THE BENEFIT OF AUTOMATIC SEGMENTATION OF INTRACRANIAL ORGANS AT RISK FOR RADIATION THERAPY: A MULTI-RATER BEHAVIORAL INVESTIGATION

By

Matthew A. Deeley

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Chemical and Physical Biology

August, 2013

Nashville, Tennessee

Approved:

Professor Charles W. Coffey, II (Co-Advisor)

Professor Benoit M. Dawant (Co-Advisor)

Professor Edwin F. Donnelly (Chair of the Dissertation Committee)

Professor David R. Pickens

Professor Ronald R. Price

To my wife, Christina, without whom this work could not have been completed.

To Broderick, Anders, and our growing family, a constant source of inspiration,

and

To my parents for providing me the opportunities they never had.

I am part of all that I have met;
Yet all experience is an arch wherethro'
Gleams that untravell'd world whose margin fades
For ever and forever when I move.
How dull it is to pause, to make an end,
To rust unburnish'd, not to shine in use!
–Alfred, Lord Tennyson

ACKNOWLEDGEMENTS

The work of this dissertation would not have been possible without the support of many people. And while it is not possible to acknowledge everyone in this space, I will for a moment indulge my natural tendency toward long-windedness. Though, as my wife sometimes suggests, I will try to suppress my other tendency toward rambling.

First, I would like to thank my advisors, Charles Coffey and Benoit Dawant, for their continuous support. I owe much to Dr. Coffey, more than I can even begin to enumerate here. He introduced me to medical physics and in many ways has been a professional father. I owe the entire trajectory of my early career to his teaching and his example. I hope to look back many years from now and ask myself whether Charlie Coffey would be proud of the work I have done, and if the answer is in the affirmative, then I will have made a difference. I thank Professor Dawant for the opportunity to grow as a scientist. He welcomed me, by all accounts an outsider physicist, into the lab with exactly the right combination of independence and oversight for me to thrive. I thank him for always making the time for our conversations, whether simple updates or dissecting a paper; these were tremendously useful to my development both in the small things and the big picture. I must thank the National Institute of Biomedical Imaging and Bioengineering and the NIH in general for funding my research assistantship and making possible so much of the science done today.

I thank my committee, Edwin Donnelly, David Pickens, and Ron Price, for their advice and patience over the years. I have been extraordinarily lucky to have had two advisors and a committee with expertise so well matched to my interests. Dr. Price taught my first medical imaging course as an undergraduate and piqued an interest that has remained, and Drs. Pickens and Price have seen me through the masters and doctorate as teachers and committee members. I thank them for sticking with me for so long. I would like to thank Dr. Donnelly for serving as my committee chair and for his continuous involvement and advice concerning this project.

In 2010 as our family grew it became increasingly clear it was going to be difficult for me to remain a full-time graduate student. At the same time the American Board of Radiology was in the midst of changing requirements for board certification of medical physicists. I sought and was offered a position as a clinical and faculty medical physicist in Vermont. This is an unusual situation for a Ph.D. candidate, especially in the physical sciences. Without the coordination and advice of the Chemical and Physical Biology program leadership at Vanderbilt, namely Professors Al Beth, Hassane Mchaourab, and Bruce Damon, my continuation of this work would not have been possible. I owe many thanks to Lindsay Meyers for the literally hundreds of emails and conversations navigating this and other aspects of the graduate school process. I also thank my committee and advisors, particularly Benoit Dawant, for supporting

my vision to grow as a clinician and scientist at the same time.

A number of individuals have been pivotal in my education and development as a scientist, and as this dissertation represents the terminal end of that education (at least formally) it seems appropriate to thank them here.

Professor John Wikswo, a man whose energy and inquisitiveness are unmatched in my experience, introduced me to research and taught me the importance of asking questions.

Professor David Weintraub was a sounding board for me throughout my entire time in Nashville spanning 12 years. I learned much from him about life and academics. His door was always open, and for that I am grateful.

