. Recently, automatic delineation has become a very active field of research. This approach could significantly reduce contouring time as well as reducing variation in the organ boundary definition.

Moreover, the recent introduction of advanced techniques in the field of radiotherapy, such as adaptive re-planning, is expected to result in significant increases in contouring workload. Therefore, there is a demand for an accelerated delineation process, which may be achieved by automatic segmentation.

Different auto-segmentation methods have been discussed in various publications. One of these methods is referred to as atlas-based auto-segmentation (ABAS). ABAS generates the contours by deforming reference images to the test image and subsequently mapping the reference contours onto the test image. Many studies have attempted to validate the accuracy and performance of ABAS for clinical work. It has been shown that the implementation of ABAS for head and neck target and normal structures contouring may save up to 60 minutes of the oncologist's time [30].

This chapter provides a brief summary of auto-segmentation methods that are generally used in radiotherapy. Also, it offers a detailed description of the atlas based auto-segmentation (ABAS) approach including an introduction to mutual information and deformable image registration to provide an understanding of the auto-segmentation process. Furthermore, it takes a closer look at various choices of atlases in the atlas-based segmentation process. In addition to that, it demonstrates more details about qualitative and quantitative methodologies used to evaluate the accuracy of ABAS performance.

2.2 OVERVIEW OF AUTO-SEGMENTATION METHODS IN RADIOTHERAPY

An auto-delineation tool is used to generate anatomic structures automatically for the purpose of treatment planning. Different automatic contouring techniques are used in the radiotherapy field using various types of image content. These techniques can be divided into two main categories: delineation methods that do not require prior information and contouring methods powered by prior information. Conventional auto-contouring methods used in the radiotherapy field do not require prior information in the contouring process. They are based on the analysis of image information and properties such as voxel intensities and/or image gradient. Among these methods, region based methods such as adaptive thresholding as well as edge detection based methods have been applied for decades [42]. These methods take into consideration the type of tissue, as different tissues are associated with different image intensities (e.g., bone vs. soft tissue). They may be used to obtain clinically acceptable contours by auto-segmentation in different sites such as whole brain, as the brain is almost a homogenous organ that is confined within the skull. However, when we are aiming to label the anatomical structures rather than tissue types, there is no direct relationship between a voxel's value and the name of the structure that should be assigned to it. For example, different structures that consist of the same type of tissue (e.g., vertebrae vs. ribs) cannot be differentiated from one another just by looking at their intensity values. But, they can be distinguished by their geometrical constraints and spatial relationship to other structures. Therefore, these spatial relationships (e.g., neighbourhood relationships) should be considered to improve the delineation results.

A large number of investigative works have participated in studying the spatial information such as size, location, and shape of the organs to improve the robustness of the automatic delineation and to compensate for the weakness of CT soft tissue contrast, which limits the accuracy of boundary definition. These studies improved delineation results by using methods that incorporate prior information used during the delineation process. These methods carry information about the shape, location, and size of the anatomical structures and the organs' appearance in different imaging modalities. They become advantageous compared to considering only the intensity information of the images. Among these methods, model-based segmentation and statistical shape model approaches have been established as one of the successful methods. These approaches work by matching a model that includes information about the estimated shape and appearance of the structure in new images. A straightforward approach to combine this information is used to examine different numbers of shapes by statistical means, leading to statistical shape models (SSMs) [43]. The correct shape of structures can be estimated by using landmarks or surface meshes in combination with statistics of voxel intensities. Another method that also incorporates prior information is referred to as atlas based auto-segmentation, which uses CT data sets that include contours validated by the user. Then, by using one of the atlas selection methods, the contours are generated automatically after using the deformable image registration process.

Hybrid segmentation approaches have been presented during the last few years. These approaches combine different methods together to compensate for the weakness

of using each approach alone. Commonly, atlas based auto-segmentation is combined with the statistical appearance models (SAMs) or statistical shape models (SSMs) that provide better segmentation results [44]. In addition, atlas-based auto-segmentation, statistical intensity, and the prior spatial model have been joined to improve atlas based segmentation results and minimize the necessity of manual post processing [48]. The results of hybrid approaches are often superior compared to non-hybrid approaches. Figure 2.1 provides a summary of the segmentation methods used in radiotherapy field.

2.3 ATLAS BASED AUTO-SEGMENTATION (ABAS)

Incorporating prior information such as geometrical constraints of each structure with the spatial relationship between different structures improves segmentation performance. A complete description of such relationships is referred as an atlas. It can be generated by contouring a single image manually or by obtaining an average segmentation from multiple segmented images of various individuals to construct an average atlas case. Other means of utilizing multiple atlas cases also exist. This process of atlas-based auto-segmentation (ABAS) will be discussed in more detail in section 2.3.4.

The automatic segmentation of an image that is obtained by mapping its coordinate space to the atlas is classified as atlas-based auto-segmentation. Therefore,

for achieving accurate mapping, each voxel of any structure in the test image can be defined by finding the structure at the corresponding location in the atlas under that mapping.

The mapping process is obtained commonly by the deformable image registration process, as there are significant differences between different individuals in the atlas itself and between the selected atlas cases and the test case. After the registration, all the contours are transferred from the chosen atlas case to the test case. In the case of multiple closest atlas matches, each structure will contain number of contours equivalent to the number of best matches. To determine if a voxel is included within the segmentation or not, different algorithms can be used for this purpose (i.e., majority vote or STAPLE).

2.3.1 The basics of atlas-based auto-segmentation (ABAS)

Atlas-based auto-segmentation is an automatic delineation of anatomical structures generated by transferring the contours from atlas images to the test image after a registration process. To delineate an image using an atlas, both images should be registered by creating a coordinate map between both images. The accuracy of the delineation strongly depends on the accuracy of the computed map.

Let us consider two 3D scalar images, a test image and an atlas image, and assume that each point in the atlas image has a corresponding point in the test image. This correspondence is defined as a coordinate transformation **T** that maps the image

coordinates of the atlas image onto those of the test image. According to this transformation map, any **x** point in the test image has a corresponding **T(x)** point in the atlas image. The transformation **T** is known as image registration. This is a very important part of the process because an accurate segmentation cannot be performed without acquiring an anatomically correct transformation map [45]. One way to evaluate the accuracy of the transformation is to analyze the mutual information of the images.

2.3.2 Mutual information and Entropy theory

Measuring of information is usually termed as entropy. It uses the probability distribution of specific information by counting the number of times each bit of information occurs and presents it as a measurement. Based on this concept, entropy can be described as the distribution of the image gray values. It can be estimated by counting the frequency of each gray value in the image and dividing it by the sum of occurrences. Such measurement in imaging registration field is called mutual information. It is based on the assumption that areas of similar tissue, which are presented as gray values, in one image would correspond to areas in the other image that also consist of similar gray values. These values might not be the same in both images but similar in the same image [46].

In the case of atlas-based auto-segmentation, the correct mapping between the

test and the atlas images is unknown a priori, as each individual has different shape and size of structures. Accordingly, the appropriate registration of any images can be quantified by measuring the mutual information MI (A, B) which is defined as:

$$\mathrm{MI}\ (\mathrm{A}, \mathrm{B}) = \mathrm{H}(\mathrm{A}) + \mathrm{H}(\mathrm{B}) - \mathrm{H}(\mathrm{A}, \mathrm{B}) \qquad (2.1)$$

Where H(A) is the entropy of the test image, H(B) is the entropy of the atlas image, and H(A,B) is the joint entropy of corresponding voxel pairs between the two images. Normalized mutual information (NMI) used as a modified version of mutual information that is found to be more robust [45,46]. It is defined as

$$\mathrm{NMI}\ (\mathrm{A}, \mathrm{B}) = \frac{\mathrm{H}(\mathrm{A}) + \mathrm{H}(\mathrm{B})}{\mathrm{H}(\mathrm{A}, \mathrm{B})} \qquad (2.2)$$

In order to improve the registration, pre-processing is recommended. It involves any image processing to prepare the images for registration. The most common example of pre-processing is to define a region or a structure of interest in the images that helps exclude other structures that might negatively affect the registration results. Other techniques reported are thresholding, which is used to remove image noise. Moreover, resampling the images isotropically attains similar voxel sizes in all dimensions. In addition, the image dimensionality and number of images are considered, as they might influence the registration.

2.3.3 Image registration techniques

Due to the improvement of image acquisition techniques, applications of image

registration have been developed significantly. Image registration is a process aim at determining the feature correspondence between two images (e.g. lining up the same anatomical structures) by finding the best geometric transformation that maximizes the similarity across the images. Therefore, the pair of images become closely matched and can be directly analyzed. This process involves several factors:

- Transformation: describes a geometrical change between the images and illustrates the way that one image overlays onto the other one. There are several classes of transformation process such as rigid and deformable transformations.
- Similarity metric: also known as a registration function, it quantifies the degree of similarity between the images. Features such as image intensities, landmarks, edges or mutual information can be used to determine the similarity.
- Optimization process: aims to maximize the similarity metrics after the transformation process.
- Validation protocol: qualitatively quantifies the performance of the registration techniques in terms of its accuracy, robustness and clinical utility [47].

For atlas-based auto-segmentation, the aim of image registration is to find a transformation $T:(x,y,z) \rightarrow (x',y',z')$, which maps any point in the atlas image into the corresponding point in the test image. There are several transformation models, ranging from simple linear transformations to complex non-linear transformations. The complexity of transformation is characterized by an increase in the number of the Degrees of Freedom (DoF) of the transformation. Linear transformation includes

rigid, affine and global transformation. A rigid transformation contains six degrees of freedom: three rotations and three translations. Affine transformation involves twelve degrees of freedom represented as rotations, translations, scaling and shears. A global transformation is another example of the linear transformation with a higher number of DoF [45]. Rigid and affine transformation models are frequently used for the registration of anatomical structures like brain or bones where a small number of DoF is sufficient to perform acceptable registration. However, these transformation models are insufficient for registering soft tissue like liver or lung, as significant deformation is expected. Furthermore, there are some considerable variations in the shape and size of anatomical structures between the atlas and the other individuals. Accordingly, non-linear or deformable transforms are required. As any registration process, a deformable registration algorithm consists of three main components: a transformation algorithm, a similarity measure, and an optimization process.

The **transformation algorithm** defines how the atlas image can be deformed to match the target one. It controls the way that image features can be transformed relative to one another. Moreover, it interpolates between those features where there is no information. Transformations used in a deformable registration field range from smooth variation that can be described by a small number of DoFs to dense displacement fields defined at each voxel.

One of the most common deformable transformations is the family of splines. Spline-based registration algorithms use control points in the atlas and target image,

then a spline function defines correspondences between these points. The "thin-plate" spline is one of the spline examples. Each one of the thin-plate spline control points has a global influence on the transformation in that perturbing the position of one control point can affect all the other points in the transformed image. On the other hand, B-splines are only defined locally in the region of each control point; affecting the position of one control point only affects the transformation in a small volume in the image space while all voxels outside this area remain in the same location. Because of this property, B-splines are often referred to as having "local support". B-spline based deformable registration techniques are widespread due to their applicability, transparency and computational efficiency.

The **similarity measure** identifies the level of image match. Image registration approaches based on the content of patient images can be divided into geometric and intensity approaches. Geometric approaches build models of certain anatomical elements in each image using corresponding point landmarks. Various deformable registration algorithms based on 3D geometric features use anatomical surfaces as landmarks. Intensity approaches match intensity patterns in both images using mathematical criteria. These criteria compute the intensity similarity measure of the atlas and the target and modify the transformation until the similarity measure is maximized. Similarity computation included differences in intensities, correlation coefficient, and mutual information. As mentioned previously, geometric registration uses anatomical information only while intensity-based registrations match intensity patterns without using any anatomical knowledge. Thus, combining both features in

the registration process results in more robust methods identified as hybrid algorithms [49].

The **optimization process** is the third component of the deformable registration process. It adjusts the parameters of the transformation in order to maximize the similarity measure. A good optimizer is one that instantly finds the best transformation. In deformable registration applications, designing an optimal optimizer can be difficult because the more deformable transformation needed, the more DoFs are required to define it, which means longer time is required [50].

2.3.4 Atlas approaches

Different atlas-based auto-segmentation approaches, widely used, with different atlas selection criteria have been discussed in several publications. The segmentation accuracy of each approach was also covered [31,32,42,39,44]. This section covers more details about the most common four different atlas approaches for atlas-based auto-segmentation. These approaches are:

a) Delineation using one single individual atlas,

b) Delineation using an average-shape atlas derived from an atlas library

c) Delineation using the single closest match atlas from an atlas library, and

d) Delineation using multiple closest matched atlases from an atlas library.

These approaches can be categorized according to the number of atlases used

per segmentation (single or multiple), the type of atlas used (individual or average), and the atlas selection (fixed, i.e., same atlas for all test cases, or variable, i.e., different individual atlases selected for each individual test case).

Single individual atlas approach (Figure 2.2) is the most straightforward strategy. It requires only a single manually segmented case. The selection of this case can be either random, or based on investigative criteria such as lack of artifacts, or normal appearance of the structures in the image. This strategy is the most commonly used atlas approach because it requires only one atlas case. Thus, the preparation effort for this atlas is less compared to the other methods.

As mentioned previously, smaller magnitudes of the deformation between test image and atlas images result in a higher accuracy of the matching. If the atlas is derived from an individual subject, then there is a high potential that this individual is an outlier in the population. In such a case, delineating other subjects using that atlas becomes insufficient and the segmentation accuracy expected to be low. An improved atlas would be one that is as similar as possible to many individuals. Such an atlas can be generated as an average case derived from many atlas subjects. This approach is described as an **average-shape atlas approach** (Figure 2.3).

One way of acquiring an average shape atlas from a population of subjects is to generate an active shape model (ASM). This provides a statistical description of a population of subjects by the average shape. Obtaining an ASM requires identifying the corresponding landmarks on all individuals [51,52]. Other methods are based

completely on deformable registration, such as active deformation models (ADM) [51].

Single closest match atlas approach is defined as finding the most similar case, from an atlas library, to the test case. There are at least two characteristics that describe the similarity between a test image and the selected atlas case. The first one is the image similarity measure using the mutual information, after either affine or non-rigid registration. The other is the magnitude of the deformation (i.e., non-rigid transformation) that is required to map the coordinates of the test image onto those of the selected atlas case. Based on these two characteristics, four different criteria for choosing the atlas subject that is most likely to produce the best segmentation of a given test image are:

1. Image similarity after acquiring affine registration (NMI affine):

After affine registration is acquired between the test case and the template, the atlas subject with the closest normalized mutual information value to the test image will be selected and used for the auto-segmentation process. Because only an affine registration needs to be computed, it is, therefore, computationally less expensive than the other criteria. It has been stated that this criterion may perform marginally better than the other three [44].

2. Image similarity after acquiring deformable image registration (NMI non-rigid):

After deformable registration is acquired between the test case and the template, the atlas subject with the closest normalized mutual information value to the

test image will be selected and used for the auto-segmentation process.

3. Average deformation over all atlas voxels:

After acquiring deformable image registration between the test case and the template, the magnitude of the deformation between the test image and template image and between each individual atlas and the template case is computed and averaged over all voxels. The atlas with the closest average deformation is selected and used for segmentation. This criterion is based on geometric similarity rather than intensity similarity, on which the previous criteria are based.

4. Maximum deformation over all atlas voxels (DEF max):

This criterion is similar to the average deformation and the only difference is that it uses the maximum deformation over all voxels rather than the average deformation. It pays more attention to outliers. The idea is that subject that matches well overall might be significantly inaccurate in some regions.

For the single closest matches approach, a template case that includes the mutual information of all the atlas subjects (by a preprocessing step) is usually used rather than registering each case individually. Thus, the process of finding the closest case becomes faster.

Multiple closest matches atlas approach is obtained by repeating a single closest match approach a number of times equivalent to a user-defined number. Each case chosen, as a close match, is non-rigidly registered to the test case, and then all the

contours are transferred from the atlas to the test case. At the end of this process, each structure will have several suggested segmentations, equivalent to the user-defined number of closest matches. In order to combine these contours into the final result, various algorithms, such as majority vote or STAPLE, can be used to make the decision. Successful applications of the multi-atlas approach have been reported in different studies [33,37,29,36]. A schematic diagram of atlas library construction and Single and multiple closest match atlas approaches are shown in Figure 2.4-2.6.

2.3.5 Voxel segmentation determination methods

As shown previously, employing multiple-atlas matches improves the segmentation accuracy in the atlas-based auto-segmentation process. Each atlas image is non-rigidly registered to the test image independently. Then, atlas contours are transferred to the test image to create a segmented version of the test image. However, several segmentations result from this process. In order to obtain a single segmentation, various algorithms can be used to combine the contours. These algorithms examine each voxel to decide if it should be included in the segmentation or not. A notable example of producing consensus contours in the context of medical image processing is the **majority vote approach**. This approach tests the degree of agreement between the individuals on incorporating each voxel in the segmentation. Therefore, for each voxel, the segmentations of the selected individual atlases are determined. Then individuals' votes agreeing or disagreeing to involve this voxel

within the segmentation are counted. Voxels with more than 50% of the total votes agreeing will be included within the contour [53,54]. Majority vote provides better results in the case of an odd number of segmentations. However, in the case of an even number segmentations, the voxel may be excluded because it will not pass the >50% criterion.

Another sophisticated method used to combine the segmentations is the expectation-maximization algorithm for simultaneous truth and performance level estimation (STAPLE) framework that was presented by Warfield et al. [55]. This approach computes a probabilistic estimate of the true segmentation by estimating an optimal combination of the segmentations and computes a measure of the performance level represented by each segmentation

2.4 ABAS PERFORMANCE ASSESSMENT TEST

The leave-one-out test is one of the most common assessment tests used to evaluate the performance of ABAS. In this test, one subject is excluded from the atlas library and used as a test case. Since it has all the structures manually contoured, contours generated by ABAS can be easily analyzed quantitatively and qualitatively. It can be used to compare the performance of using different approaches (i.e., single closest match vs. multiple closest matches), different segmentation algorithms (i.e., majority vote vs. STAPLE) or to compare different number of best matches (i.e., three closest matches vs. five closest matches).

2.5 SEGMENTATION ACCURACY ASSESSMENT METHODOLOGY

Traditionally, manual segmentation is considered to be the absolute ground truth that any automatic method has to approach. However, evaluating acceptable accuracy of auto-segmentation is not an easy task. The main reason being the absence of gold standard that can be directly derived from CT data. A variety of accuracy indices and endpoints have been adapted to provide a quantitative evaluation of ABAS performance. These use clinically generated contours as a reference for the purpose of evaluation. Some of these indices provide geometrical agreement between the reference and ABAS contours. These metrics identify the differences in contour position, shape, size or orientation. Moreover, dosimetrical endpoints are used for the accuracy evaluation such as mean and maximum dose. Qualitative analysis may also be used to assess ABAS performance.

For geometrical comparison, let us consider two segmentations, a reference and ABAS segmentation. The reference segmentation is represented by a volume V_R and a center coordinate (x,y,z). These contours could be manual segmentations contoured by an expert observer, or a representative shape based on several manual segmentations contoured by different observers [55]. The ABAS segmentation is the one that is segmented automatically. It is represented by a volume V_{ABAS} and a center coordinate (x',y',z'). The most commonly used **geometrical indices** (Figure 2.7) are:

a) Center of mass shift (CMS)

This index focuses on the location of the structures' centers. It indicates the

volume displacement in 3D. It is calculated as:

$$\Delta x = |x - x'|, \quad \Delta y = |y - y'|, \quad \Delta z = |z - z'| \qquad (2.3)$$

b) Percentage of Volume differences

This is defined as the absolute volume difference between the segmentations multiplied by 100 and divided by the reference segmentation.

$$\Delta V\% = \frac{|V_{reference} - V_{ABAS}|}{V_{reference}} \times 100 \qquad (2.4)$$

c) Coefficient of Variance (CV)

CV is a quantity used to compare the spread of data sets. High CVs represent large dispersion in the variable. CV is calculated as the ratio of the standard deviation to the average data. It could be used for geometric and dosimetric analysis by calculating volume CV and dose CV, respectively.

d) Similarity coefficients

The aim of similarity coefficients is to provide a quantitative evaluation of the amount of overlap between the two volumes. The most commonly used coefficients in this field are:

- Jacquard coefficient:

This is defined as the ratio of the overlap volume to the union volume of two delineations.

$$J = \frac{V_R \cap V_{ABAS}}{V_R \cup V_{ABAS}} \qquad (2.5)$$

- Dice similarity coefficient (DSC):

This is defined as the ratio between twice the overlap volume to the sum of volumes [52].

$$DICE = \frac{2(V_R \cap V_{ABAS})}{V_R + V_{ABAS}} \qquad (2.6)$$

- Conformity level (CL):

This coefficient is a generalization of the Jaccard coefficient. It is calculated as the ratio of the commonly segmented volume by all observers to the encompassing volume. It is used to compare atlas delineation with other observers. It is unbiased with respect to the number of contours and it equals the Jaccard coefficient in the case of two segmentations.

$$CL = \frac{V_A \cap V_B \cap V_C \cap V_{ABAS}}{V_A \cup V_B \cup V_C \cup V_{ABAS}} \qquad (2.7)$$

Some other terminologies are used to indicate different similarity coefficients, which have the same goal, such as conformity index [53], concordance index [54], and percent volume overlap [56]. In general, these indices are scalars with a value between 0 and 1. A value of 0 means full disjoint volumes, whereas a value of 1 indicates identical delineations.

e) Hausdorff distance (HD)

This index describes the maximum distance of all the distances from a point in one contour set to the closest point in the other set. However, this index is sensitive to small regions of poor segmentation [42].

Comparing the result with inter-observer variation is another common approach for ABAS performance assessment. Inter-observer variation depends on the structure volume, location, visibility, and contouring protocol. Commonly, this variability is evaluated using one of the similarity indices and the Hausdorff distance.

Although these indices are used widely for geometrical quantitative analysis, the discrepancy may not affect the clinical impact. Moreover, the different indices complement each other and may not correlate. The lack of correlation between the DSC coefficient and the Hausdorff distance for various parameter settings of atlas-based auto-segmentation algorithm was shown by Sharp G. et al. [42].

Evaluating some **dosimetrical endpoints**, such as mean dose for parallel structures and maximum dose for serial structures, is used to assess the clinical implementation of ABAS. One way to define these end points is by superimposing the dose distribution of the clinically used plan on the ABAS structures. Another way is to start an optimization process using ABAS contours and compare the dose distribution and some specific dose endpoints with the one that optimized on the reference structures.

Qualitative evaluation is an important analysis, as it demonstrates the clinical acceptability and efficiency of the generated contours. One way of performing this

evaluation is by classifying the degree of agreement of each structure into categories ranged from 'good agreement', minor to major editing required, to 'not acceptable'. This methodology quantifies the frequency at which ABAS is able to achieve clinically acceptable segmentations; therefore, this approach is considered as a suitable measure of ABAS clinical performance [57].

To summarize, it has been mentioned early in this chapter that the multi-atlas approach has a superior performance among other approaches. It was also stated that deformable image registration between atlas and test images is required because of significant variations between the individuals. However, a larger deformation requires more time to obtain accurate results. In fact, a smaller magnitude of the deformation typically results in a higher matching accuracy. In order to achieve higher accuracy with minimum deformation, one way is to construct an atlas library with a large number of individuals that cover a larger population. This will enable more ease in finding the closest matches to the test case with a smaller computational cost due to the smaller magnitude of required deformation. ABAS accuracy evaluation is not a straightforward task.

Due to the absence of the gold standard and patient positioning uncertainties and anatomical changes during the treatment, plans containing small contour deviations may still produce similar dose volume results in a treatment plan. As a result, the use of ABAS contours, which may be considered marginally unacceptable by a physician, may not have a significant negative influence on treatment efficacy

This chapter has introduced auto-segmentation concepts and segmentation evaluation metrics. In the following chapters, studies are presented that make use of the MIM Maestro™ ABAS tool to segment organs at risk in head and neck radiotherapy. The metrics introduced here are used to geometrically and dosimetrically evaluate the results and make comparisons with the results of manual segmentation processes. Also, we used these indices to compare our results with other studies.

Figure 2. 1 automatic segmentation approaches in radiotherapy field

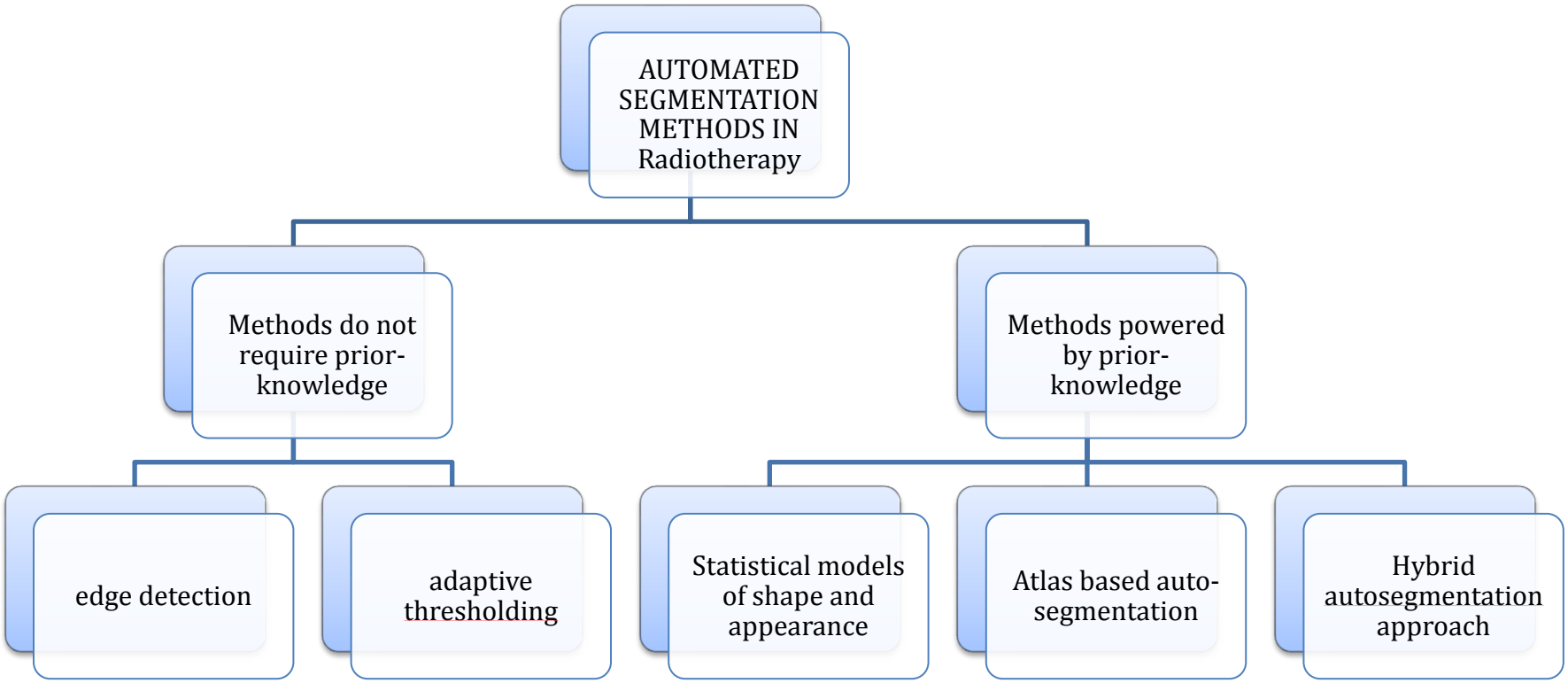


Figure 2. 2 Schematic illustration of the single subject atlas approach.

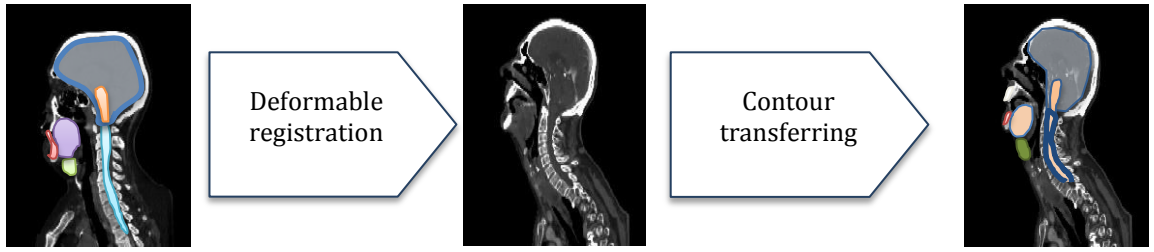


Figure 2. 3 Schematic illustration of the average atlas approach

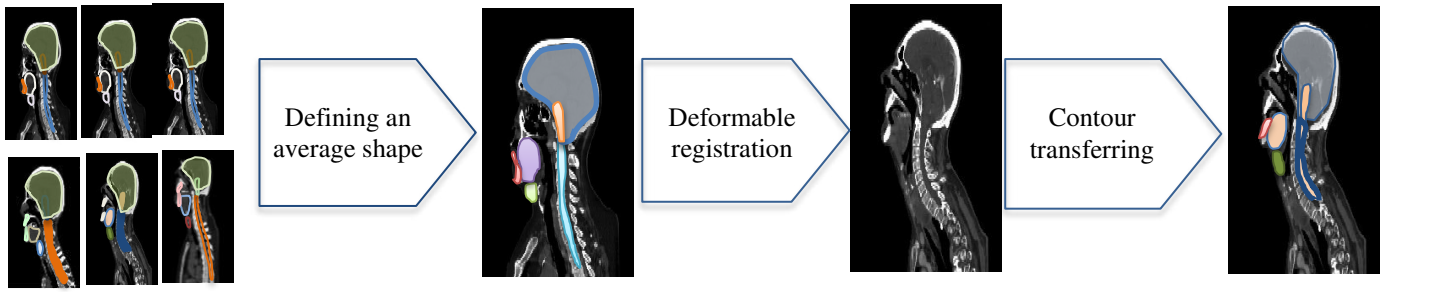


Figure 2. 4 Atlas library construction

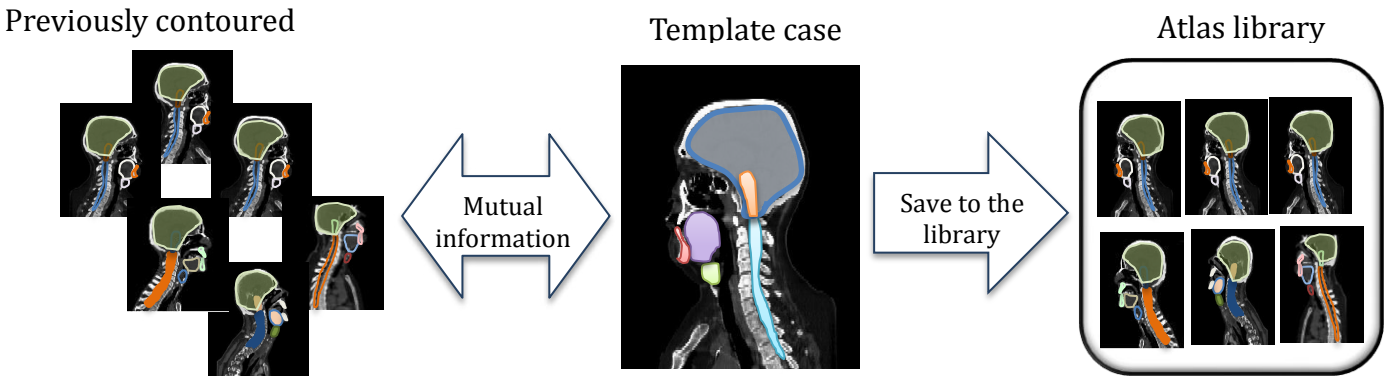


Figure 2. 5 Schematic illustration of the single closest match atlas approach

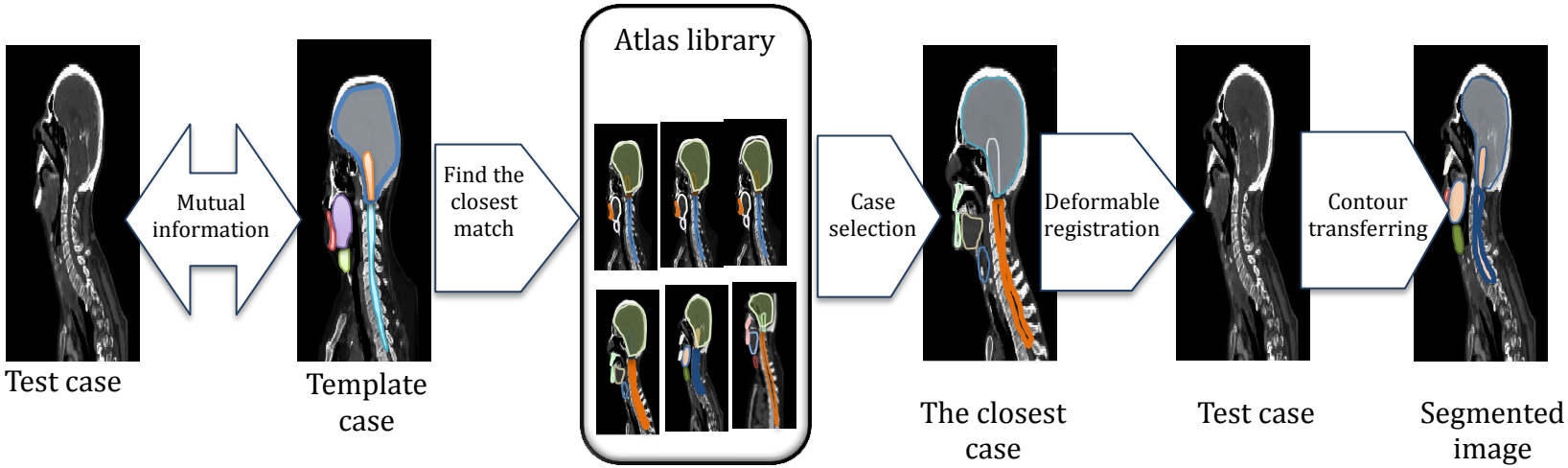


Figure 2. 6 Schematic illustration of the multiple closest matches atlas approach

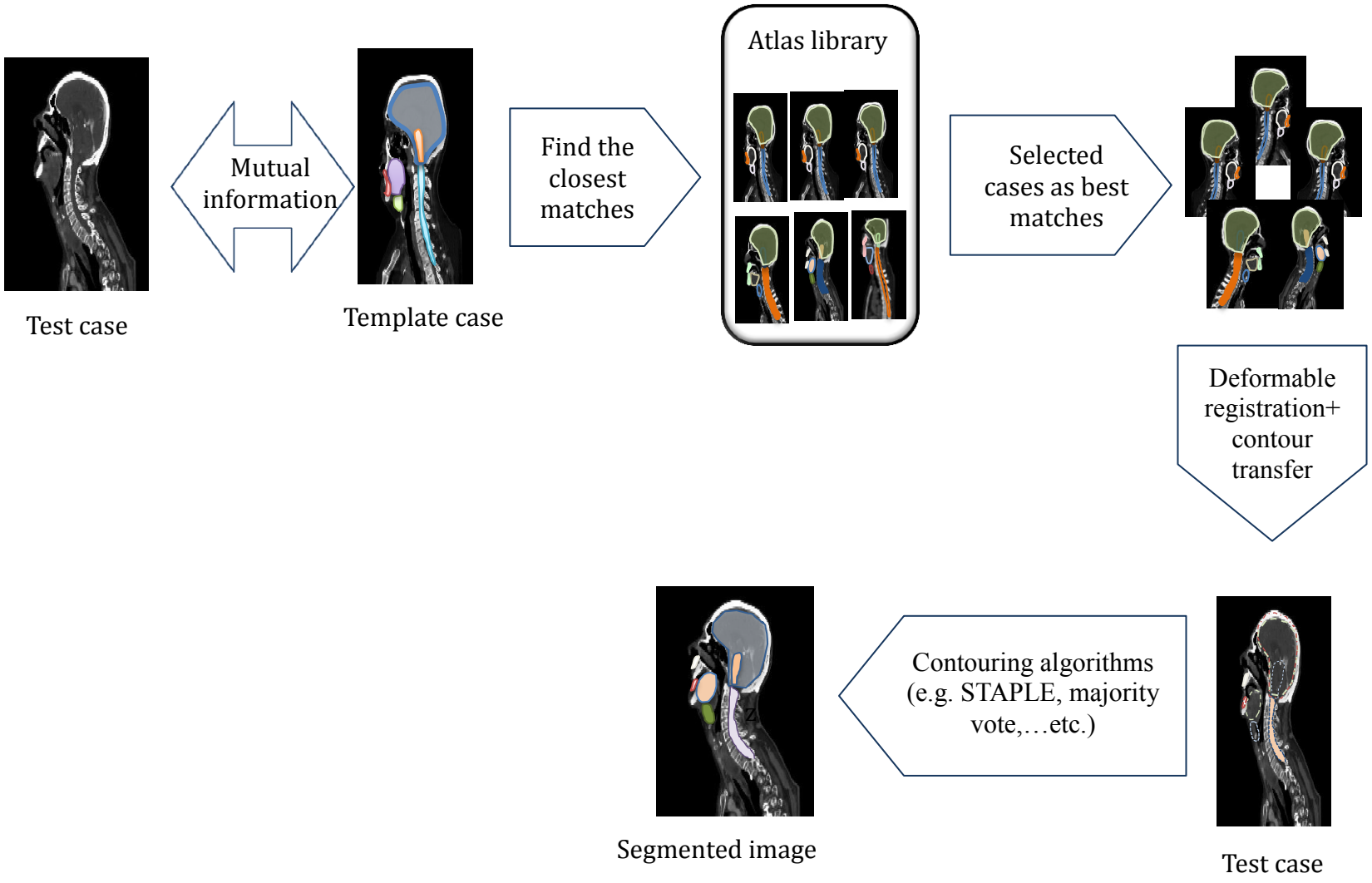
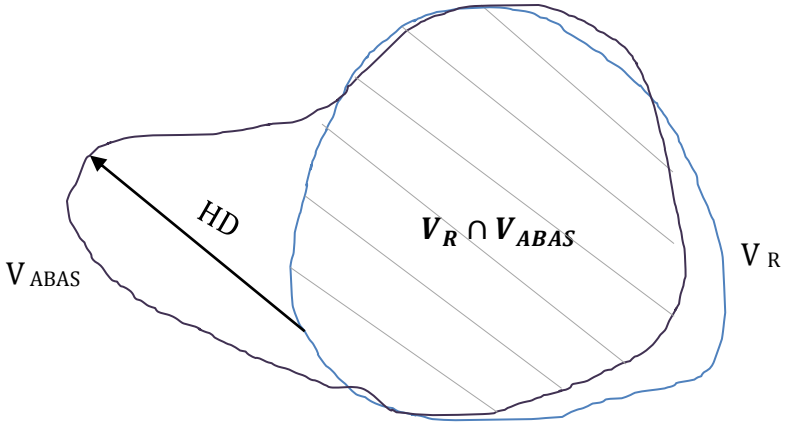


Figure 2. 7 Schematic illustration of the geometrical indices

V_R and V_{ABAS} represent the reference (blue) and the ABAS (black) volumes, $\boldsymbol{V_R \cap V_{ABAS}}$ represents the common area between V_R and V_{ABAS} and HD represents the maximum distance between V_R and V_{ABAS}.



CHAPTER 3: ASSESSMENT OF VARIABILITY IN MANUAL SEGMENTATION

3.1 INTRODUCTION

As mentioned in Chapter 1, accurate definition of target volumes and OARs is required to exploit the benefits of VMAT. However, manual segmentation, which is routinely performed, is a time-consuming task. It is also subject to intra- and inter-observer variability. The magnitude of this variation can be influenced by various factors such as the images used for contouring (image modality, resolution, contrast), human factors (individual experience, professional background), and the utilization of the delineation guidelines [17,19,20]. The study of intra- and inter-observer variation is necessary; it allows the quantification of segmentation uncertainties within an institutional delineation protocol, validation of contouring guidelines and assessment of the performance of auto-segmentation tools [18, 19, 24, 52, 53]. It is stated in several publications that organs at risk (OARs) delineation uncertainties are considered as one of the potential sources for uncertainties in dose-volume histogram data and therefore reduced performance of predictive outcome models [18, 24, 35]. Qazi et al. [44] reported acceptable accuracy for head and neck target and OAR auto-segmentation tools within a shorter delineation time compared to manual segmentation but also mentioned the need for multi-center and multi-observer studies to provide more insight in the robustness and reliability of the automated approach.

The aim of this study is to investigate the significance of head and neck OAR intra- and inter-observer variation in our institution and to quantify its impact on certain dosimetric endpoints, examine its consistency with other studies and consequently, establish benchmark data that could be used to evaluate the performance of our in-house head and neck atlas-based auto-segmentation tool.

3.2 MATERIALS AND METHODS

3.2.1 Case selection and Image data sets

This study made use of eight head and neck cancer treatment plans. All eight plans were for patients diagnosed with NasoPharyngeal Carcinoma (NPC). Each patient had been previously scanned in a supine head first position and contrast enhanced CT images were acquired on a CT scanner at 120 kV energy with a 2.5 mm slice thickness and 1.17x1.17 mm pixel size. All the cases had completed a VMAT treatment course. Treatment planning was done using Eclipse ™ AAA calculation algorithm version 11.0.31. Patient diagnosis and dose prescriptions for these eight cases are shown in Table 3.1. For the purpose of this study, each CT data set was anonymized using the MIM ™ anonymization tool.

3.2.2 Participating observers

At BCCA, Vancouver Centre, clinical dosimetrists are assigned to delineate OARs. Three dosimetrists participated in this study. These dosimetrists were trained under the supervision of an expert observer following institutional protocols. Each participating dosimetrist had significant experience with Eclipse™ software tools while MIM™ software was in the process of being implemented into clinical practice. However, training sessions on MIM™ software were provided prior this study enabling participating staff to make use of the manual segmentation tools available in MIM™.

3.2.3 OAR delineation

For each one of the test cases, each observer delineated a complete head and neck contour set consisting of fifteen structures on both the Eclipse™ version 10.6 treatment planning system (Varian Medical Systems, Palo Alto, CA, USA) and the MIM Maestro V 6.4 software (MIM Software Inc. Cleveland, OH, USA). The segmentations were performed on the axial CT images. Each structure set included whole brain, brainstem, spinal cord, eyes, optic nerves, parotids, submandibular glands, oral cavity, laryngopharynx, mandible, and lips. The original contour set used for treatment planning was not presented to the observers. Therefore, the observers were completely blinded from the original and other observers' contour sets. The time required to perform delineation of each complete contour set was recorded. The observers were protected from any interruption during the contouring process

3.2.4 Quantitative analysis of intra- and inter-observer variation

A number of indices were used to quantify intra- and inter-observer variation for each structure. Volume variation assessment, DICE, and HD were used for geometric evaluation whereas dosimetric endpoints of either mean organ dose or maximum organ dose were chosen for parallel or serial organs at risk, respectively.

Volume variation:

1) **Average ± SD** in Volume, V(cc):

   This test looked at the variation in segmented organ volume over the three observers and two software platforms. The average and SD of the volume over all observers was calculated for each organ. Because organ volume varies from patient to patient, overall volume variation does not distinguish between patient to patient and inter-observer variation. Eclipse and MIM results were generated individually, to determine if there was any bias introduced into the manual segmentation results due to the choice of software and to gain some insight into the intra-observer variation.

2) Relative volume variation on an organ by organ basis **(ΔV%)**

   a. The impact of inter-observer variation on organ volume was quantified by looking at the difference in volume, ΔV, among the 3 observers, for each organ. The steps involved were as follows, for each clinical case:

i. The maximum (V_{max}), minimum (V_{min}), and average ($V_{Average}$) absolute volumes over the three observers segmentations were calculated for each case.

ii. ΔV, was defined as $V_{max} - V_{min}$

iii. The $\Delta V\%$ was calculated $\Delta V \times 100 / V_{Average}$

iv. The overall average $\Delta V\%_{ave}$ was calculated over the eight results

b. The impact of intra-observer variation was measured for each observer for each case by comparing their Eclipse contour to the MIM contour. This metric may also include some variation due to the use of two different software platforms. The analysis was performed separately for each observer. The steps performed in this calculation were as follow, for each clinical case:

i. For each observer,

ii. $\Delta V = |(V_{Eclipse}) - (V_{MIM})|$, and

iii. $V_{Average} = ((V_{Eclipse}) + (V_{MIM})/2,)$ were calculated for each case and each organ.

iv. $\Delta V\%$ was calculated as $\Delta V \times 100 / V_{Average}$ (resulting in 8 values per organ per observer).

v. Average $\Delta V\%_{ave}$ was then calculated over the eight results for each observer

3) Coefficient of variance (**CV**) of the Volume

CV is a quantity used to compare the spread of data sets. High CVs represent large

the dispersion in the variable. Inter-observer CV was calculated as the ratio of the SD in volume to the average volume across the three observers. Intra-observer CV was calculated as the ratio of the SD in volume to the average volume across the same observer. Ultimately, each organ had eight CV values. The results represent the average ±SD of CV over the eight cases for each organ. CV was calculated to allow direct comparison with other studies.

4) Degree of contour overlap using **DICE**;
    a. Inter-observer variation: using Eclipse data, DSC was measured according to equation (2.6) for each pair of the three observers over 8 test cases. The steps of finding the result were as follow, for each clinical case:
        i. DSC for observer A and B was calculated.
        ii. DSC for observer B and C was calculated.
        iii. DSC for observer C and A was calculated.
        iv. Each organ resulted in 24 DSC pairs (3 DSC pairs × 8 test cases)
        v. The average and SD of DSC over all the 24 results were calculated.
    b. Intra-observer variation: DSC was measured for each one of the three observers over 8 test cases, comparing their Eclipse contour to the MIM contour. The steps followed to determine the results were as follows, for each clinical case:
        i. For each observer separately, DSC was calculated for the Eclipse contour compared with the MIM contour.
        ii. Each organ resulted in 8 DSC results for each observer.

iii. The average and SD DSC over all the 8 results were calculated for each observer.

iv. These steps were repeated for each organ.

5) **Hausdorff distance** (**HD**):

Using the same process used for the DSC calculations, HD was calculated for inter- and intra-observer variation.

All the geometric and dosimetric indices were measured and exported using MIM Maestro™ statistical analysis tool.

Dosimetric evaluation:

Geometric variation quantified by the organ volumes and similarity metrics would not, by itself, indicate whether such differences could be clinically relevant. Therefore, the dosimetric impact due to the OAR variation was also quantified.

In order to do this, for each structure set, the dose distribution of the original treatment plan was superimposed onto all contours to determine dose–volume parameters.

The maximum dose to brainstem, spinal cord and optic nerves (the serial organs) and the mean dose of the other structures (parallel organs) were chosen as dosimetric endpoints. Note that to simplify notation, the symbol D is used to represent either mean or maximum organ dose depending on the organ. The dose for observers A. B and C on MIM and Eclipse ($D_{A,MIM}$, $D_{A,Eclipse}$, $D_{B,MIM}$, $D_{B,Eclipse}$, $D_{C,MIM}$, and $D_{C,Eclipse}$) was calculated. Differences in organ dose ΔD, between the original treatment and those obtained using the study structures sets were then assessed. Both the SD in dose difference ΔD and

coefficient of variance of Dose (CV) for each dose endpoint were measured for each OAR.

6) The dose variation **ΔD SD (Gy):**

The dose variation ΔD (Gy) was obtained as the difference between the reference contour organ dose and organ dose associated with each study segmentation. The steps used to calculate the SD of dose variation for intra and inter-observer variation were as follows:

i. Inter-observer dose variation: using Eclipse segmentations, ΔD was measured separately for $D_{A,Eclipse}$, $D_{B,Eclipse}$ and $D_{C,Eclipse}$ over 8 test cases.
ii. The SD of ΔD among all over 24 results (3 segmentations × 8 test cases) represents SD of inter-observer dose variation using Eclipse software.
iii. The same process repeated for MIM data using $D_{A,MIM}$, $D_{B,MIM}$ and $D_{C,MIM}$
iv. Intra-observer dose variation: using Eclipse and MIM segmentations for observer A, ΔD was measured for $D_{A,Eclipse}$, and $D_{A,MIM}$ for each case per organ individually.
v. The SD of ΔD among all over 16 results (2 segmentations × 8 test cases) represents SD of intra-observer dose variation for observer A.
vi. The same process repeated for observers B and C using $D_{B,MIM,}$ $D_{B,Eclipse,}$ $D_{C,Eclipse}$ and $D_{C,MIM.}$
vii. The maximum and minimum SD among the 6 SDs (Eclipse, MIM, observer A, B and C) per organ were identified

7) Coefficient of variance **(CV)** in Dose:

Inter-observer CV was measured as the ratio of the SD in dose to the average dose across the three observers where as Intra-observer CV was calculated as the ratio of the dose SD to the average dose across the same observer. Each organ ultimately had eight CV dose values. The results represent the average ± SD of CV over the eight values.

Segmentation Time:

Using a stopwatch, each observer measured the time required to perform manual segmentation for each organ for each of the eight cases.

3.3 RESULTS AND ANALYSIS

Figures 3.1 and Figure 3.2 represent a typical segmented case, showing contours for the 15 OARs displayed on an axial, sagittal and coronal plane. Qualitative observation indicates that much of the variation among the observers occurs at the inferior or superior border of the structures, indicating a decision to include or exclude the uppermost or lowermost CT slice in the structure. Quantitative volumetric and dosimetric and time results and analysis are described in sections 3.3.1, 3.3.2 and 3.3.3 respectively.

3.3.1 OAR volume variation assessment

Table 3.2 summarizes the average and standard deviation of the structure volume on an organ by organ basis, over all observers, for the 15 commonly delineated OARs for NPC patients. This data reflects both patient-to-patient variation in organ volume as

well as inter- and intra- observer variation. Volumes delineated using Eclipse and MIM were not statistically different as indicated by p-values, which were calculated using the t-test. Inter-observer volume variation is systematically larger compared with the intra-observer/inter-platform volume variation.

Figure 3.3 provides a graphical representation of the relative differences in organ volume ($\mathbf{\Delta}$ $V\%_{ave}$) over all observers, for all OARs. Table 3.3 presents the coefficient of variation (CV) for these structure volumes. Both inter- and intra- observer data are shown. This data isolates the observer performance from the patient-to-patient organ volume variation. The structures are displayed in order of the least variation in observer performance to the most variation in observer performance. The volume of the whole brain, eyes and mandible proved to be the most consistent CV < 10%. The CV for brainstem, oral cavity, parotid glands, submandibular glands and spinal cord volumes was between 10% and 20%. Volumes of the laryngopharynx, optic nerves and lips were the most variable with CV > 20%.

Figure 3. 4 shows the inter- and intra- observer variation in the DSC coefficient for the 15 OARs, indicating the degree of overlap of contours across observers. Intra- and inter- observer performance agreed within 1 SD. The average DSC was greater than 0.8 for 11 out of 15 OARs. Consistent with the CV data, observer performance was least consistent for the laryngopharynx, lips and optic nerves, with average DSC of 0.76±0.02 vs.0.78±0.07, 0.69±0.05 vs. 0.70±0.06 and 0.61±0.01 vs. 0.62±0.01 for inter- and intra-observer, respectively.

Figure 3.5 displays the inter- and intra-observer Hausdorff Distance for all OAR's. Except for whole brain and spinal cord, inter-observer variation was systematically equal to or larger than intra-observer variation. HD <1cm for eyes, brainstem and submandibular glands. The spinal cord showed the highest variation in HD compared to the other structures. The whole brain indicated the largest HD, perhaps not a complete surprise, as it is also the largest structure.

3.3.2 The dosimetric impact of inter- and intra-observer variation:

The dosimetric impact of differences in OAR segmentation are presented in Figure 3.6 and Table 3.4. The two dosimetric endpoints, maximum OAR dose (D_{max}) and mean OAR dose (D_{mean}), are shown for serial and parallel organs respectively. Figure 3.6 shows the standard deviation of inter- and intra-observer differences of ΔD (Gy), while Table 3.4 shows the coefficient of variation of the OAR doses. The following observations are of particular importance:

(1) From Figure 3.6, the variation in inter- and intra-observer dose differences for the parotid glands, brainstem and spinal cord are notably higher than the other structures.

(2) Spinal cord showed the highest standard deviation in ΔD_{max} (up to 4 Gy) while Parotid gland showed the highest standard deviation in ΔD_{mean} (3.54 Gy).