The framework and conclusions of this dissertation have been informed by my clinical background in radiation oncology and medical physics. I had the pleasure of working with Dr. Keith ver Steeg for three years; he has always been a centering force. Dr. Dennis Duggan was involved early in this project and helped get me started. Dr. George Ding and I collaborated on several projects. George challenged my work on a number of fronts, and though we often saw things differently, I know the work benefited much from the constructive criticism. Several attending physicians and resident physicians were part of this project; without their time and effort, the data presented and analyzed herein would not exist.

I would like to thank my colleagues at Fletcher Allen Health Care and the University of Vermont. I was hired as a Ph.D. candidate soon-to-be finished. What I anticipated would be a relatively short process to defense turned into more than two years. Their patience and support has been particularly important to me in completing this work. I especially thank Dr. Yang Cai, James Goodwin, Marleen Moore, and the Chair of Radiation Oncology, Dr. H. James Wallace.

Dr. Thomas Holter has been a friend and source of motivation, encouragement, and advice throughout my masters and doctoral programs. As someone who has training both as a professional and a scholar, his perspective has been invaluable.

F. Richard Knoop has been a continuous positive presence in my life for nearly two decades. First as my teacher and then friend, Rick as always been there. As an alumnus, he is the reason I looked at Vanderbilt for undergrad. I will be forever thankful for his influence and friendship.

Kenneth Lewis has been a friend since the first day of graduate school. We made the joint decision to continue toward the Ph.D. and have reinforced that decision mutually from time to time as we navigated the ups and downs of graduate school. Luckily, our ups and downs were usually out of phase.

I would also like to thank my extended family of the Barretts and the Daudelins.

My mother- and father-in-law have been unbelievably supportive. They never even asked the standard, "So, what are you going to do with that?" My grandparents-in-law have likewise brightened my life in so many ways. And, George Barrett, thank you for all the help with the math.

My parents did not have the opportunity to go to college; they grew up in a very different reality than the one I have known. I will forever be thankful for the support they gave me as a child and the sacrifices they made to afford me opportunities they never had. My grandmother, Dorothy, who turns 89 as I submit this dissertation, was perhaps the first academic role model of my life. A single mother of ten, she attended college in her 50s and taught me that you make your own path.

No one has been more important to my sanity and the completion of this work than my wife, Christina. When we met nearly seven years ago, I was only three months into the process of the Ph.D. She could not have known, and in many ways neither could I, the challenges that lie ahead. But, she has been there to motivate, listen, and in general support me throughout it all. As our family grew Christina was always there to fill the gaps created by my late nights and seven-day work weeks. This work could never have been completed without her, and for that I will always be grateful.

Lastly, I must thank my little boys; their smiles and laughs pulled me through the toughest of times.

TABLE OF CONTENTS

Page

DEDICATION iii

ACKNOWLEDGEMENTS iv

TABLE OF CONTENTS vii

LIST OF TABLES x

LIST OF FIGURES xi

I INTRODUCTION 1

I.1 The reliance of modern radiation therapy on image segmentation . . . 2

I.2 Image registration . . . 3

I.2.1 The adaptive bases algorithm . . . 4

I.3 Medical image segmentation . . . 7

I.4 Challenges of and approaches to segmentation in the brain . . . 7

I.4.1 Atlas-based segmentation of the brainstem and eyes . . . 8

I.4.2 Model-based segmentation of the optic chiasm and optic nerves . . . 12

I.5 Evaluative framework . . . 14

I.6 Goals and contributions of the work . . . 19

II CHARACTERIZATION OF SEGMENTATION VARIANCE 21

II.1 Introduction . . . 23

II.2 Methods . . . 25

II.2.1 Study design . . . 25

II.2.2 Automatic segmentation . . . 26

II.2.3 Calculation of simulated ground truths . . . 27

II.2.4 Comparison metrics . . . 30

II.3 Results . . . 31

II.3.1 Volume . . . 35

II.3.2 Dice similarity coefficient . . . 36

II.3.3 Euclidean distance . . . 39

II.3.4 Time . . . 41
II.4 Discussion . . . 41
II.4.1 Limitations and future work . . . 46