3) From Table 3.4, the coefficient of variation, CV, in the mean or maximum dose for all structures was < 5% for all except the optic nerves and the parotid glands. The average CV in parotid gland mean dose was 6%. The variation in maximum dose for the optic nerves was up to 20%.

3.3.3 Segmentation time

Manual segmentation time on an organ by organ basis is shown in Table 3.5. Although the observers had significant experience with Eclipse software tools whereas the MIM software had not yet been clinically implemented, it took 29.73±0.73 minutes for each observer to complete full segmentation set regardless the software used for segmentation. It is interesting to note the differences in time required to segment different structures. There is a trend to longer times for the larger volume structures.

3.4 DISCUSSION AND CONCLUSION:

Numerous studies have confirmed substantial inter- and intra- observer variation in OAR delineation in the head and neck, as well as other sites. Quantitative comparisons may be made between the data in this study and data from three other studies reporting on similar geometric and dosimetric metrics. Brouwer et al. [22] investigated OAR sub-regions showing inter-observer variability across five radiation oncologists at the same institution delineating twelve study cases. The authors focused on the spinal cord, parotids, submandibular glands, thyroid cartilage and glottic larynx. In another study, Nelms et al. [24] examined inter-observer variation among thirty-two observers from multiple institutions in the delineation of a single CT dataset for an oropharyngeal patient. The examined OARs were brainstem, parotid glands, spinal cord, mandible and brain. Tao et al. [36] performed a multi-institution study to assess whether manually edited multi-subject ABAS can reduce inter-observer variation and improve dosimetric

consistency for OARs in NPC. That study examined sixteen NPC cases, each one contoured by eight radiation oncologists from eight different institutions.

Our geometric variation results were consistent with the other studies. Comparison of mean volume variation, coefficients of variation in volume and DSC are shown in Tables 3.6, 3.7 and 3.8. DSC coefficients are not statistically different among the studies and range from 0.6 to 0.99 depending on the OAR. A commonly cited benchmark for acceptable geometric performance in auto-segmentation is a DSC similarity coefficient of 0.8 [58]. It is apparent from the data in Table 3.6 that this is not achievable for all OARs, even for the best observers performing manual segmentation. Submandibular glands and optic nerves showed DSC < 0.8 in our study and the study by Tao et al. [36]. This suggests that an OAR specific DSC benchmark should be used when evaluating the performance of segmentation algorithms or manual observers.

Mean OAR volume is reported by Nelms et al. [24] and Brower et al. [22] in addition to our study. The whole brain data in the study by Nelms et al. [24] is significantly different from our study. This is due to the fact that the Nelms et al [24] CT data set did not capture the superior extent of the whole brain, likely because they used an oropharynx case whereas our study used nasopharynx cases. The oropharynx is located more inferiorly than nasopharynx and thus the full brain is not generally considered an OAR for oropharynx cancer cases. The spinal cord data from Brower et al. [22] is also significantly smaller than the other two studies. This may be due to the superior or inferior extent required of their institutional protocol. It should be pointed

out that the data in Table 3.7 for our study and Brouwer et al. [22] reflects patient-to-patient organ volume variation as well as inter-observer variation, whereas the Nelms et al [24] study used only one case so their data reflects only inter-observer variation across several institutions.

The coefficients of variation of OAR volume in Table 3.8 show that our data agrees with the Tao et al [36] data within the margin of error for all reported OARs. The Nelms et al [24] data showed systematically higher coefficients of variation in OAR volumes compared with the other authors. This is interesting because they looked at a single case with 32 observers from different institutions, whereas all others studies looked at a minimum of 8 cases.

There is no comparative data on HD in any of the other studies. An HD value of 1.0 cm is often used as a benchmark to validate the geometric performance of auto-segmentation tools. Referring back to Figure 3.5, it is noted that our data indicates that an HD of 1.0 cm is achieved for only five of the fifteen OARs, the eyes, the brainstem and the submandibular glands. Given that the HD represents the largest distance between two contours, it could be a significant indicator of differences in maximum organ dose, if the discrepancy is in a high dose gradient region. This could be more important for serial structures such as brainstem, spinal cord and optic nerves in close proximity to the PTV.

Although OAR definitions may vary from observer to another, and for the same observer on repeated attempts, one should quantify whether this variation has an effect on the dose received, or reported to be received, by each organ. In particular, the proximity of the organ to the high dose volume and to large dose gradients could have a large impact on maximum organ dose. Mean organ dose may be less sensitive to OAR segmentation than maximum dose.

Comparison of dosimetric variation in our study with that of Tao et al. [36] is shown in Table 3.9. The variation in the dose ranges from 1% to 20% depending on the OAR and there is no statistical difference between the results in the two studies. The optic nerves showed the largest coefficient of variation while the submandibular glands showed the smallest coefficient of variation.

The consistency of our results with other studies provides broader validation of inter-observer variation in manual segmentation for OAR's in the head and neck region. This data may be used with confidence to set performance benchmarks for auto-segmentation tools. Table 3.10 shows a summary or recommended performance benchmarks for geometric and dosimetric assessment of organ segmentation based on the results in Figure 3.3-3.6 in this chapter.

Table 3. 1 Patient diagnosis and dosimetric parameters

Case	Diagnosis	Prescribed dose for each PTV (Gy)	PTV size (PTV1/ PTV2/ PTV 3) (cc)	VMAT start angle	VMAT stop angle	Rotation direction	# Arcs	Collimator rotation	MUs
1	Nasopharynx	70/63/56	70/ 85.5/ 396.94	179	181	CCW	1	30	566
2	Nasopharynx	70/60/56	75.69/ 8.4/ 545.48	179/ 181	181/ 179	CCW/CW	2	30/330	495
3	Nasopharynx	70/56	316/ 606	179	181	CCW	1	30	529
4	Nasopharynx	66/54	255.7/ 741	179	181	CCW	1	30	378
5	Nasopharynx	70/63/56	84.5/ 69.6/ 421.4	179	181	CCW	1	30	519
6	Nasopharynx	70/60/56	616.5/ 13.3/ 127.7	179/ 181	181/ 179	CCW/CW	2	30/330	505
7	Nasopharynx	70/63/56	63.6/ 15.9/ 537.78	179/ 181	181/ 179	CCW/CW	2	30/330	473
8	Nasopharynx	70/63/56	67.1/12.9/ 604	179	181	CCW	1	30	390

Table 3. 2 Average ±SD of the delineated volumes (cc) over eight test cases per organ, Eclipse and MIM represent inter-observer variation, Observer A,B and C represent intra-observer variation

Structure	ECLIPSE (3 observers)	MIM (3 observers)	p-value MIM vs. Eclipse	Observer A (Eclipse vs. MIM)	Observer B (Eclipse vs. MIM)	Observer C (Eclipse vs. MIM)	Average p-value (A vs. B, B vs. C, C vs. A)	Average inter-observer	Average intra-observer
Optic nerves	0.75±0.21	0.80±0.21	0.35	0.63±0.12	0.81±0.13	0.89±0.14	0.13	0.78 ± 0.14	0.58 ± 0.09
Submandibular glands	7.35±1.29	7.50±1.50	0.83	7.12±0.78	7.13±1.63	8.03±1.06	0.32	7.42 ± 0.93	5.57 ± 0.87
Eyes	8.68±0.56	8.82±0.73	0.69	8.79±0.73	8.46±0.36	9.00±0.52	0.46	8.75 ± 0.43	6.56 ± 0.40
Spinal cord	23.32 ±4.56	22.20 ± 4.71	0.57	18.62 ±1.38	24.01 ±2.42	25.64 ±2.71	0.53	22.76 ± 3.09	17.07 ± 1.63
Brainstem	25.61 ±2.45	24.91 ± 3.34	0.69	23.58 ±1.22	26.24 ±2.01	25.95 ±2.13	0.76	25.26 ± 1.93	18.95 ± 1.34
Lips	27.35±8.15	30.28±7.34	0.52	22.42±6.30	30.40±5.60	33.63±4.24	0.55	28.82 ± 5.16	21.61 ± 4.03
Parotid glands	31.29±3.69	29.15±4.76	0.54	27.37±2.39	31.40±3.35	31.88±2.91	0.60	30.22 ± 2.82	22.66 ± 2.17
Laryngopharynx	35.91±8.88	38.08±10.31	0.66	43.36±5.74	29.70±7.94	37.92±4.31	0.67	36.99 ± 6.04	27.74 ± 4.50
Mandible	73.43±4.52	72.28±6.65	0.88	70.50±8.09	75.95±3.87	72.12±1.94	0.82	72.86 ± 3.73	54.64 ± 3.47
Oral cavity	85.75±11.09	87.83±12.22	0.77	98.54±5.41	80.32±7.29	81.52±6.40	0.44	86.79 ± 7.77	65.09 ± 4.78
Whole Brain	1369.25 ± 189.75	1386.41 ± 189.75	0.86	1366.91 ± 190.3	1381.02 ± 193.83	1385.56 ± 191.56	0.87	1377.83 ± 189.79	1033.37 ± 191.92

Table 3. 3 Average ± SD of volume coefficient of variance over eight test cases

Structure	Inter-observer CV	Intra-observer CV
Whole Brain	1% ± 0%	1% ± 1%
Eyes	7% ± 2%	6% ± 4%
Mandible	8% ±1%	6% ± 5%
Brainstem	11% ± 1%	7% ±3%
Oral cavity	14% ± 2%	8% ± 9%
Parotid glands	15% ±3%	10% ± 5%
Submandibular glands	19% ± 1%	16% ± 10%
Spinal cord	20% ± 2%	10% ± 9%
Laryngopharynx	27% ± 1%	16% ± 4%
Optic nerves	28% ± 2%	17% ± 5%
Lips	28% ± 2%	18% ± 11%

Table 3. 4 Average ± SD of dose coefficient of variance over eight test cases

		Inter-observer CV		Intra-observer CV	
Structure	Endpoint	Average	SD	Average	SD
Brainstem	$D_{max.}$	2%	3%	2%	2%
Spinal cord		3%	5%	2%	4%
LT Optic nerve		18%	12%	7%	9%
RT Optic nerve		20%	13%	10%	13%
RT Submandibular gland	D_{mean}	1%	1%	1%	1%
LT Submandibular gland		1%	1%	1%	1%
Laryngopharynx		2%	1%	1%	1%
Mandible		2%	2%	1%	2%
Oral cavity		3%	2%	2%	2%
Whole Brain		3%	4%	2%	5%
RT Eye		3%	2%	3%	2%
LT Eye		4%	4%	3%	4%
Lips		5%	3%	4%	2%
RT Parotid gland		5%	4%	5%	4%
LT Parotid gland		6%	6%	4%	8%

Table 3. 5 Average and SD of manual segmentation time on an organ-by-organ basis among the three observers

Structure	Average time (min)	SD (min)
Body	4.03	1.15
Brain	3.52	0.30
Brainstem	1.90	0.55
Cord	2.17	0.35
LT eye	0.80	0.05
RT eye	0.70	0.08
LT Optic nerve	0.73	0.25
RT Optic nerve	0.57	0.20
LT Parotid gland	2.37	0.27
RT Parotid gland	2.07	0.18
LT Submandibular gland	1.43	0.13
RT Submandibular gland	1.25	0.37
Laryngopharynx	1.78	0.37
Lips	1.70	0.10
Mandible	3.13	0.48
Oral Cavity	1.50	0.13
Total segmentation time / case (min)	29.73	0.73

Table 3. 6 comparison of DSC result with other studies

Structure	Khawandanh[a]	Tao [40][b]	Nelms et al. [27][c]
Whole Brain	0.99 ± 0.00	N/A	0.98 ± 0.01
Eyes	0.91 ± 0.00	0.89 ± 0.01	N/A
Mandible	0.86 ± 0.03	0.89 ± 0.02	0.87 ± 0.07
Oral cavity	0.83 ± 0.00	0.81 ± 0.04	N/A
Brainstem	0.86 ± 0.01	0.83 ± 0.03	0.66 ± 0.17
Spinal cord	0.80 ± 0.02	0.77 ± 0.04	0.8 ± 0.07
Parotid glands	0.84 ± 0.02	0.83 ± 0.03	0.77 ± 0.09
Submandibular glands	0.78 ± 0.01	0.77 ± 0.06	N/A
Optic nerves	0.61 ± 0.01	0.57± 0.09	N/A

a) 1 institution 3 observers, 8 cases

b) 8 institutions, 8 observers, 16 cases

c) 32 institutions, 32 observers, 1 case

Table 3. 7 comparison of mean volume result with other studies

Structure	Khawandanh[a]	Nelms et al. [24][b]	Brouwer et al [22][c]
Spinal cord	22.76 ± 4.64 cc	27.54 ± 5.51 cc	17.6 ± 1.1 cc
Brainstem	25.26 ± 2.89 cc	25.57 ± 10.59 cc	N/A
Submandibular glands	7.42 ± 0.93 cc	N/A	10.5 ± 0.9 cc
Parotid glands	30.22 ± 4.22 cc	23.445 ± 7.295 cc	29 ± 3.3 cc
Mandible	72.86 ± 5.59 cc	67.67 ± 10.56 cc	N/A
Whole Brain	1377.83 ± 12.77 cc	802.83 ± 22.56 cc	N/A

a) 1 institution 3 observers, 8 cases

b) 32 institutions, 32 observers, 1 case

c) 1 institution, 5 observers ,12 cases

Table 3. 8 comparison of volume CV result with other studies

Structure	Khawandanh [a]	Tao et al. [36][b]	Nelms et al. [24][c]	Brouwer et al [22][d]
Whole Brain	1% ± 0%	N/A	3%	N/A
Eye	7% ± 2%	9% ± 3%	N/A	N/A
Mandible	8% ±1%	11% ± 2%	16%	N/A
Brainstem	11% ± 1%	12% ± 4%	41%	N/A
Oral cavity	14% ± 2%	11% ± 4%	N/A	N/A
Parotid glands	15% ±3%	13% ± 5%	31%	14%
Submandibular glands	19% ± 1%	20% ± 12%	N/A	16 %
Spinal cord	20% ± 2%	26% ± 4%	20%	16%
Optic nerve	28% ± 2%	66% ± 20%	N/A	N/A

a) 1 institution 3 observers, 8 cases

b) 8 institutions, 8 observers, 16 cases

c) 32 institutions, 32 observers, 1 case

d) 1 institution, 5 observers ,12 cases

Table 3. 9 comparison of Dose CV result with other studies

Structure	Endpoint	Khawandanh	Tao et al. [36]
RT Optic nerve	D_{max}	20%±13%	19%±18%
LT Optic nerve		18%±12%	17%±16%
Brainstem		2%±3%	4%±2%
Spinal cord		3%±5%	7%±4%
RT Parotid	D_{mean}	5%±4%	3%±1%
LT Parotid		6%±6%	4%±1%
Oral cavity		3%±2%	3%±1%
RT Eye		3%±2%	4%±2%
LT Eye		4%±4%	5%±3%
RT Submandibular glands		1%±1%	1%±1%
LT Submandibular glands		1%±1%	2%±1%

Table 3. 10 benchmark data extracted from inter- and intra-observer variation. This data will be used for auto-segmentation tool evaluation

	DICE		HD (cm)		ΔV (%) Maximum volume difference between observers		SD of ΔD (Gy) Maximum variation in dose	
Structure	Average	SD	Average	SD	Maximum ΔV(%)	SD	Dosimetric endpoint	SD
Whole Brain	0.99	0.01	1.45	0.36	2%	2%	D mean	0.98
Eyes	0.91	0.01	0.48	0.11	15%	5%	D mean	0.36
Mandible	0.86	0.06	1.34	0.10	15%	18%	D mean	1.38
Brainstem	0.86	0.03	0.67	0.09	25%	13%	D max	2.71
Oral cavity	0.83	0.03	1.29	0.07	27%	15%	D mean	2.84
Parotid glands	0.84	0.02	1.23	0.16	37%	14%	D mean	3.54
Spinal cord	0.80	0.03	0.86	0.54	43%	11%	D max	4.01
Submandibular glands	0.78	0.06	0.80	0.05	42%	10%	D mean	1.14
Laryngopharynx	0.76	0.08	1.41	0.19	54%	27%	D mean	2.88
Lips	0.69	0.06	1.37	0.24	60%	25%	D mean	2.28
Optic nerves	0.61	0.05	1.28	0.03	57%	13%	D max	2.64

Figure 3. 1 axial views of a test case show the anterior-posterior and lateral contouring variation between the observers A, B and C

(1) Shows the eyes, optic nerves, whole brain and brainstem, (2) Shows the lips, both parotid glands, spinal cord, oral cavity, and part of mandible, (3) Shows the mandible, both submandibular glands, and spinal cord for observer A (pink), observer B (blue) and observer C (yellow).

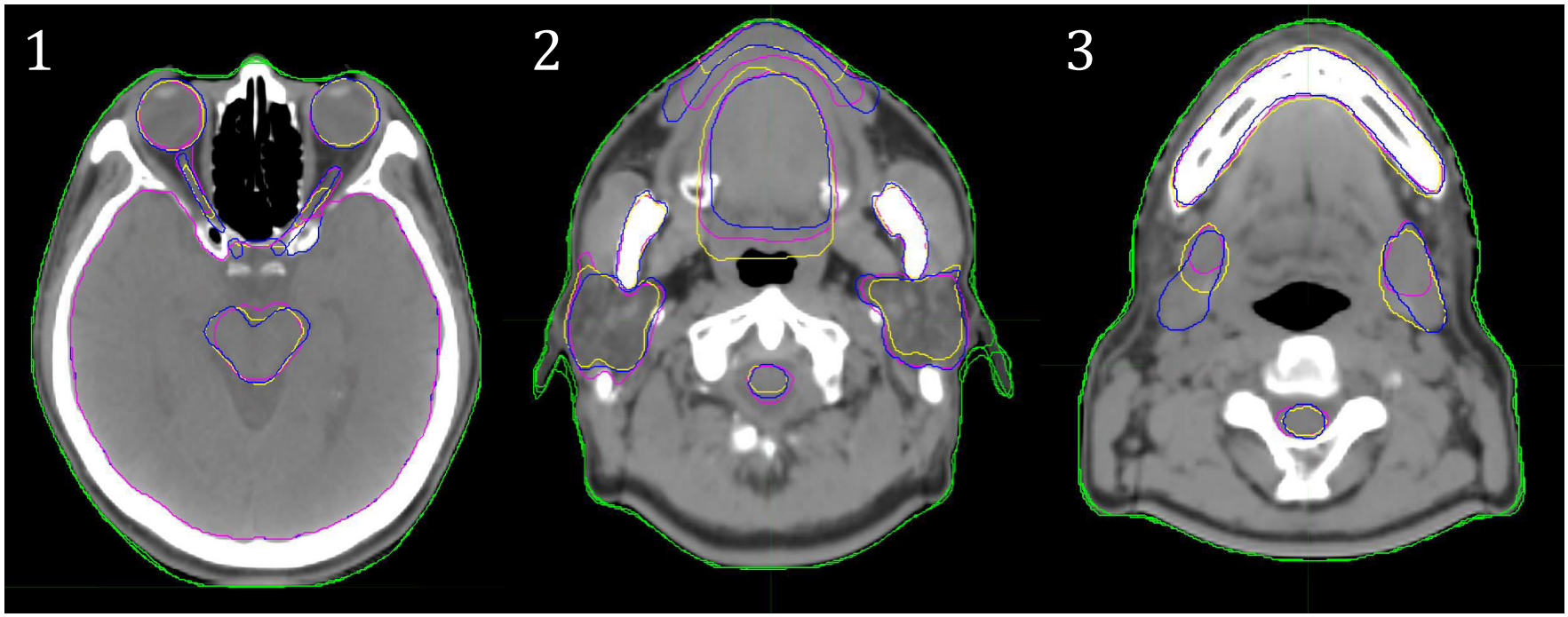


Figure 3. 2 Sagittal (1) and frontal (2) views of a test case show anterior-posterior and longitudinal contouring variation between the observers.

(1) Shows the whole brain, brainstem, lips, mandible, oral cavity, laryngopharynx and spinal cord. (2) Shows the whole brain and both parotid glands for observer A (pink), observer B (blue) and observer C (yellow).

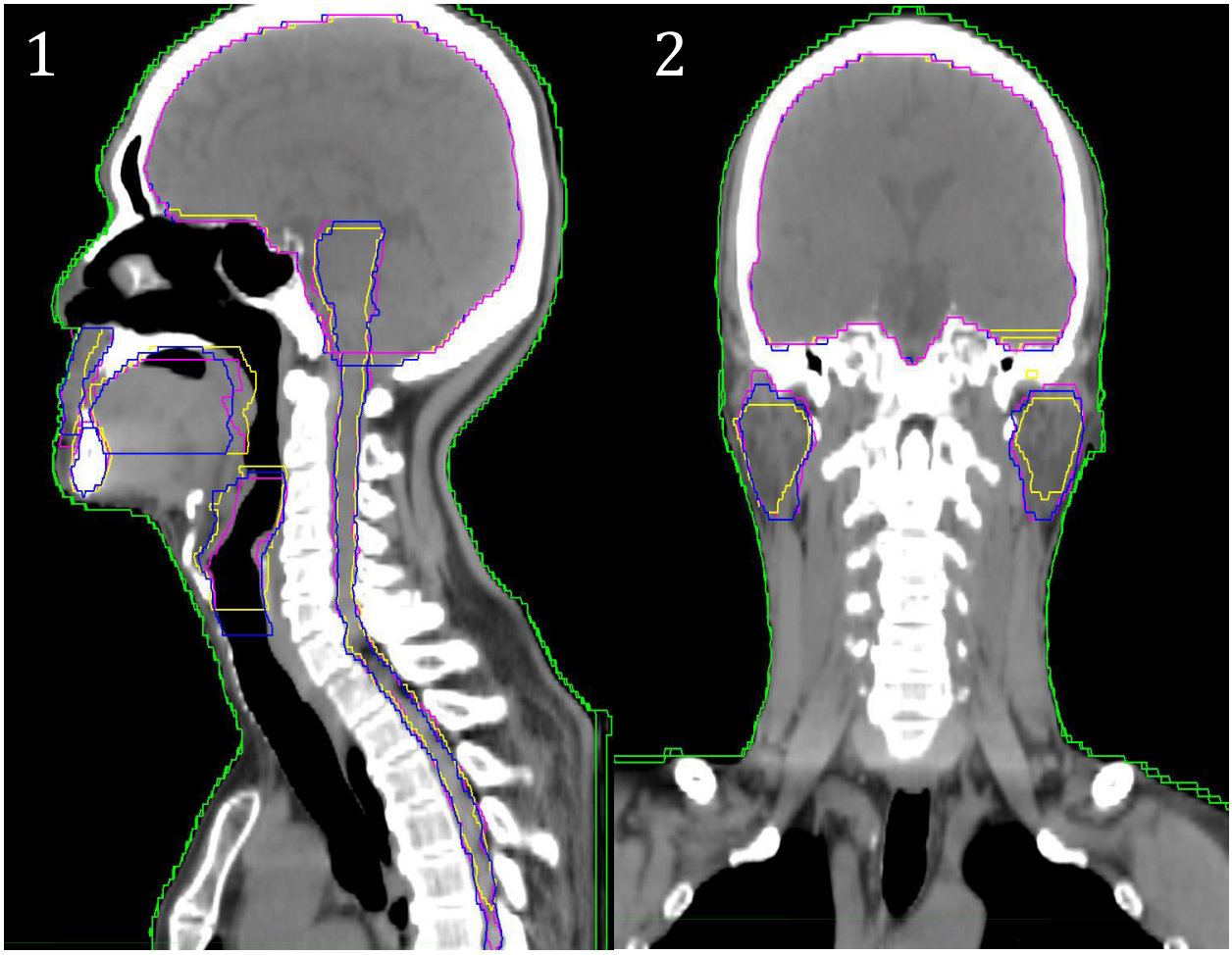


Figure 3. 3 average percentage of ΔV per organ over all eight cases, Eclipse and MIM represent inter-observer variation whereas observer A, B and C represent intra-observer variation

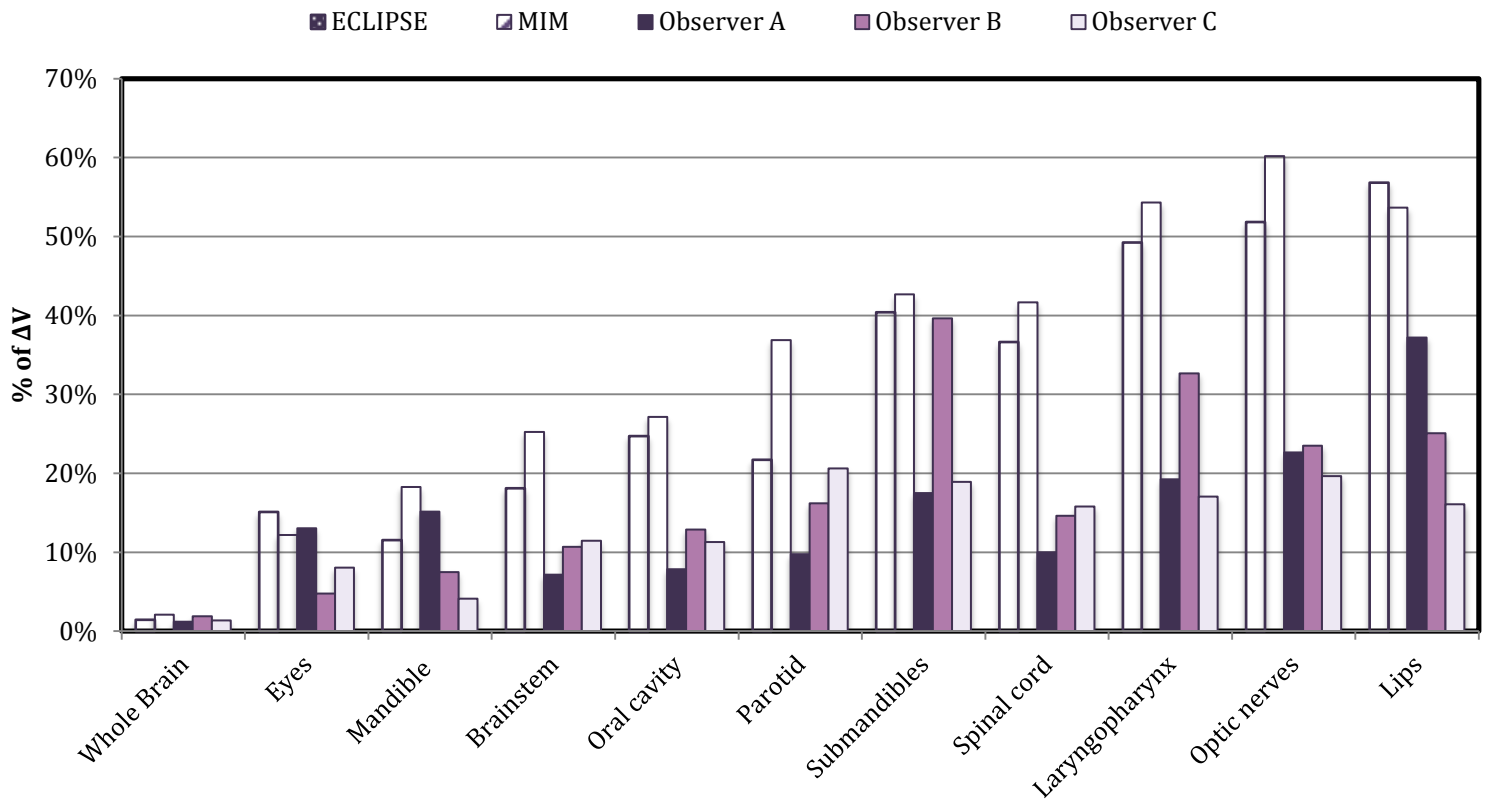


Figure 3. 4 Average DSC over the eight test cases, inter-observer represents average DSC over Eclipse and MIM results whereas intra-observer variation represents average DSC over observer A, B and C results

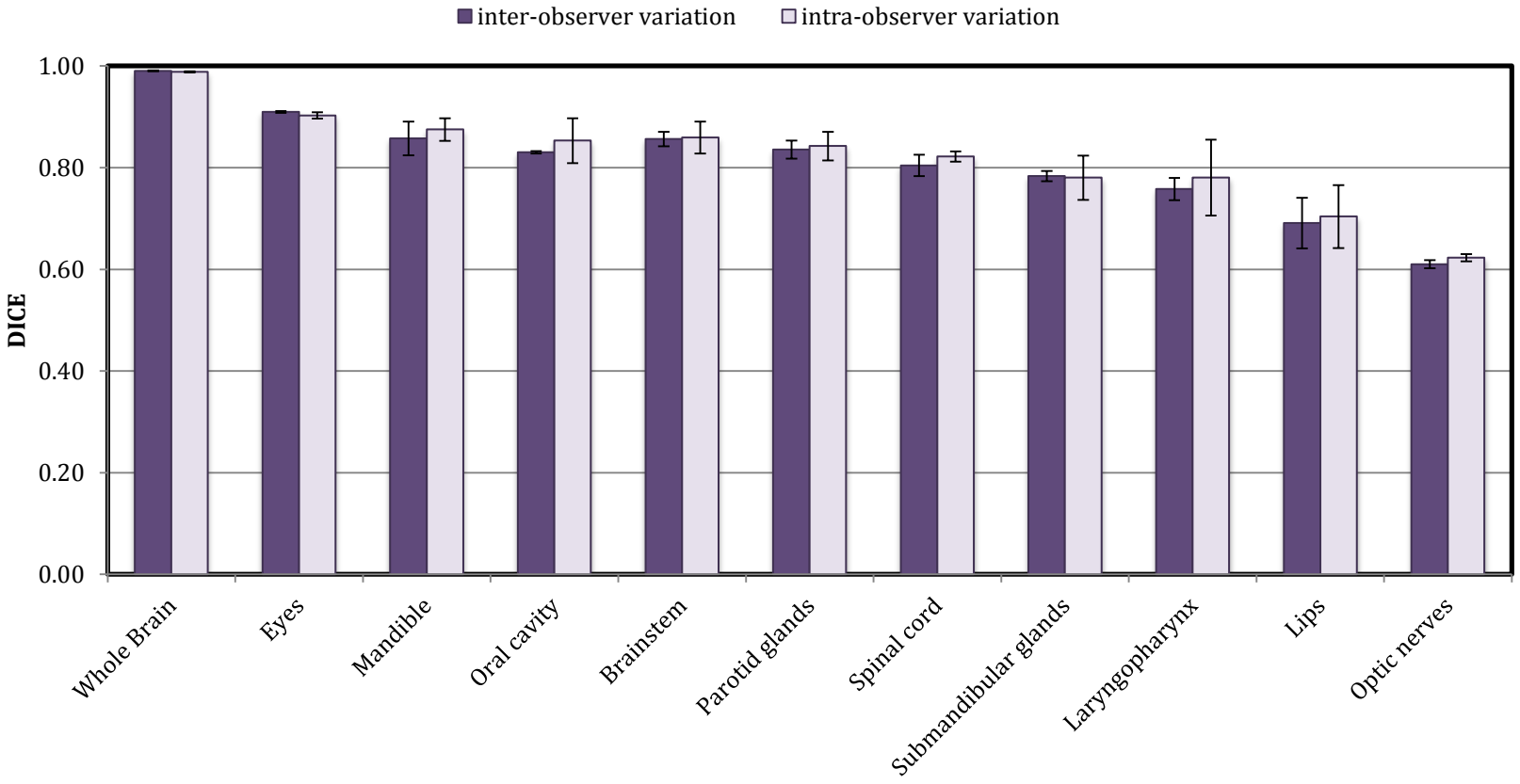


Figure 3. 5 Average HD (cm) over eight test cases, inter-observer represents average DSC over Eclipse and MIM results whereas intra-observer variation represents average DSC over observer A, B and C results

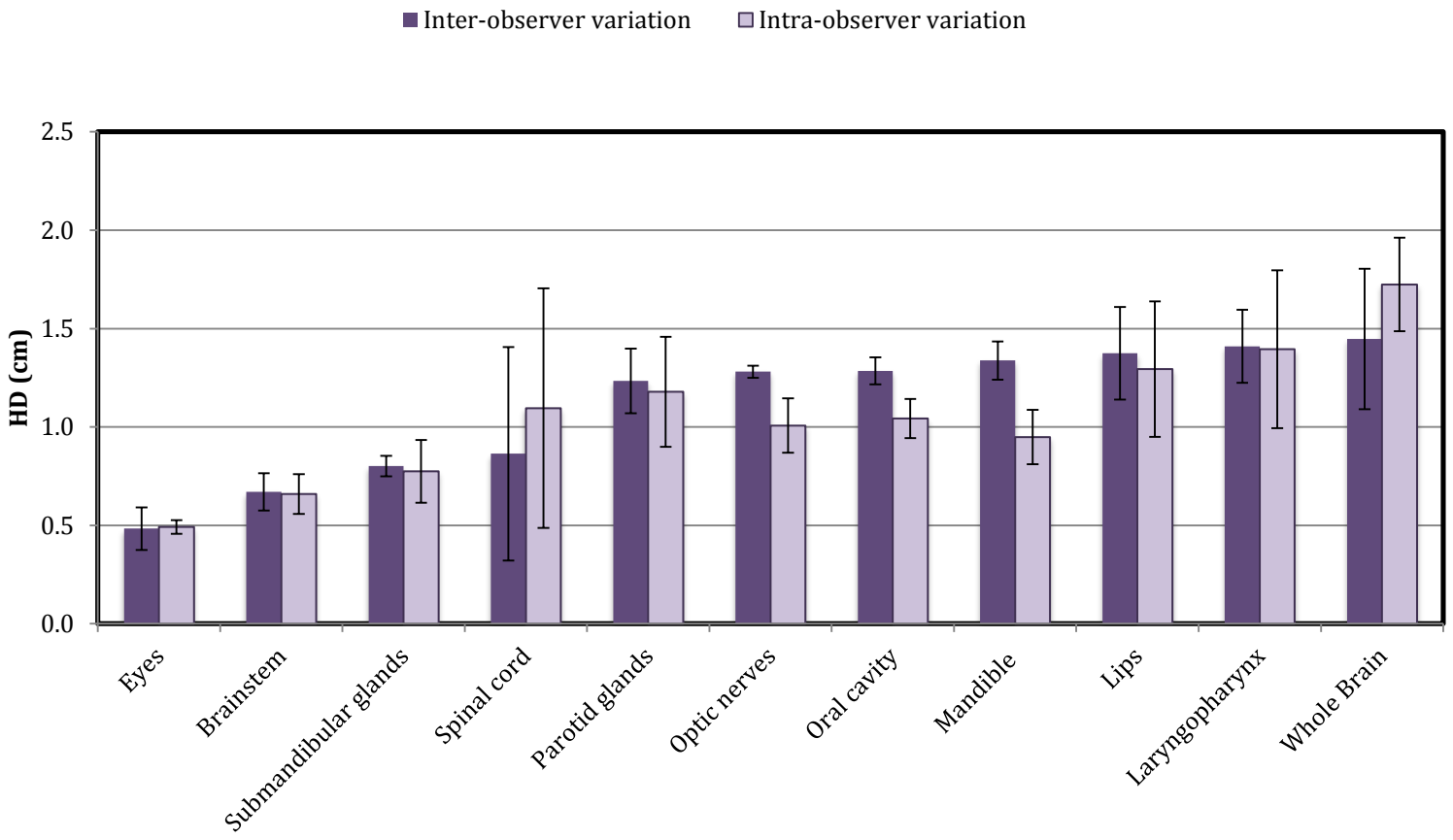
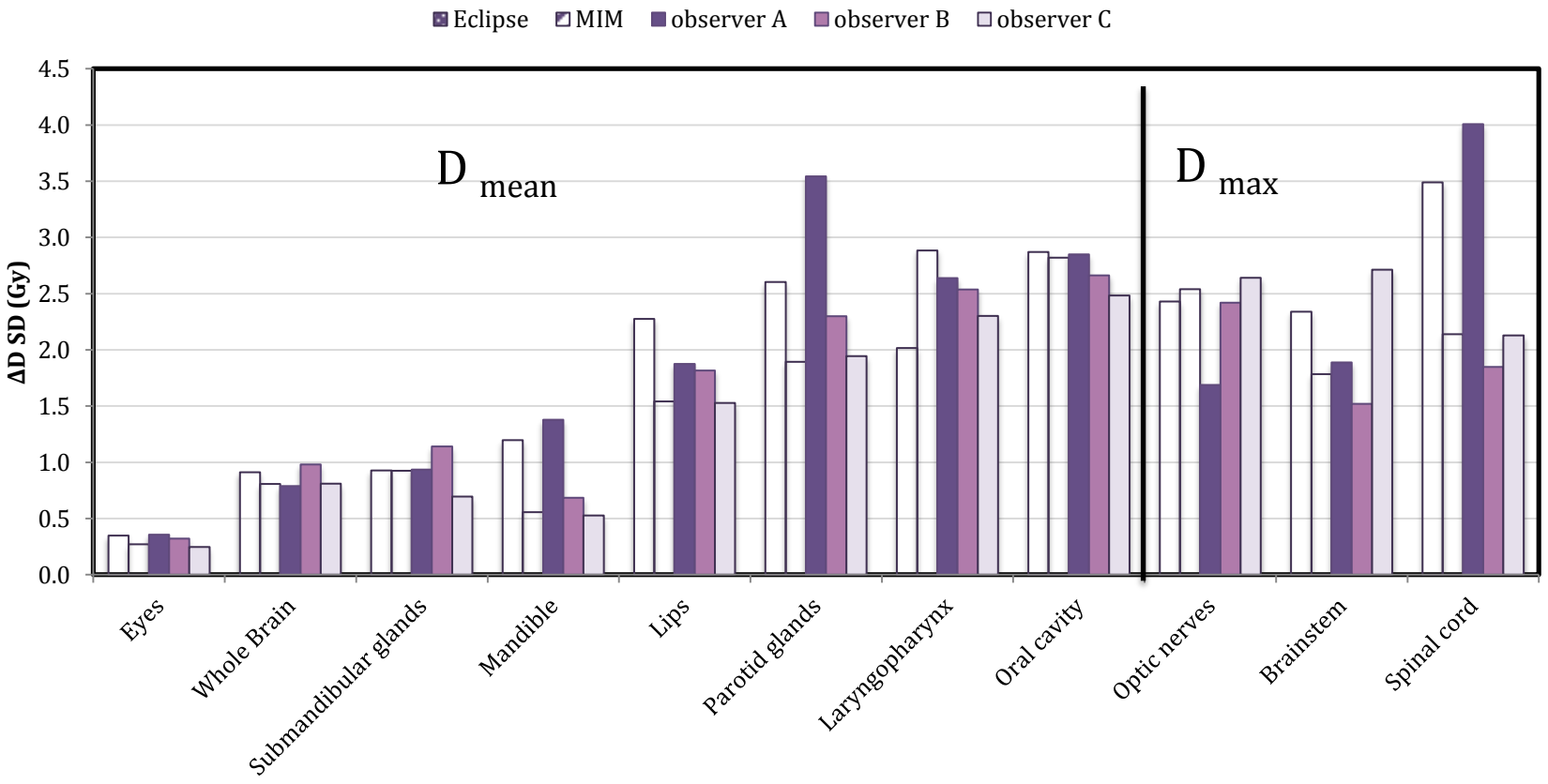


Figure 3. 6 SD of ΔD per organ over all eight cases, Eclipse and MIM represent inter-observer variation whereas observer A, B and C represent intra-observer variation



CHAPTER 4: PRE-CLINICAL ASSESSMENT OF IN-HOUSE ATLAS-BASED AUTO-SEGMENTATION (ABAS) PERFORMANCE FOR ORGANS AT RISK (OAR) IN HEAD AND NECK RADIATION THERAPY

4.1 INTRODUCTION

In Chapter 1, we discussed the advantages of using volumetric modulated radiotherapy (VMAT) for head and neck cancer, providing better sparing of organs at risk (OARs) and, therefore, reduced normal tissue toxicity. Auto-segmentation approaches developed for contouring in radiation therapy have been described in Chapter 2. These approaches seek to reduce contouring time and workload as well as reduce inter-observer variation. Prior to clinical implementation of auto-segmentation, it is advisable to evaluate the performance of the method, both in terms of both time-savings and the quality of the result.

In Chapter 3 manual delineation time and inter-observer variation for head and neck OARs were investigated. Manual delineation of a standard set of 15 OARs requires 30 minutes in our institution. La Macchia et al. [30] reported 2.7 hours on average to segment a single locally advanced head-neck cancer case, including the PTVs. Although OAR segmentation only represents one component of the segmentation work, inter-observer segmentation variability in OAR segmentation can be large source of uncertainty in radiation treatment planning [59]. The findings reported in Chapter 3

indicate that mean or maximum dose within an OAR can vary up to ±4 Gy due to inter-observer variation in our institution.

The purpose of this part of the project is to validate the performance of our in-house head and neck atlas, using the MIM Maestro™ auto-segmentation tool, by utilizing manual dosimetric and geometric variation reported in Chapter 3 as benchmarks. The ABAS OAR contours, without using any post-processing or manual interference, were assessed against reference contours from the original treatment plans. The timing of ABAS segmentation was also compared with timing of manual segmentation determined in Chapter 3.

4.2 MATERIALS AND METHODS

4.2.1. Atlas subject classification

An in-house atlas specifically for head and neck OAR contouring was built Using MIM Maestro™ software version 6.5. This atlas consists of 36 segmented CT data sets from previously treated cases. All cases were anonymized using the MIM Maestro™ anonymization tool. Each subject added to the atlas was reviewed carefully for completeness of structure contours and no tumor invasion to any organ at risk.

In order to assess how well our test cases were represented by the range of atlas subjects, all atlas and test cases were characterized by their age, gender, head tilt angle and some craniofacial indices [61,62] (e.g. cephalic index and facial index). Head tilt was defined as the angle between the horizontal and the line connecting the hard palate with

the dens on the C2 cervical vertebral body and measured on the CT sagittal plane (Figure 4.1).

The cephalic index is defined as the ratio of the maximum width of the head, to its maximum length (Figure 4.2), expressed as a %. The classification ranges from dolichocephalic (long-headed, with cephalic index < 75%.), to mesaticephalic (moderate-headed, with cephalic index 75% - 80%.), to brachycephalic (short-headed with cephalic index > 80%.). Sub-classifications of brachycephalic such as hyperbrachycephalic (cephalic index 85% - 90%) and ultrabrachycephalic (cephalic index > 90%) may also be used.

The facial index is defined as the ratio of the morphological face length to the morphological face width (Figure 4.3), also expressed as a %. The morphological face length is measured as the distance from the nasal root to the lowest point of the lower jaw. The morphological face width was defined as the most laterally located point on the zygomatic arch. According to facial index, individuals are categorized as euryproscopic (broad face with facial index < 85%), mesoproscopic (round face with facial index 85% - 90%.), and leptoproscopic (long face with facial index >90%). A subcategory of euryproscopic utilized to describe very broad faces is called hypereuryprosopic (facial index < 80%). On the other hand, hyperleptoprosopic (facial index < 95%) is used to describe a very long face.

4.2.2. *Atlas performance evaluation*

Using the MIM Maestro™ ABAS tool with user defined settings of five closest matches and the majority vote algorithm, our in-house head and neck atlas was used to auto-segment the same eight head and neck cases used in Chapter 3. The original contours and dose distributions, which had been used clinically, were used as a 'reference volumes' and 'reference doses', respectively. The results of Chapter 3 were used to provide performance benchmarks for the geometric and dosimetric performance of the ABAS tool.

Quantitative evaluation

Geometric evaluation was performed in a similar manner as described in Chapter 3 for evaluation of manual segmentation. Average volume as well as the percentage volume difference ΔV (%), DSC and Hausdorff distance (HD) comparing the auto-segmented structures to the reference structures were calculated. The average volume (cc) was calculated as the average absolute volume for each organ over the eight study cases, thus including patient –to- patient variation. ΔV (%) between the ABAS volume V_{ABAS} and the reference volume $V_{reference}$ was determined, per organ per case, by calculating:

$\Delta V\ (\%) = (V_{reference} - V_{ABAS}) \times 100 / V_{reference}$ ……………………………………………………..

(4.1)

The average ΔV (%) was calculated over the eight cases, per OAR.

In order to validate the performance dosimetrically, the original plans were superimposed onto the auto-segmented structure sets. The dosimetric endpoints were mean or maximum structure dose, depending on the structure. For the whole brain, eyes,

parotids, submandibular glands, oral cavity, laryngopharynx, mandible, and lips, the mean dose was evaluated whereas the maximum dose to the brainstem, spinal cord, and optic nerves was assessed. The standard deviation of dose differences (SD ΔD (Gy)) was assessed over all eight cases.

Quasi-quantitative evaluation

3D quasi-quantitative analysis of variation between auto-segmented and reference contours was performed to investigate which sub-regions were associated with the highest discrepancy for each OAR. This would indicate the region where manual modification might be needed. Cranial, caudal, medial, lateral, anterior, and posterior directions were evaluated for each structure using a simple scoring method. The integer scoring scale ranged from zero to two. A zero score represents variation <0.5cm. The score of one is assigned for a difference between 0.5 cm and 1 cm. A score of two is assigned if the discrepancy is more than 1 cm. Then, the scores of each structure were summed over the eight cases. Accordingly, zero will represent no significant variation and 16 will represent the highest rate of discrepancy.

4.2.3. Time assessment

The time was recorded for each case starting from the click on the segmentation button (t_{start}) until the segmentation process was fully completed (t_{end}). Δt was calculated as $t_{end} - t_{start}$. The average time (min: sec) over eight ABAS studies was then calculated.

4.3 RESULTS AND ANALYSIS

4.3.1. Subject classification

Atlas Subjects

Our atlas consists of 36% and 64% female and male data sets, respectively. The median age of the subjects was 60 years (32-97). The median head tilt angle was 15.6° (2°-25°). More than 55% of the subjects had head tilt angle between 15° and 25°. More than 40% of the atlas cases were classified as brachycephalic (short-headed), while 36% were mesaticephalic (moderate-headed) (Figure 4.4). Facial classification indicates a similar number of euryprosopic (broad faces) and leptoproscopic (long faces) with 17% of the atlas population each, whereas mesoproscopic (round faces) accounted for more than 25% of the population (Figure 4.5).

Test cases

Two female and six male NPC cases were tested in this study. In regards to the cephalic index, two cases were classified as mesaticephalic, three brachycephalic, two hyperbrachycephalic and one ultrabrachycephalic case. Thus, 75% were short or very short headed and 25 % were moderate headed. Face shape classification indicates two euryproscopic, three leptoproscopic and three hyperleptoproscopic cases (three broad faces and five long and very long faces). Cephalic and facial indices of the test cases covered almost all the categories and they fit within the distribution of the atlas cases as shown in Figure 4.6 Head tilt indicates three cases <10°, two cases between 10° and 15° and three cases between 15° and 20°. Table 4.1 summarizes these findings.

4.3.2. Auto-segmentation result analysis

Table 4.2 represents the average absolute volume ±SD across the eight study cases. ABAS results for this metric were within inter- and intra- observer variation for the majority of the structures. The largest discrepancy between ABAS and inter-observer average volume was associated with the laryngopharynx., with ABAS performing less well compared with the manual observers.

Figure 4.7 presented the percentage difference in volume (ΔV %) for ABAS segmentation, in comparison with the maximum and minimum inter-observer volume differences. From the Eclipse and MIM results from Chapter 3, the graph demonstrates ABAS (ΔV %) <30% overall, for all the structures, which was significantly less than the maximum inter-observer ΔV % results. The largest improvement of ABAS over the maximum inter-observer ΔV % was associated with optic nerves, spinal cord, lips, and laryngopharynx. Moreover, the ABAS results were within only 1SD from the minimum inter-observer ΔV %.

The auto-segmentation tool successfully achieved average DSC >0.8 for the whole brain, eyes, mandible, brainstem, oral cavity, and spinal cord as shown in Figure 4.8. Furthermore, it was able to outperform the manual inter-observer segmentation for lip contours with an average DSC 0.75±0.11 vs. 0.69±0.05. Optic nerves auto-contours achieved DSC results consistent with the manual segmentation (0.61±0.09 vs. 0.61±0.01 for ABAS and manual segmentation, respectively). On the other hand, manual

segmentation outperforms auto-segmentation in the delineation of parotid glands, submandibular glands, and laryngopharynx.

Further, auto-segmentation was able to achieve HD < 1cm for the spinal cord, eyes, brainstem and submandibular glands contours (Figure 4.9). ABAS significantly outperforms the manual segmentation of spinal cord with the least HD on average. On the other hand, HD of ABAS was minimally higher than the manual variation of lips and submandibular glands.

Dosimetric results are displayed in Figure 4.10. Auto-segmentation showed SD ΔD less than the worse manual segmentation for the mean dose of the mandible, lips, and laryngopharynx, and the maximum dose of the spinal cord and optic nerves. A significant improvement in SD ΔD of ABAS spinal cord, lips, and mandible compared to manual segmentation ones. On contrary, ABAS parotid and submandibular glands demonstrated considerably higher SD ΔD compared to the manual segmentation.

Table 4.3 represents the results of the quasi-quantitative analysis. This analysis demonstrated high variation in the cranial and caudal direction compared to the other directions. Furthermore, it indicated that the highest discrepancy was associated with the cranial and caudal direction of ABAS laryngopharynx while the caudal and posterior borders of the ABAS oral cavity presented the maximum variation compared to the other directions. ABAS parotid glands indicated significant differences in the cranial and caudal direction and less discrepancy in the medial regions. The largest variation associated with ABAS lips contours were in the cranial, caudal and lateral.

Table 4.4 represents the time requires to auto-segment complete head and neck structure set. The shortest auto-segmentation time was one minute and forty-seven seconds while the average segmentation time over the eight cases was only two minutes and twenty-nine seconds compared to 30 minutes for manual segmentation, which was reported in Table 3.5.

Figures 4.11 and 4.12 represent a typical segmented case, demonstrating head and neck OARs on an axial, sagittal and frontal plane. According to these results, our in-house ABAS was able to generate structures within manual segmentation variation for most of the time.

4.4 DISCUSSION AND CONCLUSION

Our in-house atlas was constructed to segment a complete head and neck OARs structure set including the most frequently delineated structures. These structures are the whole brain, brainstem, spinal cord, eyes, optic nerves, optic chiasm, parotid glands, submandibular glands, mandible, lips, oral cavity, and laryngopharynx. The assessment of ABAS optic chiasm contours was excluded from this analysis due to the lack of the reference optic chiasm contour on the test cases.

First of all, we classified our atlas subjects in terms of gender, age, head shape, face shape, and the degree of head tilt to ensure that our atlas subject covered a wide variety of features. Thus, the chance of finding the similar cases to our patients might be increased. The head tilt was expected to impact on the delineation of the optic chiasm because of its very small volume. This was confirmed when we asked the system to

delineate the optic chiasm for a case with head tilt angle out of the ABAS range (>40°). ABAS failed to segment it as shown in Figure 4.13. So, it could be expected that for any case with head angle out of the atlas subjects' head angles range, the optic chiasm either could not be auto-segmented or ABAS would perform poor segmentation. Also, we tested each selected beast match case to see which index has an effect on the choice of best match. We found that head shape and tilt angle had some impact on the chosen close match cases. Facial index did not particularly impact the best match selection.

In comparison with benchmark data in Table 3.9, ΔV % of our ABAS structures was superior to manual segmentation for all structures. Also, DSC and HD results indicated that ABAS performs as well as the manual segmentations for most of the structures. Thus, it would be reasonable to expect that the need for manual correction would be reduced using ABAS in place of manual segmentation. . Further, this supports the idea that using ABAS minimizes inter-observer variation [36]. In addition, the dosimetric results showed that the use of ABAS contours did not negatively impact on dosimetric endpoints for more than 80% of the structures. Significant improvement was shown in the SD of ΔD_{max} (Gy) of ABAS spinal cord compared to manual segmentation. Except for parotids, the results demonstrated that SD of ΔD (Gy) for all auto-segmented organs was < ±3Gy. It is of note that these results were achieved for structures created in less than three minutes compared to 30 minutes for manual segmentation.

Our quasi-quantitative study indicated a higher discrepancy in the cranial and caudal direction compared with medial and lateral directions. This was consistent with Brouwer et al. [22] who also reported that the largest inter-observer variation was

observed in the cranial, caudal and medial direction for all of the studied structures. These results could help the user of this atlas based software tool to predict the area of most common mismatch. Thus, the reviewing and modification process, if required, could be made easier.

Further, many studies have confirmed substantial time saving in using ABAS either for re-planning or when starting a case from scratch. All of these studies focused on comparison with manually corrected ABAS structures. Quantitative comparisons (represented in Table 4.5, 4.6 and 4.7) were performed on similar geometric indices between:

[1] The result of unedited ABAS in this study and the unedited ABAS data from three other studies., and

[2] The result of unedited ABAS in this study and the manually modified ABAS structures from four studies; these studies are:

La Macchi et al. [30] investigated the performance of three available commercial ABAS tools, VelocityAl, ABAS (Elekta), and MIM, using five head and neck planning CT as atlas cases to segment the re-CT of the same five cases. They used a single atlas approach; with each planning CT representing the atlas subject used to segment its re-CT study.

Mattiucci et al. [58] aimed to present benchmark data by comparing the manually modified ABAS structures with the structures delineated with the common agreement of five expert operators. Also, they intended to quantify the reliability of using auto-contouring for re-planning. They examined the performance of ABAS in the contouring of

10 re-planning NPC cases. The deformable registration and re-planning ABAS was performed using VelocityAI 2.3™ software.

Sims et al. [40] constructed head and neck ABAS consisting of 45 head and neck cases, using the ISOgray™ ABAS tool, to delineate brainstem, spinal cord, parotid and submandibular glands, mandible and the lymph nodes. They compared the manually modified results of brainstem, parotid glands and mandible only.

Tao et al. [36] offer a multi-institution clinical study to investigate the ability of ABAS in minimizing inter-observer variation. They used an atlas consisting of 50 head and neck cases to automatically delineate 20 head and neck OARs of 16 patients that had been contoured by eight radiation oncologists from 8 institutes. They used multi-atlas subjects and the STAPLE algorithm. Again, they compared the manually edited ABAS with manual contouring.

DSC results of our unedited ABAS were consistent or superior compared to all unedited ABAS results of La Macchi et al. [30], Sims et al. [40] and Mattiucci et al. [58]. More interestingly, the results of our unedited ABAS structures, except for salivary glands, were consistent or better than the manually modified ABAS of La Macchi et al. [30], Sims et al. [40] Mattiucci et al. [58] and Tao et al. [36]. For salivary glands, DSC results of all the four studies ranged between 0.79 and 0.85 for parotid glands and Tao et al. [36] reported 0.81 for submandibular glands. These results were notably superior compared to our results (0.74 and 0.69 for parotid and submandibular glands, respectively).

La Macchi et al. [30] reported the percentage of volume variation (ΔV%) of the three software platforms. Although their results were expected to provide high quality results as they used a single atlas approach to segment re-CT of the same case, our results were consistent with their stated results for all the structures.

These studies reported automatic segmentation time ranged from 2 - 10 minutes. This time depends on the number of best matches, registration process, and processor speed of the used workstation. The required time to manually modify ABAS structures was reported by Mattiucci et al. [58]. They reported 22.2 minutes on average, ranging from 10.6 – 44 minutes, to manually modify the 14 ABAS OARs. They also reported that the mean time to manually segment the same 14 OARs from scratch was 29.7 minutes (ranged from 20.5 – 35.2 minutes). This was consistent with the average time for the segmentation of 15 head and neck OARs, which was 29.73 minutes at our institution, as reported in Chapter 3. This indicates that manual editing of ABAS contours is also a time consuming task and could reduce the potential timesaving achievable using an ABAS approach.

It is interesting to highlight that the evaluated structures in this study were the auto-generated structures without using any manual interference. Consequently, most of these structures, which were segmented within less than three minutes, provided geometric and dosimetric results close to the structures delineated manually in almost 30 minutes. This encouraging result provided the motivation to explore clinical

implementation of an auto-segmentation workflow, in a small pilot study, the subject of future work.

Table 4. 1 The test cases' craniofacial and head tilt angle classification

	Cephalic index	Facial index	Head tilt angle
1	Hyperbrachycephalic (very short headed)	Euryprosopic (broad face)	15.69
2	Hyperbrachycephalic (very short headed)	Leptoprosopic (long face)	17.36
3	Mesocephalic (medium headed)	Hyperleptoprosopic (very long face)	8.78
4	Brachycephalic (short headed)	Hyperleptoprosopic (very long face)	4.27
5	Ultrabrachycephalic	Euryprosopic (broad face)	12.10
6	Brachycephalic (short headed)	Leptoprosopic (long face)	15.38
7	Brachycephalic (short headed)	Leptoprosopic (long face)	3.87
8	Mesocephalic (medium headed)	Hyperleptoprosopic (very long face)	14.29

Table 4. 2 Average volume ±SD (cc) of the delineated volumes over eight test cases per organ for ABAS, compared with inter-observer and intra-observer results from Chapter 3.

Structure	Auto-segmentation	Inter-observer	Intra-observer
Optic nerves	0.80±0.26	0.78 ± 0.21	0.58 ± 0.09
Submandibular glands	6.94±1.65	7.42 ± 1.39	5.57 ± 0.87
Eyes	9.22±1.33	8.75 ± 0.64	6.56 ± 0.40
Spinal cord	22.50±5.13	22.76 ± 4.64	17.07 ± 1.63
Brainstem	30.07±8.44	25.26 ± 2.89	18.95 ± 1.34
Lips	28.38±7.47	28.82 ± 7.74	21.61 ± 4.03
Parotid glands	27.17±8.92	30.22 ± 10.42	22.66 ± 10.19
Laryngopharynx	46.71±10.49	36.99 ± 9.59	27.74 ± 4.50
Mandible	77.14±14.85	72.86 ± 5.59	54.64 ± 3.47
Oral cavity	81.17±11.28	86.79 ± 11.66	65.09 ± 4.78
Whole Brain	1388.52±192.69	1377.83 ± 189.79	1033.37 ± 191.92

Table 4. 3 3D quasi-quantitative evaluations represent the sum of the scores per organ over the eight cases. Each organ in each case received a score of 0, 1 or 2 representing <5 mm, 5mm to 10mm, and > 10mm discrepancy between the auto-segmented and reference contour. Thus 0 represents the best score while 16 represents the worst.

Structure	Cranial	Caudal	Medial	Lateral	Anterior	Posterior
Whole Brain	0	0	0	0	0	0
RT Eye	0	0	0	0	0	0
LT Eye	0	0	0	0	0	0
RT Optic nerve	0	0	0	0	0	0
LT Optic nerve	0	0	0	0	0	0
Mandible	0	0	0	0	0	0
Brainstem	4	3	0	0	0	0
Spinal cord	4	0	0	0	0	0
Laryngopharynx	12	13	0	0	1	0
Lips	7	7	0	8	0	0
RT Submandible	6	3	1	1	5	6
LT Submandible	6	3	1	1	6	6
Oral cavity	3	12	0	0	0	13
RT Parotid	9	8	7	3	3	0
LT Parotid	8	8	7	3	2	0

Table 4. 4 ABAS timing data for each case

	t (start)	t (end)	Δ t
1	4:28:00	4:31:02	0:03:02
2	4:32:35	4:35:00	0:02:25
3	4:35:57	4:38:27	0:02:30
4	4:39:01	4:42:06	0:03:05
5	4:43:13	4:45:43	0:02:30
6	4:46:38	4:49:10	0:02:32
7	4:50:03	4:51:50	0:01:47
8	4:52:53	4:54:57	0:02:04
Average time ± SD (hr:min:sec)			0:02:29 ± 0:00:26

Table 4. 5 DSC comparisons with unedited atlas results from other studies:

Structure	La Macchi [42][a]	Sims [31][b]	Mattiucci [58][c]	Khawandanh [d]
	Unedited ABAS	Unedited ABAS	Unedited ABAS	Unedited ABAS
Mandible	0.84 - 0.89	0.78	0.83	0.89
Spinal cord	0.7 - 0.81	N/A	0.76	0.84
Parotid	0.73 - 0.79	0.68	0.74	0.74
Brainstem	0.77 - 0.81	0.58	0.84	0.85
Brain	N/A	N/A	0.96	0.97
Eyes	N/A	N/A	0.84	0.90
Oral cavity	N/A	N/A	0.76	0.83

a) Atlas of 5 planning CT created to segment Re-CT images, single best match, 3 software platforms (min-max).

b) 45- atlas cases, multi-atlas.

c) Atlas of 10 planning CT created to segment Re-CT images, single best match.

d) 36- Atlas cases, multi-atlas (5 best matches), Majority vote algorithm.

Table 4. 6 DSC comparisons with edited atlas results from other studies.

Structure	La Macchi [42][a]	Sims [31][b]	Tao [40][c]	Mattiucci [58][d]	Khawandanh[e]
	Edited ABAS	Edited ABAS	Edited ABAS	Edited ABAS	Unedited ABAS
Mandible	0.89- 0.9	0.82	0.90	0.89	0.89
Spinal cord	0.84 - 0.87	N/A	0.82	0.79	0.84
Parotid	0.8 – 0.82	0.85	0.84	0.79	0.74
Brainstem	0.88 – 0.89	0.77	0.86	0.85	0.85
Submandibular glands	N/A	N/A	0.81	N/A	0.69
Brain	N/A	N/A	N/A	0.97	0.97
Eyes	N/A	N/A	0.90	0.89	0.90
Optic nerves	N/A	N/A	0.66	N/A	0.61
Oral cavity	N/A	N/A	0.88	0.80	0.83

a) Atlas of 5 planning CT created to segment Re-CT images, single best match, 3 software platforms we displayed minimum - maximum results.

b) 45- atlas cases, multi-atlas.

c) 50- atlas cases, multi-atlas, STAPLE

d) Atlas of 10 planning CT created to segment Re-CT images, single best match.

e) 36- Atlas cases, multi-atlas (5 best matches), Majority vote algorithm.

Table 4. 7 ΔV (%) comparison with other studies

Structure	La Macchi [42] ΔV (%) (For three different platforms)			Khawandanh ΔV (%)
	ABAS (Electa)	MIM	VelocityAI	
Mandible	-5 ±7	-1 ±10	- 4 ±12	10 ±8
Spinal cord	-33 ±19	-9 ±17	-8 ±15	11 ±9
RT Parotid	−1 ±22	9 ±22	13 ±30	20 ±14
LT Parotid	-6 ±17	4 ±21	13 ±5	
Brainstem	-18 ±8	14 ±20	14 ±27	19 ±18

Figure 4. 1 Head tilt angle measurement.

This figure showed the sagittal view where the head tilt was measured as the angle between the upper jaw and horizontal line crossing C2.

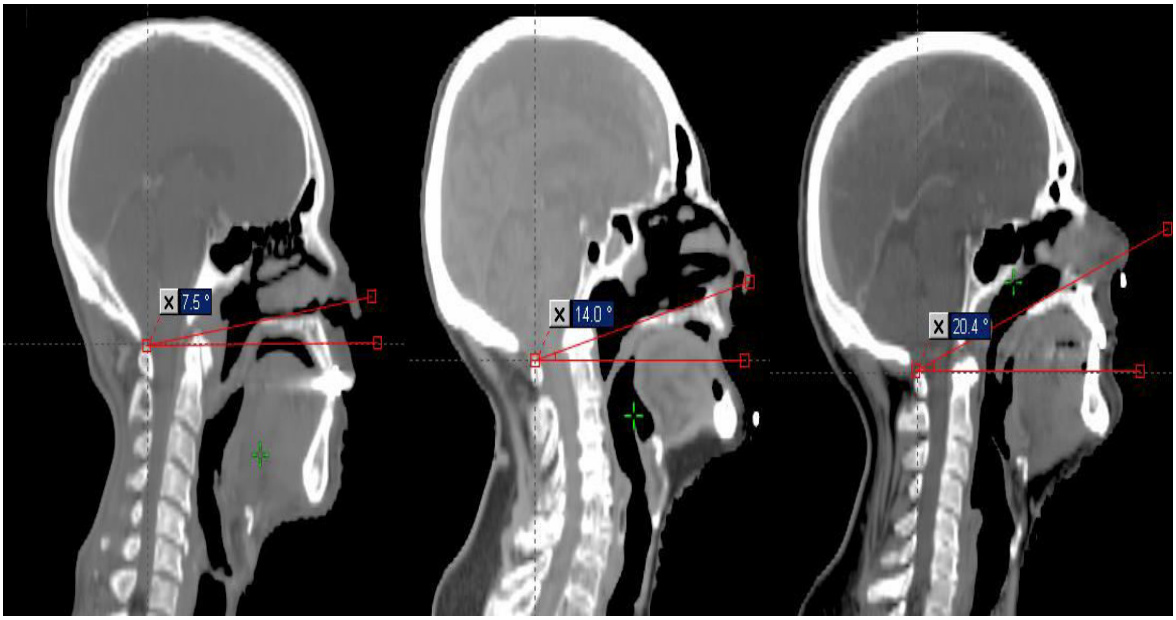


Figure 4. 2 Cephalic index measurement

The red line (A) represents the maximum head width and the red line (B) represents the maximum length

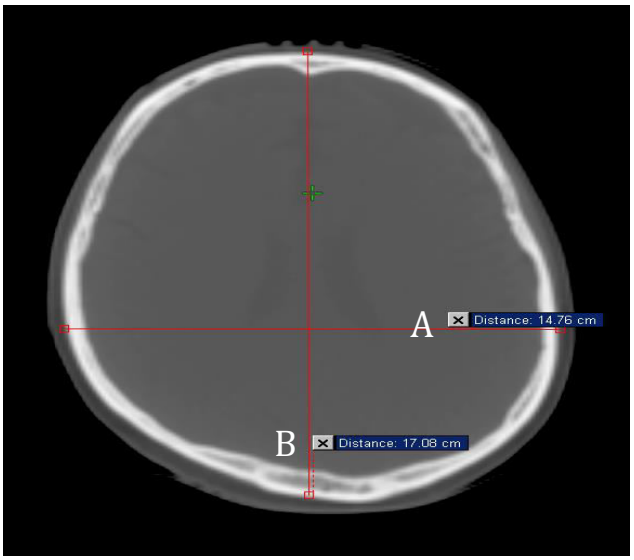


Figure 4. 3 Facial index measurement

The red line (A) represents the morphological face width and the red line of (B) represents the morphological face length

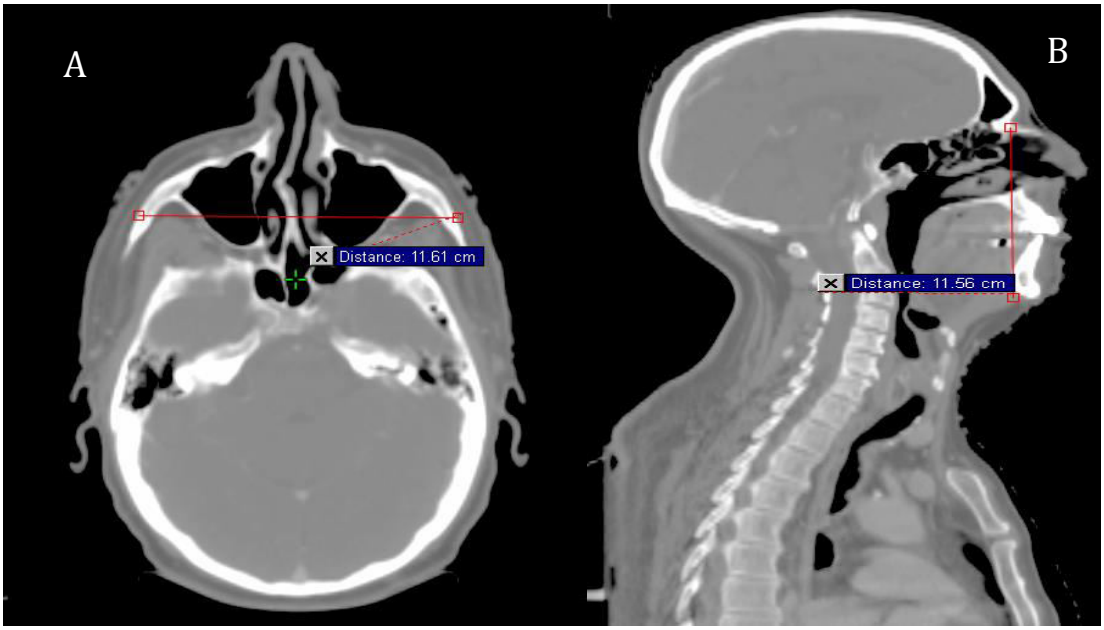


Figure 4. 4 ABAS subjects' classification according to head shape categorization

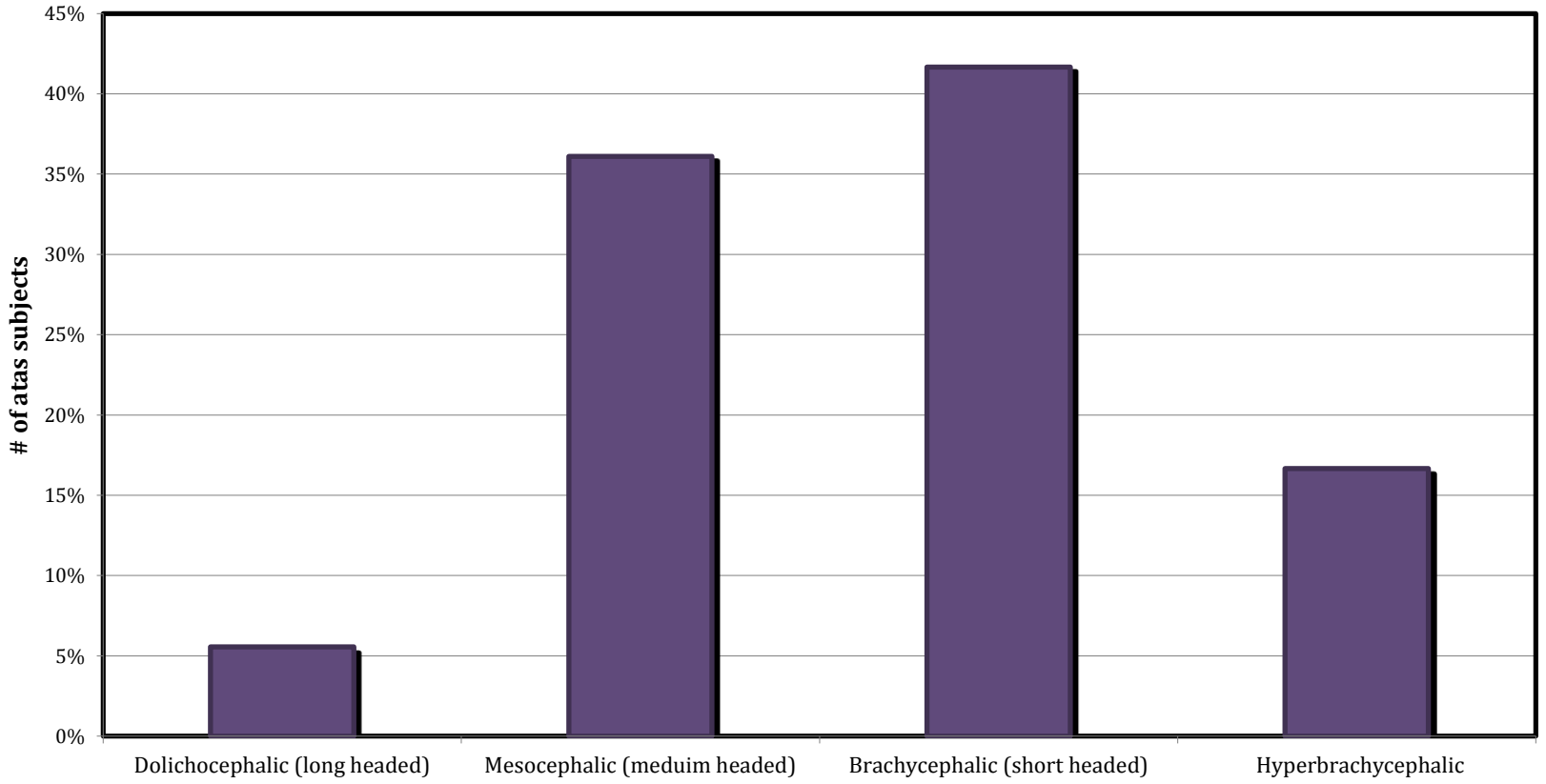


Figure 4. 5 ABAS subjects' classification according to face shape categorization

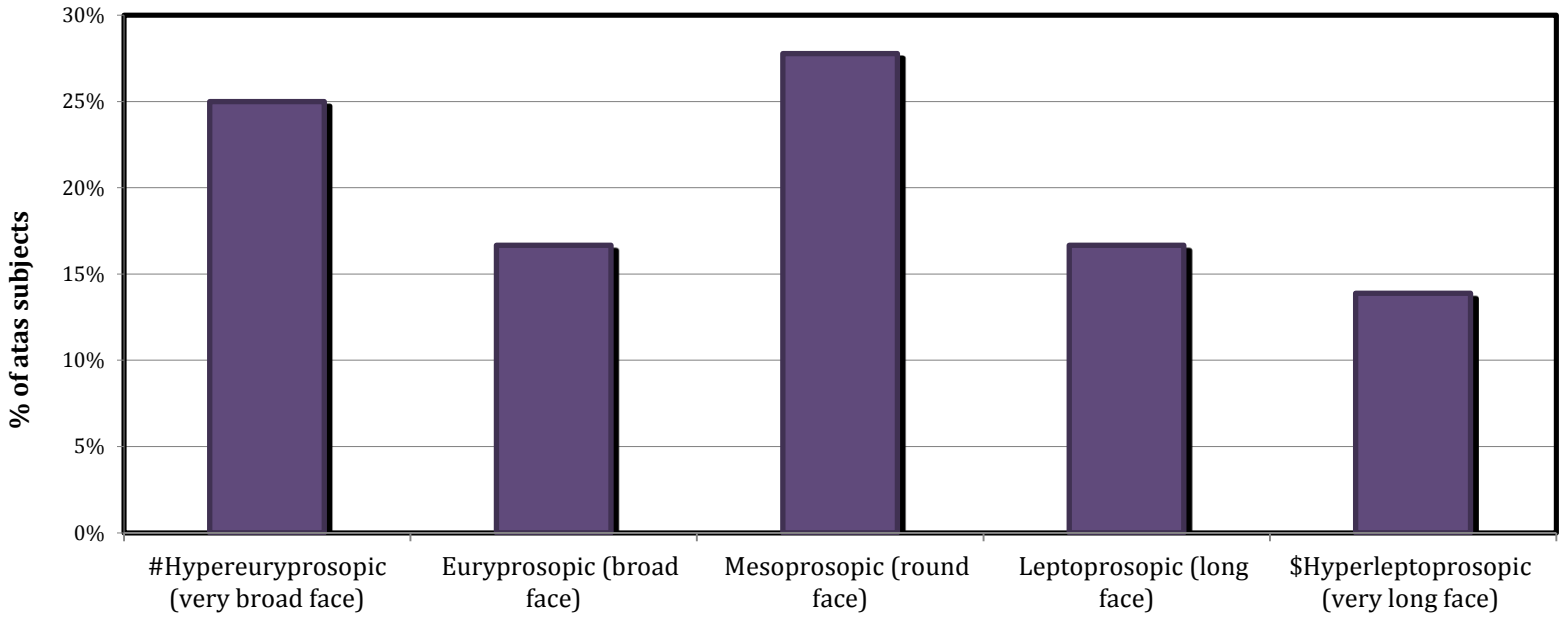


Figure 4. 6 Face and head classification of atlas (cross ✖) and test (square) cases.

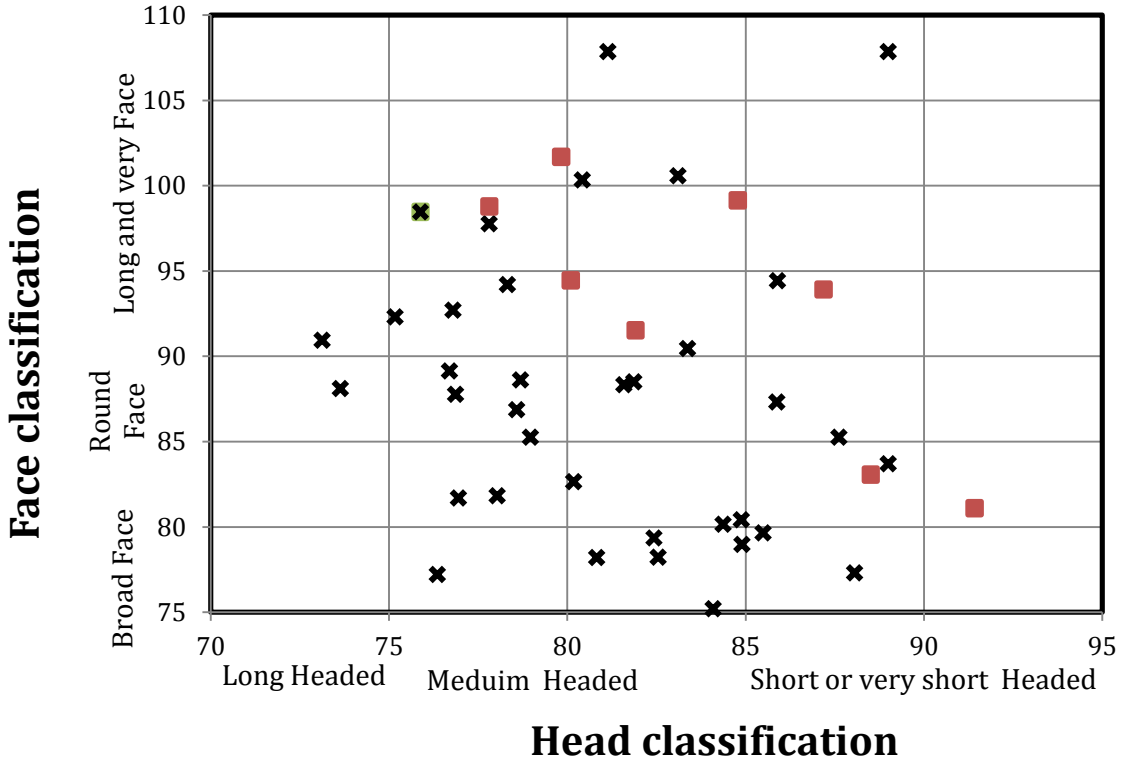


Figure 4. 7 The average ΔV% per organ overall eight cases, comparing ABAS to the maximum and minimum manual variation reported in Chapter 3.

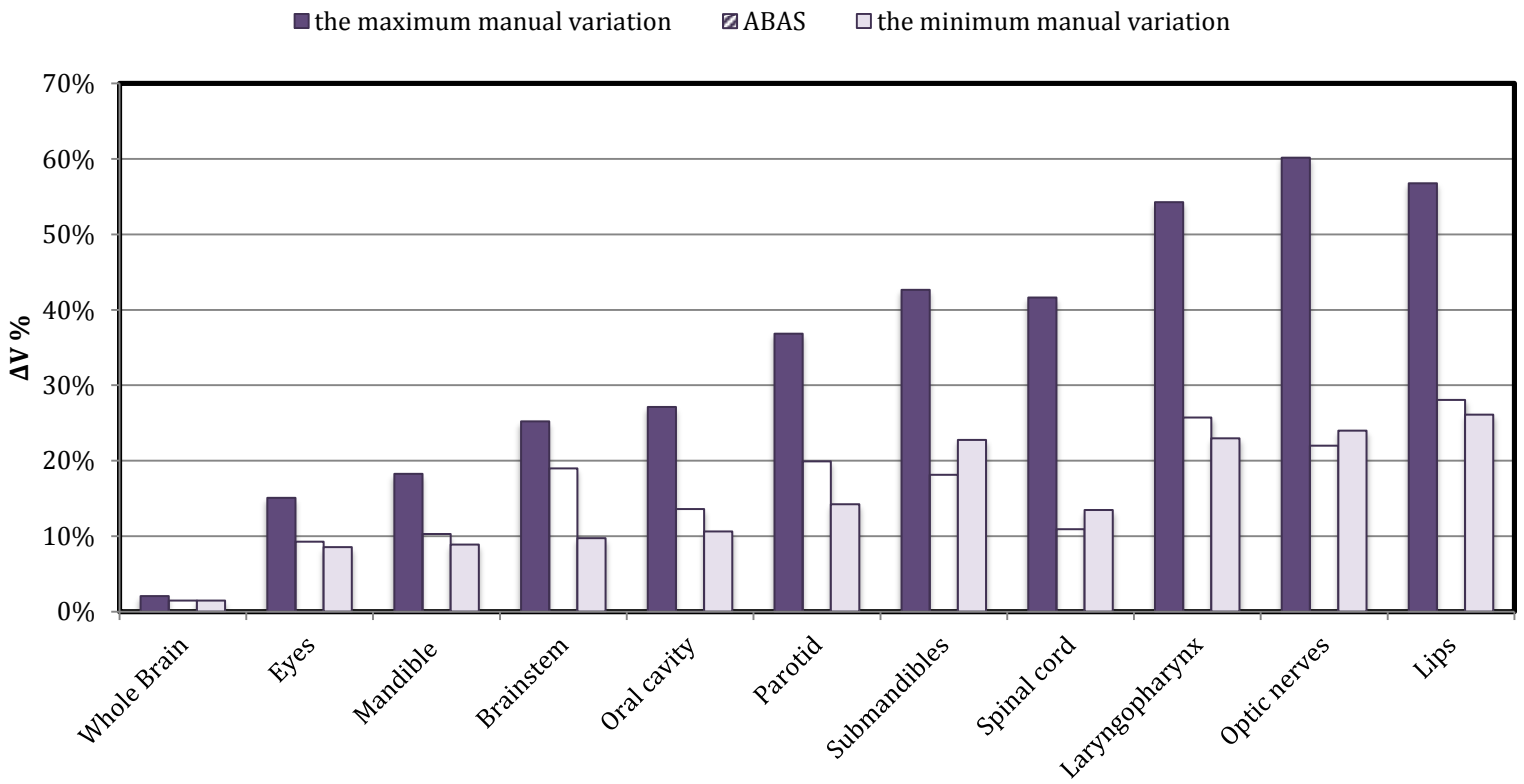


Figure 4. 8 Average DSC over eight test cases, comparing the ABAS results with inter-observer and intra-observer variation reported in Chapter 3.

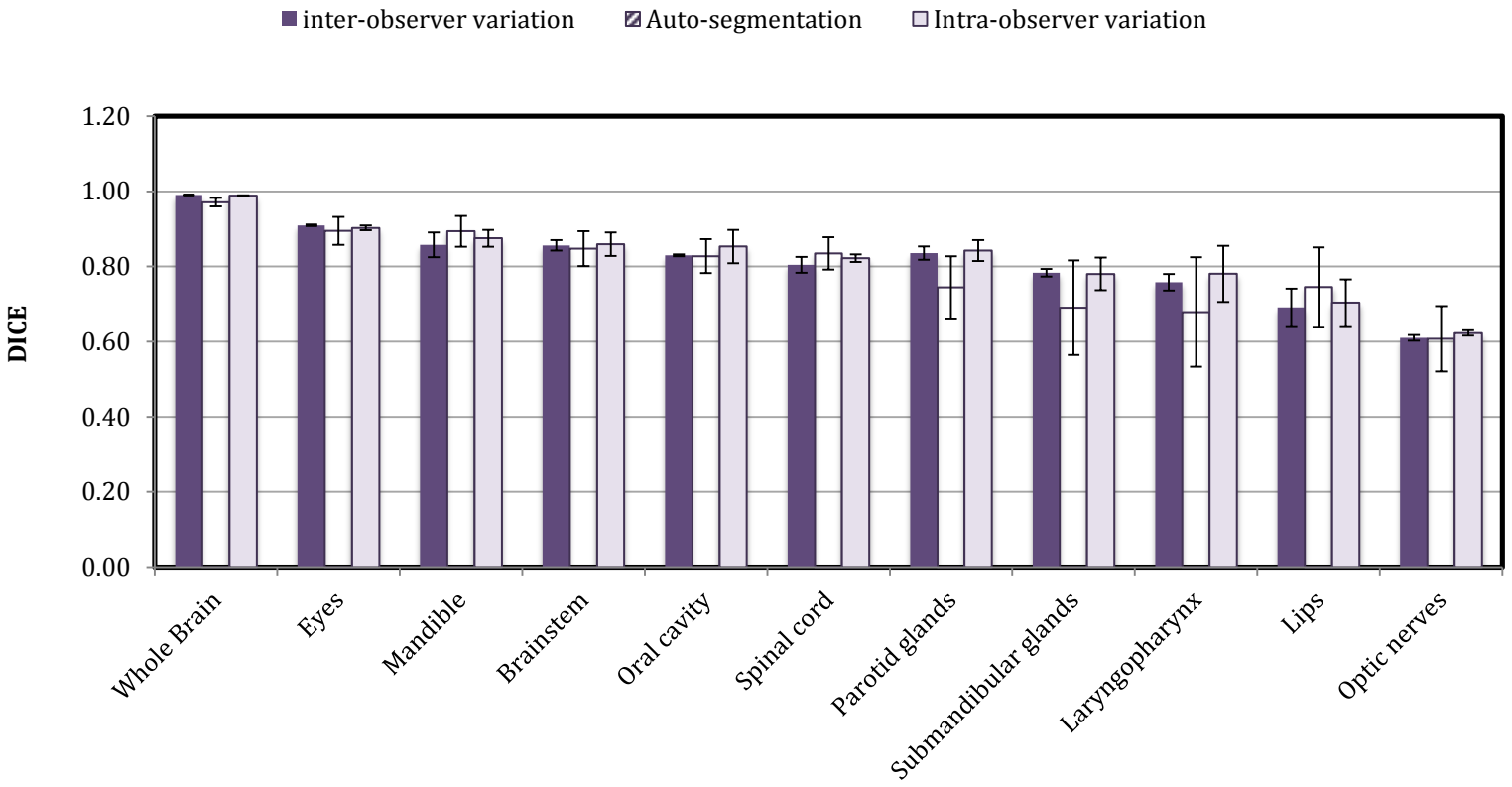


Figure 4. 9 Average HD (cm) over eight test cases, comparing ABAS results with inter-observer and intra-observer variation reported in Chapter 3

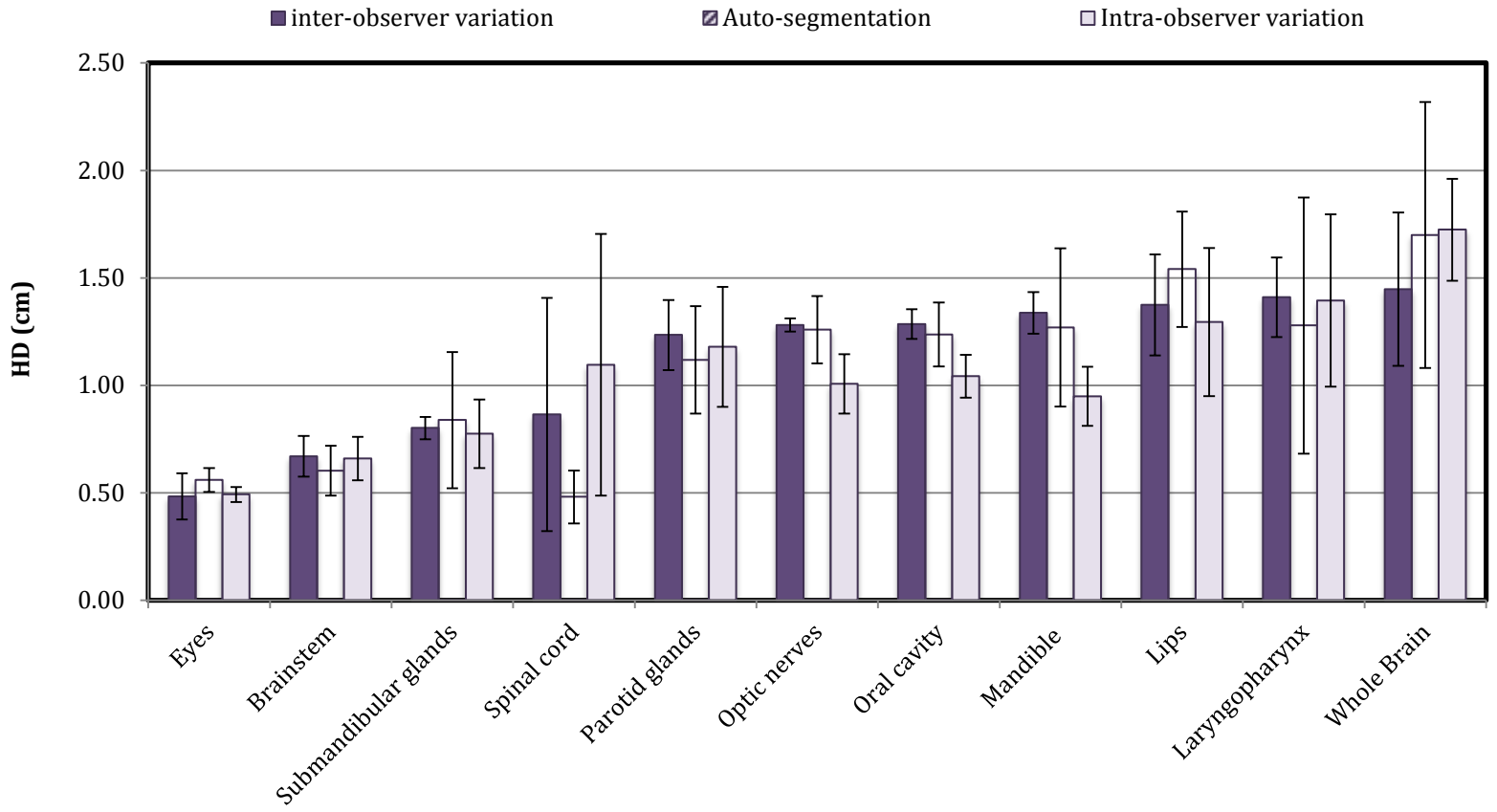


Figure 4. 10 Average HD (cm) over eight test cases, comparing ABAS results with inter-observer and intra-observer variation reported in Chapter 3.

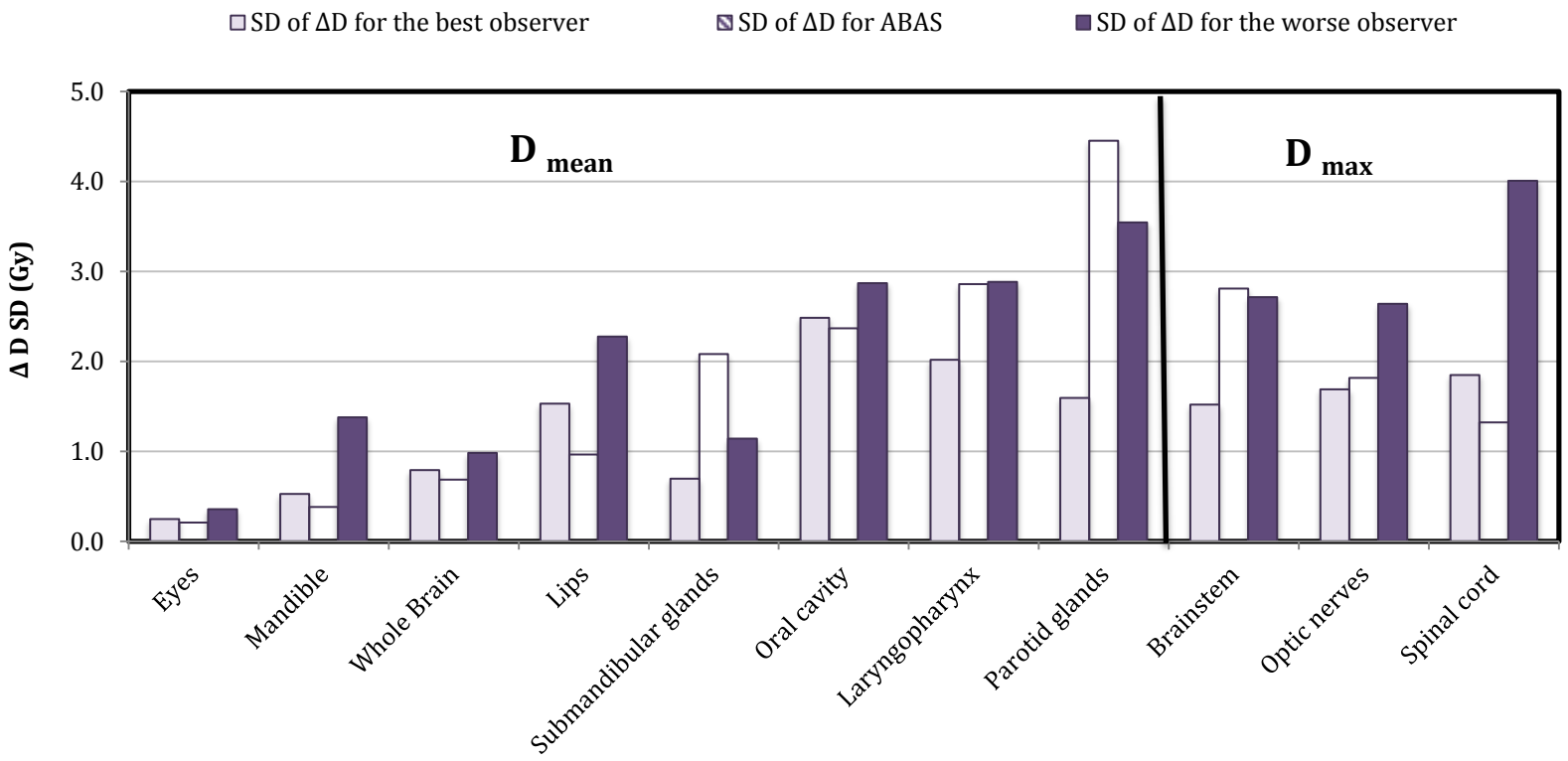


Figure 4. 11 axial view of a test case showing auto-segmentation compared to inter-observer variation and reference contour.

(1) show eyes, optic nerves, whole brain and brainstem, (2) show lips, both parotid glands, spinal cord, oral cavity, and part of mandible, (3) show mandible, both submandibular glands, laryngopharynx and spinal cord for auto-segmentation (yellow), inter-observer (pink) variation and reference (blue) contour.

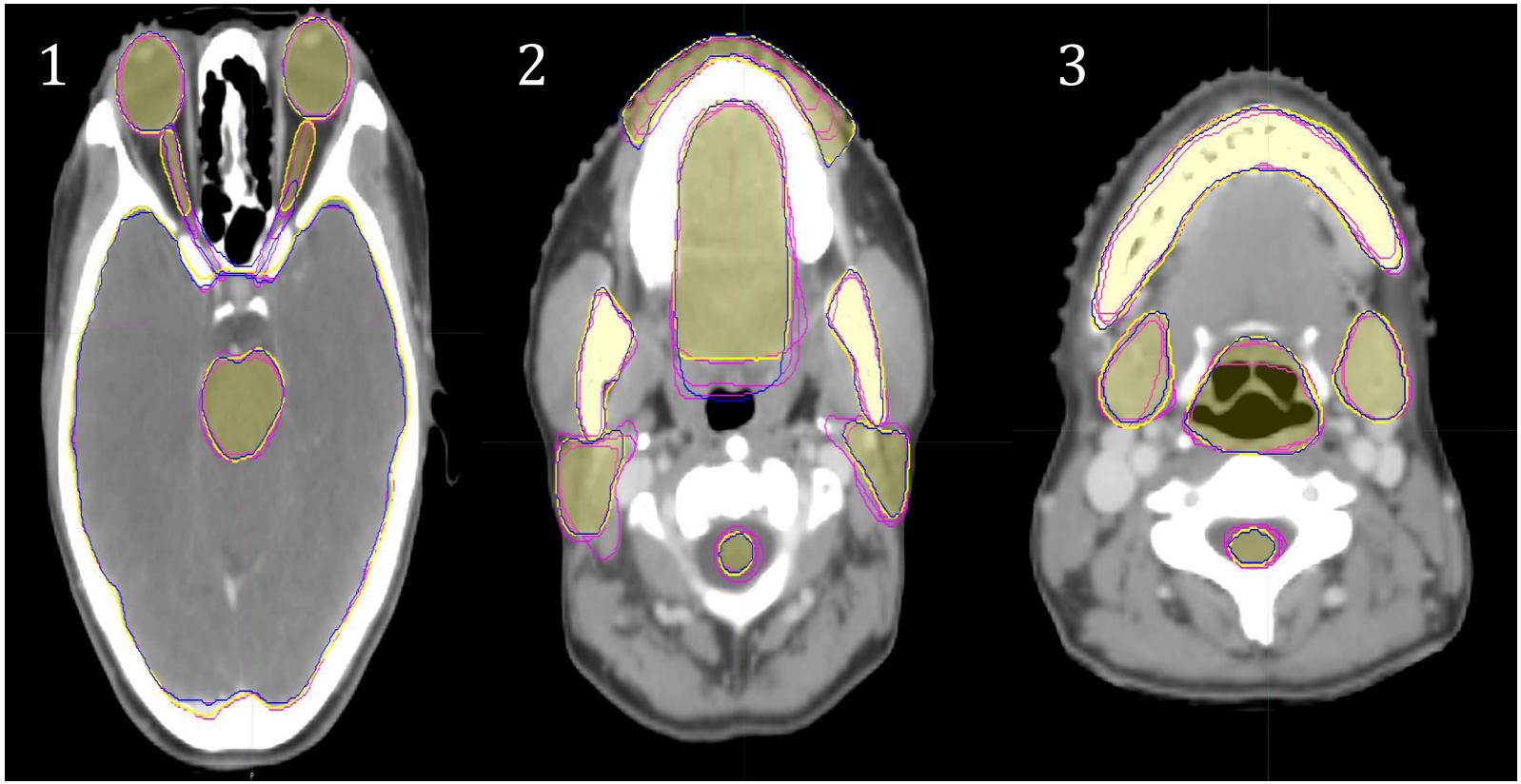


Figure 4. 12 Sagittal (1) and frontal (2) views of a test case showing the longitudinal extent of auto-segmentation compared to inter-observer variation and reference contour.

(1) Shows whole brain, brainstem, lips, mandible, oral cavity, laryngopharynx and spinal cord. (2) Shows whole brain, brainstem and both parotid glands for auto-segmentation (yellow), inter-observer (pink) variation and reference (blue) contour.

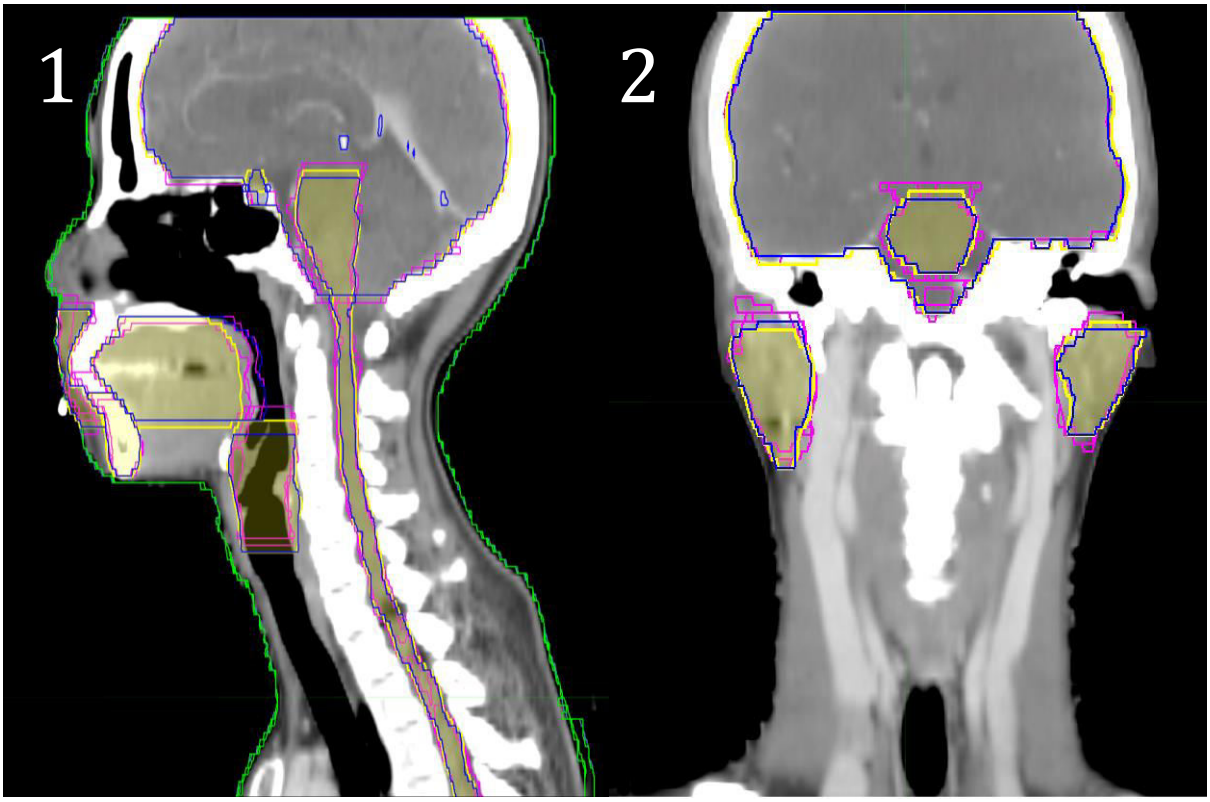
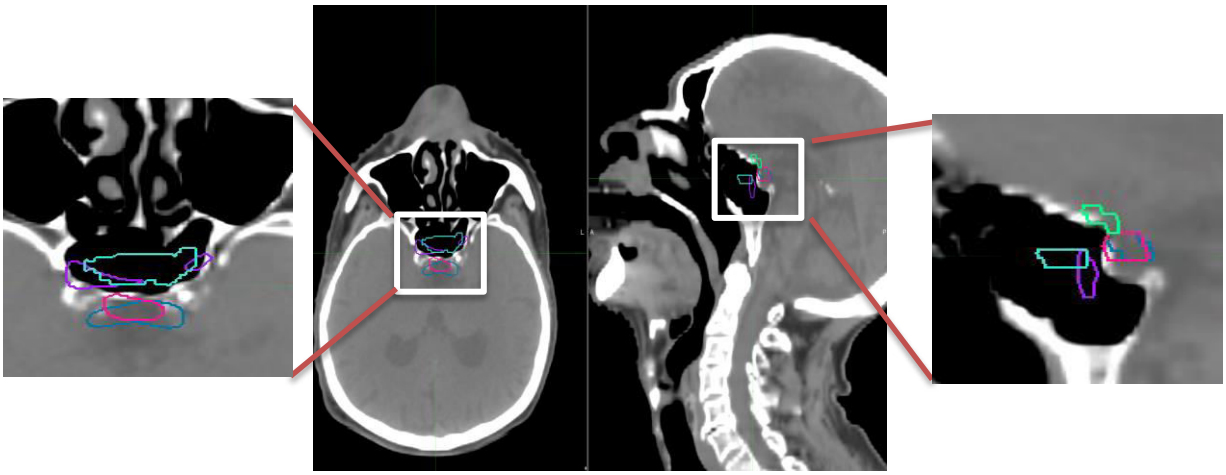


Figure 4. 13 The effect of head tilt angle on optic chiasm auto-contour.

Each outline represents the optic chiasm of one best match. The result of majority overlap equals to zero, so the structure could not be segmented



CHAPTER 5: IMPROVEMENT OF PAROTID GLAND AUTO-DELINEATION USING IN-HOUSE ATLAS BASED AUTO-SEGMENTATION TOOLS

5.1 INTRODUCTION

Parotid glands are the largest salivary glands, contributing more than two-thirds of the salivary output [23]. They are located close to the parapharyngeal space and soft palate, and are likely to receive a significant dose during head and neck radiation treatment. Radiation damage can affect saliva flow by causing xerostomia [26]. The severity of this effect is related to the irradiated volume and the amount of dose received by this volume. Xerostomia has a significant effect on oral health and can cause complications in chewing and swallowing [23,26], resulting in reduction of quality of life.

Parotid-Sparing IMRT demonstrated that significant recovery of saliva flow is achievable in state of the art radiation therapy [11,23]. However, an accurate definition of parotid structures is required to achieve this goal. Despite the importance of accurate segmentation, parotid gland contouring suffers from high inter-observer variation [18,23,26,64]. As mentioned previously, fast and accurate auto-segmentation tools have become the subject of recent research. Studies show significant reduction in contouring time is achievable using the ABAS approach compared to manual segmentation. However, most studies indicate the need for manual modifications [35,36, 40] following auto-segmentation. DSC results for auto-segmented parotid glands are generally poor

compared to DSC results for other organs at risk. After manual editing, DSC results of the edited auto-segmented parotid glands may become similar to manually segmented parotid glands [31, 42, 58], however this negates the time saving achievable using auto-segmentation.

As demonstrated in Chapters 3 and 4, both manual segmentation and auto-segmentation of parotid glands at our institution is more variable compared with segmentation of other organs at risk in head and neck radiation therapy. Our institution has a particular interest in salivary function in radiation therapy for head and neck cancer [Wu et al]. This provided motivation to look further into the problem of accurate auto-segmentation of the parotid gland. At the time of this study, a unique data set consisting of 103 previously treated head and neck cases that had been manually segmented by a single expert radiation oncologist, was available.

The aim of the work described in the current chapter was to improve parotid gland auto-segmentation by using an atlas customized specifically for this purpose using the 103 validated case referred to above. This study examines different parameters that might influence auto-segmentation results such as the rigor used in creating the atlas (the number of subjects and the quality of the segmentation), the number of best matches, and the choice of algorithm for ABAS segmentation. The specific objectives of this study were

1) To gain an understanding of the performance of different user-defined sittings;

2) To confirm if a large, single expert observer ABAS data set could provide better geometric and dosimetric results compared to a small one comprised of data from several observers;

3) To determine how auto-segmentation time is influenced by the number of best matches, and hence the deformation process, versus the number of segmented structures.

5.2 MATERIALS AND METHODS

5.2.1 Atlas construction and parotid volume characterization

Our in-house atlas was constructed for the MIM Maestro™ V 6.4 auto-segmentation tool, using 103 head and neck data sets with parotid gland contours previously segmented by a single radiation oncologist with particular expertise in salivary function. The distribution of parotid gland volumes was characterized to aid in the assessment of ABAS performance as a function of the number of atlas cases in the volume category corresponding to the case being segmented. The parotid volume distribution was divided into three sub-groups: small, intermediate and large parotid volumes. The first group, (small parotid volume), contains subjects with parotid volume < 20 cc (20.4% of the atlas population). The second group, (intermediate parotid volume), includes subjects with the parotid volume between 20 cc and 40 cc (62.1% of the atlas population). The third group, (large parotid volume), contains

subjects with parotid volume > 40 cc (17.5% of the atlas population). Auto-segmentation performance for each group was tested individually.

5.2.2 Atlas performance evaluation versus user-defined settings (the leave-one-out test) and a ABAS Segmentation Quality Index (Q)

The performance of the two available segmentation algorithms (STAPLE and majority vote) was compared. Both algorithms seek a number of best matches from the atlas for the case under study and use these matched cases in the auto-segmentation process after deformable registration is performed. The number of best matches was varied from 3 to 21 to assess the impact of this parameter.

For each volume sub-group, a subject was removed from the atlas and used as a test case. Then, the atlas was used (with the remaining 102 cases) to segment this case several times. This is known as the 'leave one out test'. Segmentations were performed varying the number of best matches first using the majority vote algorithm, then, using the STAPLE algorithm. The test case was then returned to the atlas and another case removed as the process was repeated. Three cases from each volume subgroup were tested. The time for each segmentation process was recorded.

We defined a quality metric, Q, for scoring the test results. By definition, the target values of DSC and HD are 1 and 0, respectively. Thus, Q was defined as:

$$Q = (1\text{-}DICE) + HD,$$

Where Q values closer to 0 indicate higher quality contours. For each volume subgroup, the average Q over the three test cases was reported as the test score. If two

segmentations had the same Q, the one segmented in a shorter time was considered to be the better performer.

5.2.3 Performance comparison with another available ABAS tool and manual inter-observer variation

In order to assess the significance of increasing the number of atlas subjects segmented by a single expert observer, we compared our results with the results generated by the in-house 36-subject atlas used in Chapter 4. Note that the subjects in the 36-subject atlas were delineated in our institution by different radiation therapists.

Using the same methodology and the 8 test cases used in Chapter 4, we compared the auto-segmented parotid results with the manual segmentation variation geometrically, using the percentage of volume variation ΔV (%), DSC and Hausdorff distance (HD), and dosimetrically, using the standard deviation of dose variation.

5.3 RESULTS AND ANALYSIS

The 103 case single expert observer atlas consists of 69% and 31% male and female data sets, respectively. The median age of the atlas population was 57 years (19 – 82 years). As indicated in Table 5.1 and Figure 5.1, parotid volumes incorporated in this study ranged from 8.5 cc to 72 cc with median volume 29.36 cc, 25th and 75th percentiles were 22.17cc and 36.02cc, respectively, and standard deviation 11.41cc.

This represents a slightly right skewed distribution, with a longer tail at the large volume end.

5.3.1 Leave-one-out test of the expert ABAS parotid atlas

The results of the leave-one-out test are shown in Table 5.2 for small, medium and large volume categories. The average DSC and HD over the three test cases for each volume category are reported. It was demonstrated that for the small parotid volume group (1st group) both algorithms showed relatively poor performance described by DSC < 0.8 (ranging between 0.64 – 0.74) and large HD variation ranging between 1.21-1.80 cm. Majority vote indicated overall better performance compared to STAPLE results.

The results of the intermediate parotid volume group (2nd group) showed similar performance for both algorithms. DSC and HD results were very consistent and ranged between 0.8-0.83 and 0.88-1.09cm for DSC and HD, respectively.

The results of the large parotid volume group (3rd group) showed acceptable performance for both algorithms. DSC and HD of STAPLE demonstrated considerably better performance compared to majority vote.

The figures 5.2-5.6 demonstrate the ABAS segmentation quality index scores (Q). Figures 5.2-5.4 show the effect of increasing the number of best matches on the results for the small, intermediate and large groups. Figures 5.5 and 5.6 show the sensitivity of each algorithm on changing the number of best matches.

For the small parotid volume group (figure 5.2), the results showed superior performance of majority vote compared to STAPLE. For all tested numbers of best matches, majority vote segmentation quality remains stable. However, STAPLE showed poorer performance when the number of best matches increased. The best overall performance was observed with 9 best matches and the majority vote algorithm.

For the intermediate parotid volume group (figure 5.3), the group represented with the highest percentage of atlas cases, the results showed similar performance for both algorithms with minimal improvement when using a larger number of best matches. 13, 17 and 21 best matches showed similar segmentation quality, but longer time for a larger number of best matches, as shown in Table 5.3. The best overall performance was observed with 13 best matches and the STAPLE algorithm.

For the large parotid volume group (figure 5.3), the results indicated remarkably better performance of STAPLE compared to majority vote. The quality of the majority vote segmentations remains stable for all tested numbers of best matches. However, STAPLE segmentations showed significant improvement with increasing numbers of best matches. The best overall performance was observed with 17 best matches and the STAPLE algorithm.

Figure 5.5 showed the performance of majority vote for the small, intermediate and large parotid volume groups. This figure indicates that majority vote does not demonstrate high sensitivity to a large number of best matches. Which means that increasing the number of best matches resulted in longer time but the quality of the segmentation remains the same. Further, it showed that the best performance was

observed with intermediate parotid volume whereas the worse performance was observed with the small parotid volume group. These results are very consistent with the distribution of the atlas population. As presented in figure 5.1, 62% of the atlas population represented intermediate parotid volume and the distribution of the population was skewed to the right (limited data for very small volumes and a tail on the large volume side).

Figure 5.6 showed the performance of STAPLE for the small, intermediate and large parotid volume groups. STAPLE demonstrated considerable sensitivity to a large number of best matches for small and large parotid volumes. For both groups, the Q improved by increasing the number of best matches up to a certain point. Increasing the number of best matches (> 9 best matches) significantly improves the segmentation quality for the large parotid volume group but for the small parotid volume, the segmentation quality becomes worse. Overall good quality (small Q) was observed with the intermediate group with minimal dependence on the number of best matches.

This behavior may be partially explained by the sensitivity of the algorithm to the distribution of atlas cases. For each voxel, majority vote includes the voxel if more than half of the input segmentations include the current voxel. Otherwise, that voxel will be excluded. STAPLE computes a probabilistic estimation of the true representation and measures the performance level achieved by each one of best matches. Thus, increasing the number of best matches for a small population, such as the small parotid volume group, would increase the false positive results if all of the best matches lie to one side of the test case within the distribution. For large parotids, despite a relatively small

population, because the distribution has a tail on the large volume end, best matches may be found on both sides of the test case and increasing the number of matches may improve the estimation of the true segmentation.

5.3.2 Geometric and dosimetric comparison with the 36-subjects atlas and manual inter-observer variation

Table 5.4 shows the average percentage of volume variation ΔV (%), DSC and HD over the eight NPC cases. ΔV (%) indicated that 103-subjects ABAS tends to generate larger contours compared to the three manual observers while the 36-subject ABAS tends to delineate smaller volumes. Three different trained therapists segmented all eight test cases. The 36-subjects ABAS segmentations were performed by a number of therapists in a standard clinical practice setting, whereas, as mentioned, a single expert observer contoured the 103-subjects atlas cases. The methodology of the expert observer for defining the parotid was influenced by advanced study of parotid gland anatomy and was expected to be systematically different from the other manual segmentations. The expert tended to segment larger volumes by including the less clearly defined regions of the gland. This segmentation methodology is intended to become our new institutional standard.

DSC results for the 103-subject ABAS were relatively good, with a small improvement in DSC and more consistency compared with the 36-subjects ABAS (0.79±0.04 vs. 0.74±0.08). However, the 103-subjects ABAS contours indicated the largest HD

compared to all the other contours. This is likely due to the systematically larger volumes and the sensitivity of HD to small regions of poor segmentation.

Further, the 103-subject ABAS parotid gland segmentations showed dose variation SD of ΔD (Gy) falling within the SD of ΔD (Gy) among observers (3.10 Gy for 103-subjects ABAS vs. 2.4 Gy, 3.14 Gy, and 1.86 Gy, for observer A, B, and C, respectively). The results showed up to 30% improvement in the SD of ΔD (Gy) when using 103-subjects ABAS compared to 36-subjects ABAS (3.10 Gy vs. 4.41Gy).

5.3.3 Time comparison with the 36-subjects ABAS

As shown in table 5.3, the delineation time is strongly dependent on the number of best matches and the deformation process. It was mentioned in chapter 4 that the time required to segment a complete head and neck OAR structure set, consisting of 16 structures, using five best matches is 2.25 minutes, on average. Our results showed that, using the same workstation, the time required to segment parotid glands only using the same number of best matches was 1.5 minutes, on average, which demonstrates weak time dependency on the number of segmented structures.

Results showed that the segmentation time using a large number of best matches (e.g. 21) was still less than 5 minutes, which is significantly less than the manual segmentation. Thus, if a larger number of best matches provides improved results, ABAS will be a valuable tool to provide both improved segmentation quality and reduced operator workload.

5.4 DISCUSSION AND CONCLUSION

As discussed in chapter 2, in deformable registration applications, the more deformation required, the more degrees of freedom are required to define it, which means a longer time is required to perform the registration [50]. Accordingly, in order to get an acceptable segmentation in shorter time, the atlas images and the test images should be as similar as possible. Hence, increasing the number of subjects in the atlas could increase the possibility of finding better matches cases to the test case.

As presented in chapter 4, ABAS of parotid glands using a generalized head and neck organ at risk atlas with 36 subjects contoured by a group of radiation therapists was associated with smaller DSC compared with manual segmentation by the same staff group. Moreover, the standard deviation of ΔD was the highest for parotid glands compared with all other organs at risk. Several other studies showed inferior DSC results for auto-segmented parotid glands, using different ABAS approaches and software, compared to the other auto-segmented structures. These results improved after manual editing. The reported DSC for parotid glands in different studies (unedited vs. edited ABAS) were as follows: Sims et al. [40] reported 0.69 vs. 0.84, Mattiucci et al. [58] reported 0.74 vs. 0.79 and La Macchia [30] reported 0.79, 0.79 and 0.73 vs. 0.80, 0.82 and 0.81 for three different available software tools.

The results of this study showed superior performance using the103-subject atlas with DSC results close to 0.8, which is considered good agreement in several publications [58]. As well, up to 30% improvement in dose consistency was accomplished using this atlas in comparison to the standard atlas. Thus, the need for

manual modification following auto-segmentation could be minimized using the advanced 103-subject ABAS tool.

We introduced an ABAS segmentation quality index, Q, which combines both DSC and Hausdorff distance to aid in evaluating various segmentation parameters. Scoring this index against the number of best matches for the two different segmentation algorithms provides interesting insight into the how these algorithms behave with respect to the distribution of atlas cases. The data indicate that a value of Q <1.2 represents high quality segmentation. The majority vote algorithm resulted in Q < 1.2 for medium volume parotid glands, which are well represented in the distribution of atlas cases, regardless of the number of best matches. The STAPLE algorithm approaches Q<1.2 for large parotids for large numbers of best matches, likely a result of the skew in the atlas case distribution toward larger volumes. Q for small parotids did not fall below 1.4 due to the fact that small volumes were not well represented in the atlas case distribution.

The results indicate that the 103 subject ABAS volumes were systematically larger than manual segmentation of the test cases. It was confirmed by the expert observer, who delineated the 103 cases, that he tends to segment large volumes and his methodology for defining the parotid will become our new institutional standard. Thus, this atlas will replace the 36-subjects atlas for parotid glands in the clinical workflow.

It is worthy of mention that the trend in saliva recovery research focuses on salivary gland sub-regions responsible for saliva regeneration after irradiation. Van Luijk et al. [67] indicated that stem cells in mice, rats, and humans are located in the major ducts of the parotid gland. They also showed that in rats, the involvement of the ducts in the

radiation field cause a loss of regenerative capacity due to the loss of stem cells, which can lead to long-term gland dysfunction. Miah et al. [68] found that the incidence of high-grade xerostomia is significantly lower in patients who received a bilateral superficial lobe parotid-sparing intensity-modulated radiotherapy (BSLPS- IMRT) technique compared with patients who received contralateral parotid-sparing IMRT (CLPS-IMRT). This was explained by the mean doses delivered to the superficial lobes and the ipsilateral parotid gland. BSLPS-IMRT patients received less mean doses to both the superficial lobes, which minimized the probability of ipsilateral parotid gland damage. A future goal of this auto-segmentation work would be to ensure that the region responsible for saliva recovery is efficiently delineated by ABAS. This could be done by a careful review of the ABAS contours by the expert radiation oncologist.

This work demonstrates that segmentation time shows a high dependence on the deformation process and the number of best matches. The results indicated significant time reduction compared to the protected manual segmentation time, even for a large number of best matches. In fact, the actual elapsed time for manual segmentation requires an even longer time in practice. In our institution, the CT-simulator therapists are assigned to delineate the OARs in parallel with patient scanning. Due to the workload, it takes about 24 hours on average to make the OARs available for the oncologist. Implementing an ABAS tool into the clinical process will reduce the segmentation load on the therapist and make the cases ready to the oncologist, using a clinical workflow, in a much shorter time.

To conclude, using an atlas of a large number of subjects delineated by a single expert observer provided superior results to an atlas comprised of a smaller number of cases segmented by multiple staff members. These results are encouraging and suggest that by using a customized, high quality ABAS tool we may be able to reduce the need for manual intervention and obtain clinical quality auto-segmentations of head and neck OARS.

Table 5. 1 Parotid gland volume characteristics of the **ABAS atlas cases**

Average volume (cc)	30.34
STD	11.41
1st Quarter volume (cc)	22.17
Median volume (cc)	29.36
3rd Quarter volume (cc)	36.02
Maximum volume (cc)	71.76
Minimum volume (cc)	8.45

Table 5. 2 Results of the leave-one-out test for 1st, 2nd and 3rd group (small, intermediate and large volume parotid glands, respectively). DSC represents average DSC coefficient over three test cases and HD represents average Hausdorff distance over three test cases

		1st group		2nd group		3rd group	
Algorithm	# Best matches	DICE	HD (cm)	DICE	HD (cm)	DICE	HD (cm)
Majority vote	3	0.71	1.35	0.8	1.07	0.79	1.62
	5	0.7	1.29	0.81	0.97	0.77	1.7
	7	0.71	1.26	0.8	0.98	0.78	1.54
	9	0.72	1.21	0.8	0.99	0.76	1.68
	13	0.74	1.29	0.81	0.97	0.77	1.59
	17	0.73	1.24	0.82	0.89	0.8	1.52
	21	0.73	1.22	0.82	0.9	0.8	1.58
STAPLE	3	0.64	1.56	0.81	1.02	0.78	1.65
	5	0.7	1.36	0.8	1.09	0.82	1.49
	7	0.71	1.27	0.82	0.93	0.83	1.43
	9	0.71	1.36	0.82	1.03	0.83	1.48
	13	0.68	1.63	0.83	0.9	0.85	1.06
	17	0.67	1.8	0.83	0.88	0.85	1
	21	0.67	1.58	0.83	0.89	0.85	0.99

Table 5. 3 Geometric and dosimetric comparison between parotid gland segmentations using the 103-subject ABAS, 36-subject ABAS and manual observers

	Volume variation (%)		DICE		HD (cm)		Dose variation (Gy)
	Average	STD	Average	STD	Average	STD	STD
Observer A	-5%	17%	0.84	0.04	1.49	0.84	2.4
Observer B	1.3%	25.9%	0.83	0.05	1.26	0.57	3.14
Observer C	10.7%	21.2%	0.85	0.04	1.3	0.52	1.86
36-subjects ABAS	-8.8%	27.1%	0.74	0.08	1.12	0.25	4.41
103-subjects ABAS	21.3%	23%	0.79	0.04	1.71	0.73	3.1

Table 5. 4 The results of leave-one-out test auto-segmentation time

# Of best matches	Average time (hours: min: sec)	SD (hours: min: sec)
3	0:01:03	0:00:06
5	0:01:26	0:00:06
7	0:01:52	0:00:09
9	0:02:19	0:00:07
13	0:03:06	0:00:09
17	0:04:05	0:00:13
21	0:04:49	0:00:13

Figure 5. 1 The distribution of atlas subjects as a function of parotid volume

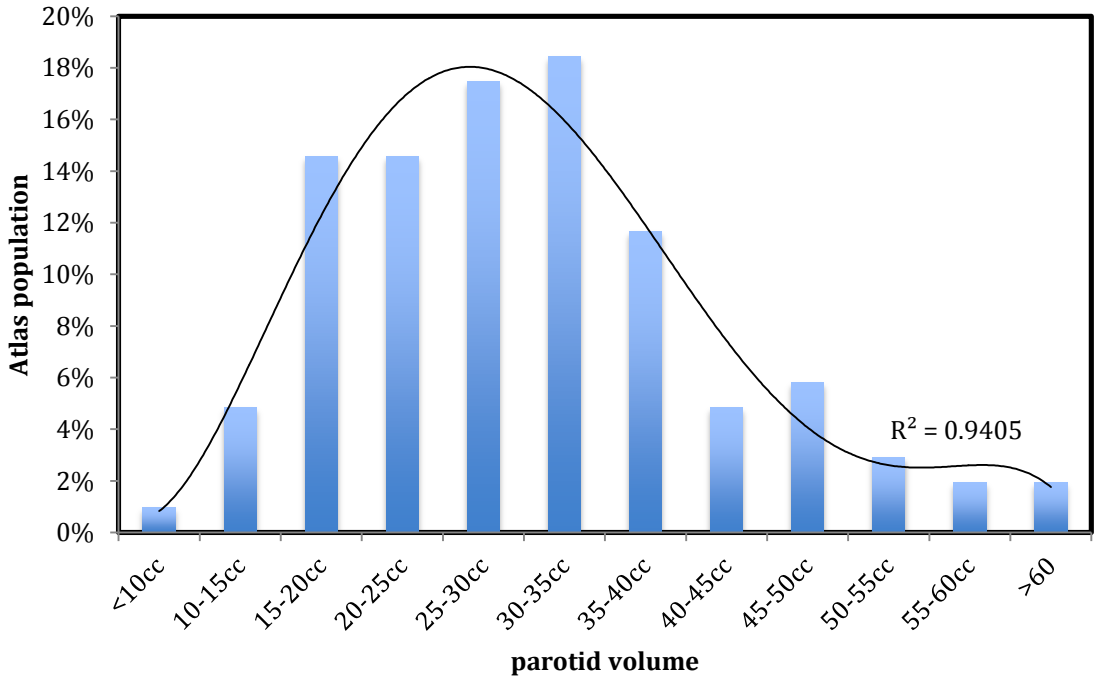


Figure 5. 2 ABAS Segmentation Quality Index (Q) scoring results for Majority vote and STAPLE as a function of number of best matches for small parotid volume.

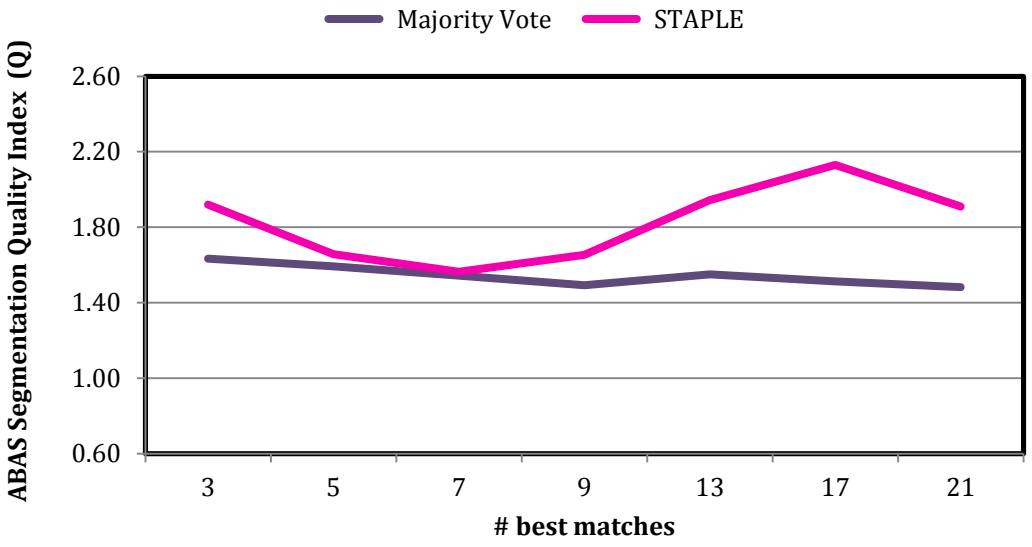


Figure 5. 3 ABAS Segmentation Quality Index (Q) scoring results for Majority vote and STAPLE as a function of number of best matches for intermediate parotid volume.

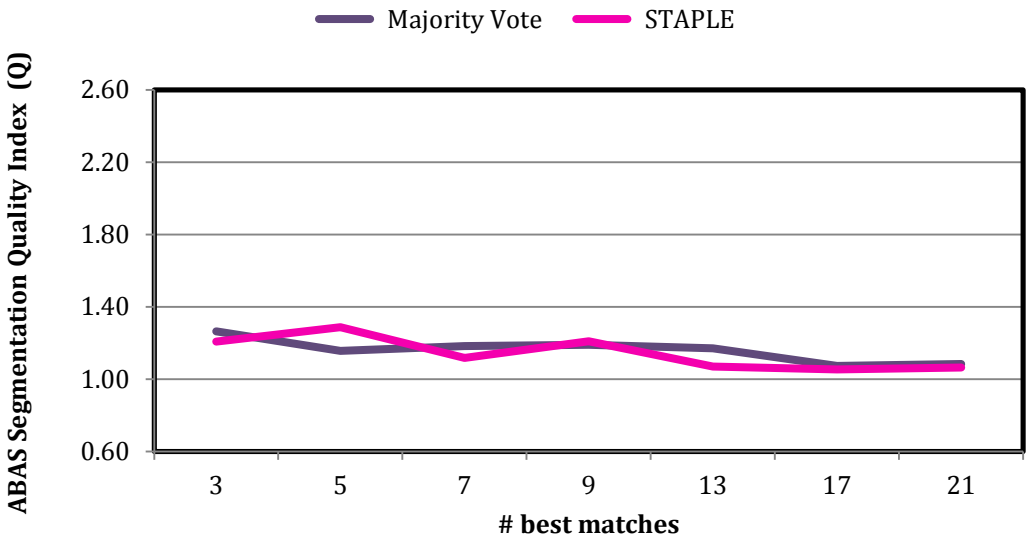


Figure 5. 4 ABAS Segmentation Quality Index (Q) scoring results for Majority vote and STAPLE as a function of number of best matches for large parotid volume.

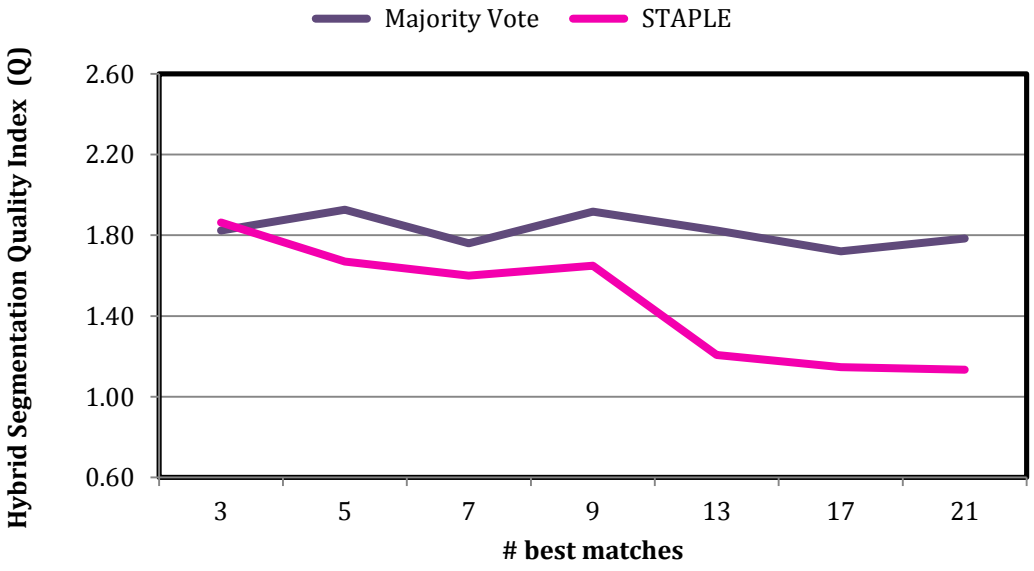


Figure 5. 5 ABAS Segmentation Quality Index (Q) scoring results for Majority vote performance as a function of number of best matches for small, intermediate and large parotid volume groups.

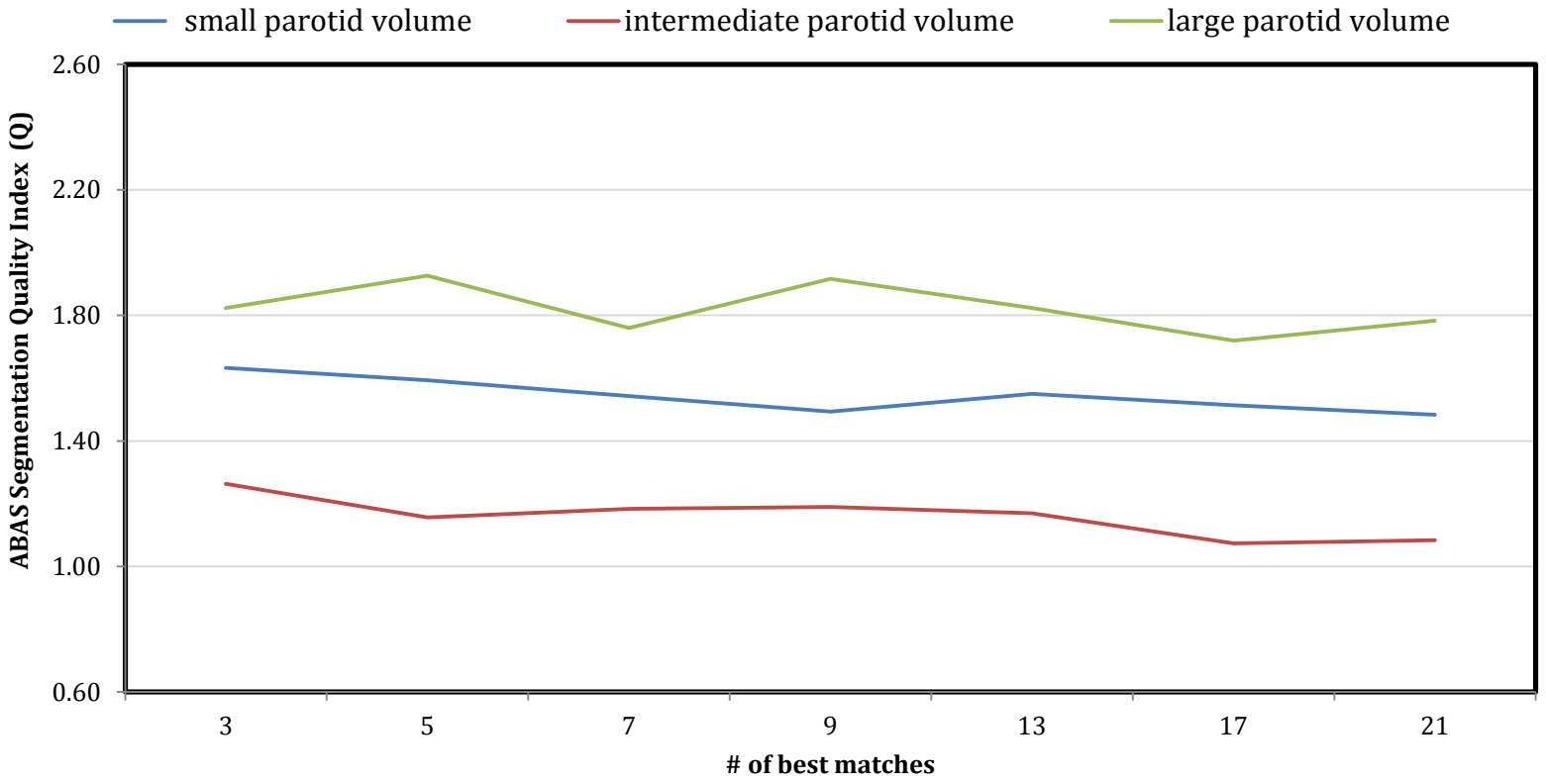
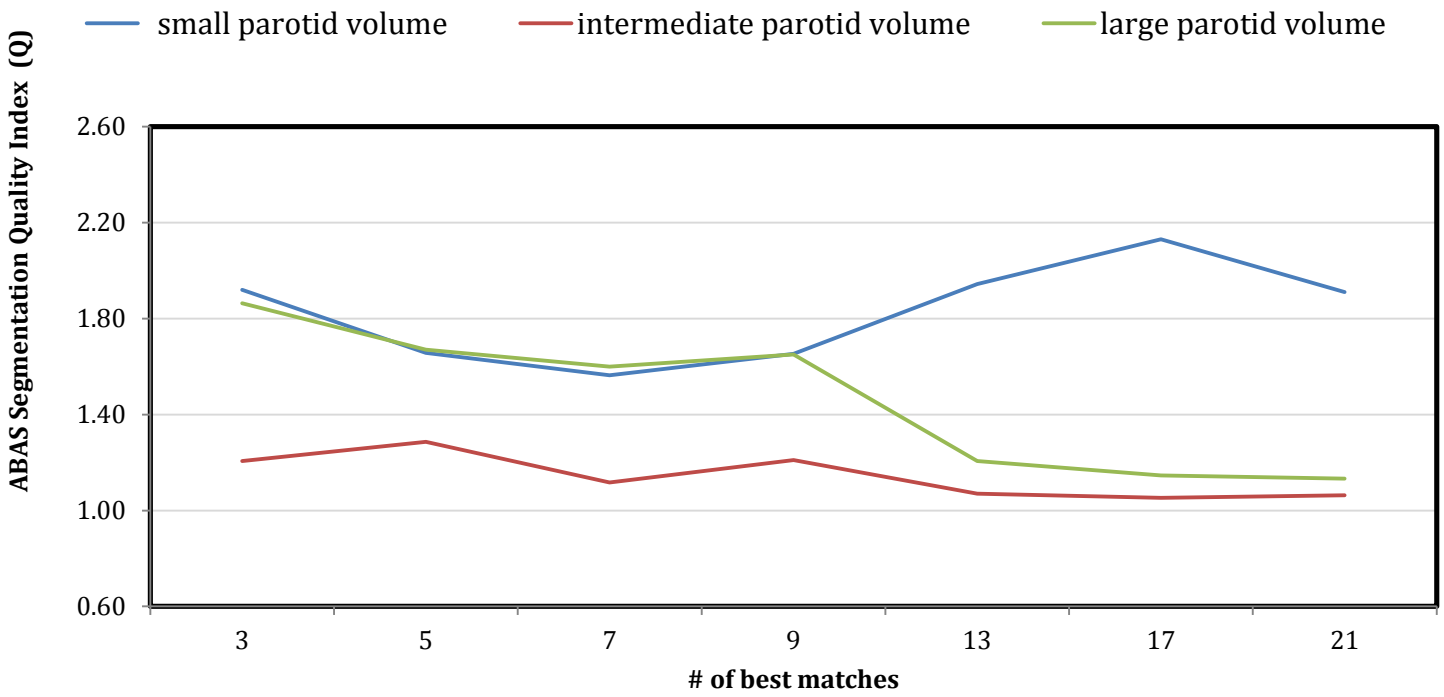


Figure 5. 6 ABAS Segmentation Quality Index (Q) scoring results for STAPLE performance as a function of number of best matches for small, intermediate and large parotid volume groups.



CHAPTER 6: CONCLUSION AND FUTURE WORK

6.1 Conclusion

This project aimed at evaluating and improving auto-segmentation of fifteen head and neck OARs in radiation therapy treatment planning by conducting three studies.

The first study quantified the inter- and intra-observer geometric variation in the manual contours between three observers, and the dosimetric variation resulting from this geometric variation. Further, the results of this study provided organ-specific geometric and dosimetric benchmark data that was used for auto-segmentation evaluation in the second and third studies.

Our results showed significant inter- and intra-observer variation with different levels of variability for different OAR's in the head and neck region. Of these OARs, segmentations of the laryngopharynx, lips and optic nerves were the most variable, whereas whole brain and eyes were the least variable. Inter-observer differences were systematically larger than intra-observer differences for all OARs. The protected time per observer to segment a complete OAR set was almost 30 minutes per patient. The dosimetric evaluation indicated a standard deviation of dose variation (SD of ΔD) up to 4 Gy and 3.5 Gy for the maximum and mean OAR dose, respectively. These findings emphasize the importance of the accuracy and consistency in the delineation of head and neck OARs.

Furthermore, a commonly cited acceptable benchmark geometric assessment in auto-segmentation is a DSC similarity coefficient of 0.8 [58]. It was apparent from our results

that this is not achievable for all OARs, even for the best observers performing manual segmentation. This suggests that an organ-specific DSC benchmark should be used when evaluating the performance of segmentation algorithms or manual observers.

The aim of the second study was to build and evaluate an in-house head and neck atlas for ABAS auto-segmentation of organs at risk. Organ by organ assessment of contouring quality was performed and compared with results for the manual segmentation reported in the first study and other published studies. Also, overall timing for auto-segmentation was performed.

Our in-house ABAS tool was constructed for head and neck cancer patients using 36 consecutive previously treated cases. The results showed acceptable geometrical agreement for most of the delineated structures compared to the benchmark manual segmentation data. More importantly, the dosimetric impact of using this tool on the clinical results was indicated to be within the manual segmentation variation for all the structures, except for salivary glands. It is worth mentioning that these results were obtained from the original auto-segmented structures without any manual modification, and a complete set of OARs was obtained in less than three minutes.

The third study was an investigation into strategies to improve ABAS of parotid glands. Another in-house atlas designed specifically for parotid gland auto-contour consisting of 103 subjects was constructed. A single expert oncologist validated these contours to overcome the impact of inter-observer variation. Different user-defined

settings were evaluated to define the parameters associated with the best performance. The ABAS results were compared geometrically and dosimetrically with those of the first two studies. A new geometric quality index, the ABAS segmentation quality index (Q), combining DSC and HD is proposed in this work. A value of Q <1.2 represents high quality segmentation. The results of the Q evaluation indicate that increasing the number of atlas subjects improves the quality of the segmented structure. It can be inferred that the chance of finding better-matched subjects increases with the number of atlas cases in the vicinity of the test case within the distribution. Additionally, it was demonstrated that STAPLE was more sensitive to the number of best matches than majority vote. When using the STAPLE algorithm, increasing the number of best matches provides better results when the test case is near the center of the atlas distribution (medium volumes) or in the tail of the distribution (large volumes), but not at the lower bound of the distribution, as was the case for small volumes. Majority vote indicated a high correlation between the density of the population around the test cases, but only weak correlation between the number of best matches and the quality of the results. The geometric and dosimetric auto-segmentation results were within the manual segmentation variation.

6.2 Future work

Based on the results of this study, the advanced parotid gland ABAS tool will become the institutional standard for parotid gland segmentation. It is recommended that the

36-subject ABAS tool continue to be expanded by increasing the number and range of atlas cases, for the remaining head and OARs.

Ongoing work aims to assess the ABAS structures qualitatively. Expert radiation oncologists will develop a scoring scheme to evaluate the ABAS results clinically. In this process, the oncologist will review both the manual and the ABAS structures. Both structure sets will be anonymized and presented blindly to the oncologist. This process will help us to understand if the quantitative analysis matches the qualitative one. Also, it will indicate the clinical robustness of the ABAS tool in segmenting head and neck OARs.

These study results encourage us to build a clinical workflow designed for head and neck OARs auto-segmentation for all head and neck cancer patients. An ongoing study aims to minimize the time elapsed for a case to be ready for the radiation oncologist, from patient image acquisition to assigning this case to a radiation oncologist to define the target volumes, using an ABAS clinical workflow.

Furthermore, the impact of automatic post processing of auto-segmented contours has not been assessed. Several tools, such as smoothing and avoidance of bone, exist in the MIM Maestro segmentation software. These may prove useful for further improving the auto-segmentation result of some structures. The dosimetric evaluation could also be more valuable if the auto-segmented structures were used for treatment planning optimization, as this process would better simulate the clinical process.

REFERENCES

[1] Advisory Committee on Cancer Statistics. Canadian Cancer Statistics 2015. Accessed January 7, 2015 at http://www.cancer.ca/~/media/cancer.ca/CW/publications/Canadian%20Cancer%20Statistics/Canadian-Cancer-Statistics-2015-EN.pdf

[2] Małecka–Massalska, T., A. Smoleń, and K. Morshed. "Extracellular–to–body cell mass ratio and subjective global assessment in head-and-neck cancers." *Current Oncology* 21.1 (2014): e62.

[3] Jemal, Ahmedin, et al. "Global cancer statistics." *CA: a cancer journal for clinicians* 61.2 (2011): 69-90.

[4] Bhide, S. A., K. J. Harrington, and C. M. Nutting. "Otological toxicity after postoperative radiotherapy for parotid tumours." *Clinical Oncology* 19.1 (2007): 77-82.

[5] Emami, B., et al. "Tolerance of normal tissue to therapeutic irradiation." *International Journal of Radiation Oncology Biology Physics* 21.1 (1991): 109-122.

[6] Friberg, Sten, and Bengt-Inge Rudén. "Hypofractionation in radiotherapy. An investigation of injured Swedish women, treated for cancer of the breast." *Acta Oncologica* 48.6 (2009): 822-831

[7] Grimm, Jimm, et al. "Dose tolerance limits and dose volume histogram evaluation for stereotactic body radiotherapy." *Journal of Applied Clinical Medical Physics* 12.2 (2011).

[8] Bentzen, Søren M., et al. "Quantitative Analyses of Normal Tissue Effects in the Clinic (QUANTEC): an introduction to the scientific issues." *International Journal of Radiation Oncology Biology Physics* 76.3 (2010): S3-S9.

[9] Kam, Michael KM, et al. "Intensity-modulated radiotherapy in nasopharyngeal carcinoma: dosimetric advantage over conventional plans and feasibility of dose escalation." *International Journal of Radiation Oncology Biology Physics* 56.1 (2003): 145-157.

[10] Bucci, M. Kara, Alison Bevan, and Mack Roach. "Advances in radiation therapy: conventional to 3D, to IMRT, to 4D, and beyond." *CA: a cancer journal for clinicians* 55.2 (2005): 117-134.

[11] Nutting, Christopher M., et al. "Parotid-sparing intensity modulated versus conventional radiotherapy in head and neck cancer (PARSPORT): a phase 3 multicentre randomised controlled trial." *The Lancet Oncology* 12.2 (2011): 127-136.

[12] Otto, Karl. "Volumetric modulated arc therapy: IMRT in a single gantry arc." *Medical Physics* 35.1 (2008): 310-317.

[13] Marur S, Forastiere AA. Head and neck cancer: changing epidemiology, diagnosis, and treatment. Mayo Clin Proc. (2008);83:489-501.

[14] Kan, Monica WK, et al. "A comprehensive dosimetric evaluation of using RapidArc volumetric-modulated arc therapy for the treatment of early-stage nasopharyngeal carcinoma." *Journal of Applied Clinical Medical Physics* 13.6 (2012).

[15] Lee, Tsair-Fwu, et al. "Comparative analysis of SmartArc-based dual arc volumetric-modulated arc radiotherapy (VMAT) versus intensity-modulated

radiotherapy (IMRT) for nasopharyngeal carcinoma." *Journal of Applied Clinical Medical Physics* 12.4 (2011).

[16] Teoh, M et al. "Volumetric Modulated Arc Therapy: A Review of Current Literature and Clinical Use in Practice." *The British Journal of Radiology,* (2011) 84(1007): 967–96.

[17] Anderson, Carryn M., et al. "Interobserver and intermodality variability in GTV delineation on simulation CT, FDG-PET, and MR Images of Head and Neck Cancer." *Jacobs Journal of Radiation Oncology* 1.1 (2014): 006.

[18] Mukesh M, et al. "Interobserver variation in clinical target volume and organs at risk segmentation in post-parotidectomy radiotherapy: can segmentation protocols help?" *The British Journal of Radiology*. (2012);85(1016):e530-e536.

[19] Fotina, I., et al. "Critical discussion of evaluation parameters for inter-observer variability in target definition for radiation therapy." *Strahlentherapie und Onkologie* 188.2 (2012): 160-167.

[20] Berthelet, Eric, et al. "Computed tomography determination of prostate volume and maximum dimensions: a study of interobserver variability." *Radiotherapy and Oncology* 63.1 (2002): 37-40

[21] Thomson, David, et al. "Evaluation of an automatic segmentation algorithm for definition of head and neck organs at risk." *Radiation Oncology* 9.1 (2014):173

[22] Brouwer, Charlotte L., et al. "3D variation in delineation of head and neck organs at risk." *Radiation Oncology* 7.1 (2012): 32.

[23] Loo, S. W., et al. "Interobserver variation in parotid gland delineation: a study of

its impact on intensity-modulated radiotherapy solutions with a systematic review of the literature." *The British Journal of Radiology* (2012);85(1016):1070-1077

[24] Nelms, Benjamin E., et al. "Variations in the contouring of organs at risk: test case from a patient with oropharyngeal cancer." *International Journal of Radiation Oncology Biology Physics* 82.1 (2012): 368-378.

[25] Feng, Mary et al. "Normal Tissue Anatomy for Oropharyngeal Cancer: Contouring Variability and Its Impact on Optimization." *International Journal of Radiation Oncology Biology Physics* (2012) 84(2): e245–49.

[26] Liu, Chengxin et al. 2014. "Error in the Parotid Contour Delineated Using Computed Tomography Images rather than Magnetic Resonance Images during Radiotherapy Planning for Nasopharyngeal Carcinoma." *Japanese Journal of Radiology* 32(4): 211–16.

[27] Daisne, Jean-François, et al. "Tumor volume in pharyngolaryngeal squamous cell carcinoma: comparison at CT, MR imaging, and FDG PET and validation with surgical Specimen 1." *Radiology* 233.1 (2004): 93-100.

[28] Anders, Lisanne C. et al. 2012. "Performance of an Atlas-Based Autosegmentation Software for Delineation of Target Volumes for Radiotherapy of Breast and Anorectal Cancer." *Radiotherapy and Oncology* 102(1): 68–73.

[29] Speight, R et al. 2014. "Evaluation of Atlas Based Auto-Segmentation for Head and Neck Target Volume Delineation in Adaptive/replan IMRT." *Journal of Physics: Conference Series* 489: 012060.

[30] La Macchia, Mariangela et al. 2012. “Systematic Evaluation of Three Different Commercial Software Solutions for Automatic Segmentation for Adaptive Therapy in Head-and-Neck, Prostate and Pleural Cancer.” *Radiation Oncology* 7(1): 160.

[31] Geets, Xavier, et al. "Inter-observer variability in the delineation of pharyngo-laryngeal tumor, parotid glands and cervical spinal cord: comparison between CT-scan and MRI." *Radiotherapy and oncology* 77.1 (2005): 25-31.

[32] Haas, Benjamin, et al. "Automatic segmentation of thoracic and pelvic CT images for radiotherapy planning using implicit anatomic knowledge and organ-specific segmentation strategies." *Physics in Medicine and Biology* 53.6 (2008): 1751.

[33] Walker, Gary V. et al. 2014. “Prospective Randomized Double-Blind Study of Atlas-Based Organ-at-Risk Autosegmentation-Assisted Radiation Planning in Head and Neck Cancer.” *Radiotherapy and Oncology* 112(3): 321–25

[34] Commowick, Olivier, Vincent Grégoire, and Grégoire Malandain. "Atlas-based delineation of lymph node levels in head and neck computed tomography images." *Radiotherapy and Oncology* 87.2 (2008): 281-289.

[35] Stapleford, Liza J. et al. 2010. “Evaluation of Automatic Atlas-Based Lymph Node Segmentation for Head-and-Neck Cancer.” *International Journal of Radiation Oncology Biology Physics* 77(3): 959–66

[36] Tao, Chang-Juan, 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." *Radiotherapy and Oncology* 115.3 (2015): 407-411.

[37] Teguh, David N. et al. 2011. “Clinical Validation of Atlas-Based Auto-Segmentation of Multiple Target Volumes and Normal Tissue (Swallowing/Mastication) Structures in the Head and Neck.” *International Journal of Radiation Oncology Biology Physics* 81(4): 950–57.

[38] Yang, Jinzhong et al. 2014. “Auto-Segmentation of Low-Risk Clinical Target Volume for Head and Neck Radiation Therapy.” *Practical Radiation Oncology* 4(1): e31–37.

[39] Pirozzi, S., et al. "Atlas-based segmentation: evaluation of a multi-atlas approach for prostate cancer." *International Journal of Radiation Oncology Biology Physics* 84.3 (2012): S799.

[40] Sims, Richard et al. 2009. “A Pre-Clinical Assessment of an Atlas-Based Automatic Segmentation Tool for the Head and Neck.” *Radiotherapy and Oncology* 93(3): 474–78.

[41] Voet, Peter WJ, et al. "Does atlas-based autosegmentation of neck levels require subsequent manual contour editing to avoid risk of severe target underdosage? A dosimetric analysis." *Radiotherapy and Oncology* 98.3 (2011): 373-377.

[42] Sharp, Gregory, et al. "Vision 20/20: Perspectives on automated image segmentation for radiotherapy." *Medical Physics* 41.5 (2014): 050902.

[43] Heimann, Tobias, and Hans-Peter Meinzer. "Statistical shape models for 3D medical image segmentation: a review." *Medical Image Analysis* 13.4 (2009): 543-563.

[44] Qazi, Arish A., et al. "Auto-segmentation of normal and target structures in head and neck CT images: a feature-driven model-based approach." *Medical Physics* 38.11 (2011): 6160-6170.

[45] Rohlfing, Torsten, et al. "Quo vadis, atlas-based segmentation?." *Handbook of Biomedical Image Analysis*. Springer US, 2005. 435-486.

[46] Pluim, Josien PW, JB Antoine Maintz, and Max A. Viergever. "Mutual-information-based registration of medical images: a survey." *Medical Imaging, IEEE Transactions on* 22.8 (2003): 986-1004.

[47] Rueckert, Daniel, and Julia A. Schnabel. "Medical image registration." *Biomedical Image Processing*. Springer Berlin Heidelberg, (2010). 131-154.

[48] Fortunati, Valerio, et al. "Tissue segmentation of head and neck CT images for treatment planning: a multiatlas approach combined with intensity modeling." *Medical Physics* 40.7 (2013): 071905.

[49] Collins, D. Louis, et al. "Cortical constraints for non-linear cortical registration." *Visualization in Biomedical Computing*. Springer Berlin Heidelberg, (1996).

[50] Crum, William R., Thomas Hartkens, and D. L. G. Hill. "Non-rigid image registration: theory and practice." *The British Journal of Radiology* (2004);77(suppl_2):S140-S153

[51] Rueckert, Daniel, Alejandro F. Frangi, and Julia A. Schnabel. "Automatic construction of 3-D statistical deformation models of the brain using nonrigid registration." *Medical Imaging, IEEE Transactions on* 22.8 (2003): 1014-1025.

[52] Sørensen, Thorvald. "{A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons}." *Biol. Skr.* 5 (1948): 1-34.

[53] Struikmans, Henk, et al. "Interobserver variability of clinical target volume

delineation of glandular breast tissue and of boost volume in tangential breast irradiation." *Radiotherapy and Oncology* 76.3 (2005): 293-299.

[54] Fox, Jana L., et al. "Does registration of PET and planning CT images decrease interobserver and intraobserver variation in delineating tumor volumes for non–small-cell lung cancer?." *International Journal of Radiation Oncology Biology Physics* 62.1 (2005): 70-75.

[55] Warfield, Simon K., Kelly H. Zou, and William M. Wells. "Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation." *Medical Imaging, IEEE Transactions on* 23.7 (2004): 903-921.

[56] Landis, D. M., et al. "Variability Among Breast Radiation Oncologists in Delineation of the Post-Surgical Breast Lumpectomy Cavity Volume." *International Journal of Radiation Oncology Biology Physics* 63 (2005): S8.

[57] Greenham, Stuart, et al. "Evaluation of atlas‐based auto‐segmentation software in prostate cancer patients." *Journal of Medical Radiation Sciences* 61.3 (2014): 151-158.

[58] Mattiucci, Gian Carlo, et al. "Automatic delineation for replanning in nasopharynx radiotherapy: What is the agreement among experts to be considered as benchmark?." *Acta Oncologica* 52.7 (2013): 1417-1422.

[59] Jameson, Michael G., et al. "A review of methods of analysis in contouring studies for radiation oncology." *Journal of Medical Imaging and Radiation Oncology* 54.5 (2010): 401-410.

[60] Whitfield, Gillian A., et al. "Automated delineation of radiotherapy volumes: are we going in the right direction?." *The British Journal of Radiology* (2013) 86.1021:20110718

[61] Franco, Fernanda Catharino Menezes, et al. "Brachycephalic, dolichocephalic and mesocephalic: Is it appropriate to describe the face using skull patterns?." *Dental Press Journal of Orthodontics* 18.3 (2013): 159-163.

[62] Shah, Twisha, Manish B. Thaker, and Shobhana K. Menon. "Assessment of Cephalic and Facial Indices: A proof for Ethnic and Sexual Dimorphism." *Journal of Forensic Science & Criminology* 3.1 (2015)

[63] Chao, KS Clifford, et al. "Reduce in variation and improve efficiency of target volume delineation by a computer-assisted system using a deformable image registration approach." *International Journal of Radiation Oncology* Biology* Physics* 68.5 (2007): 1512-1521.

[64] O'Daniel, Jennifer C., et al. "Parotid gland dose in intensity-modulated radiotherapy for head and neck cancer: is what you plan what you get?." *International Journal of Radiation Oncology Biology Physics* 69.4 (2007): 1290-1296.

[65] Van de Water, Tara A., et al. "Delineation guidelines for organs at risk involved in radiation-induced salivary dysfunction and xerostomia." *Radiotherapy and Oncology* 93.3 (2009): 545-552.

[66] Tai, Patricia, et al. "Improving the consistency in cervical esophageal target volume definition by special training." *International Journal of Radiation Oncology Biology Physics* 53.3 (2002): 766-774.

[67] Van Luijk, Peter, et al. " Sparing the region of the salivary gland containing stem cells preserves saliva production after radiotherapy for head and neck cancer." *Science Translational Medicine* 7.305 (2015): 305ra147-305ra147.

[68] Miah, A. B., et al. "Recovery of Salivary Function: Contralateral Parotid-sparing Intensity-modulated Radiotherapy versus Bilateral Superficial Lobe Parotid-sparing Intensity-modulated Radiotherapy." *Clinical Oncology* (2016).