III IMPACT OF EDITING ON SEGMENTATION VARIANCE AND ACCURACY 48

III.1 Introduction . . . 50
III.2 Methods . . . 52
III.2.1 Study design . . . 52
III.2.2 Automatic segmentation . . . 54
III.2.3 Ground truth estimation . . . 56
III.2.4 Metrics for comparison . . . 57
III.2.5 Analytical framework . . . 59
III.3 Results . . . 59
III.3.1 Assessing editing efficiency . . . 59
III.3.2 Evidence regarding hypothesis: Editing of automatic segmentations (A_1) reduces inter-rater variance . . . 65
III.3.3 Evidence regarding hypothesis: Editing of automatic segmentations (A_1) maintains or improves accuracy . . . 65
III.3.4 Evidence regarding hypothesis: Editing of automatic segmentations (A_1) salvages the results of low performing raters . . . 69
III.3.5 Evidence regarding hypothesis: Contour editing reduces inter-rater variation while maintaining or improving accuracy irrespective of the source segmentation . . . 70
III.4 Discussion . . . 74
III.4.1 Comparison to previous studies . . . 75
III.4.2 Limitations and future work . . . 79

IV DOSIMETRIC IMPACT OF AUTOMATIC SEGMENTATION 85

IV.1 Introduction . . . 85
IV.2 Materials and Methods . . . 88
IV.2.1 Segmentation . . . 88
IV.2.2 Treatment planning . . . 88
IV.2.3 Data analysis . . . 89
IV.3 Results . . . 92
IV.3.1 Impact on target coverage . . . 92

IV.3.2 Impact of segmentation on plan quality as measured by dose to ground truth . . . 96
IV.3.3 Discrepancy in dose reporting . . . 96
IV.4 Discussion . . . 99
IV.4.1 Limitations and future work . . . 102
IV.5 Conclusions . . . 103
IV.6 Supplemental Material . . . 103

V DISCUSSION AND FUTURE DIRECTIONS OF RESEARCH 108

BIBLIOGRAPHY 112

LIST OF TABLES

Page

II.1 Mean DSC, 95% CI, and standard deviation 40
II.2 Distance errors: minimum, mean, and maximum 42
II.3 Time to contour: mean and standard deviation 44

III.1 Amount of editing required as function of source 62
III.2 Accuracy as a function of source by rater and structure 67
III.3 DSC for brainstem and chiasm 81
III.4 DSC for eyes and optic nerves 82
III.5 Volume as a function of source 83
III.6 Distance errors as a function of source 84

IV.1 Doses to targets 92
IV.2 P-values from Friedman's test 95
IV.3 P-values Wilcoxon sign-rank tests 95
IV.4 Discrepancies in dose reporting 99
IV.5 Dosimetric figures of merit for brainstem 104
IV.6 Dosimetric figures of merit for the optic chiasm 105
IV.7 Dosimetric figures of merit for the eyes 106
IV.8 Dosimetric figures of merit for the optic nerves 107

LIST OF FIGURES

Page

I.1 Non-rigid registration using the adaptive bases algorithm 5
I.2 Brainstem: imaging sections of medulla 8
I.3 Brainstem: imaging sections of the pons 9
I.4 Brainstem: imaging sections of the midbrain 10
I.5 Eyes: imaging sections of the eye 13
I.6 Imaging sections of the optic nerve 15
I.7 Imaging sections of the optic chiasm 16
I.8 3D rendering of segmentations 17

II.1 Atlas-based segmentation process for the brainstem and eyes 28
II.2 A qualitative example chosen from the 20 patients 32
II.3 Axial slice showing an area of high physician variability 33
II.4 Volume by rater and structure 34
II.5 Inter-rater variance (DSC) 37
II.6 Accuracy (DSC) by rater and structure 38
II.7 Signed distances errors by rater and structure 39
II.8 True positive rate by rater and structure 43

III.1 Ground truth estimation 55
III.2 Impact of editing on efficiency 60
III.3 DSC as a function of source segmentation 61
III.4 Qualitative results of editing 63
III.5 DSC editing A_1 by rater and structure 64
III.6 Volumes pre- and post-editing by rater and structure 66
III.7 Accuracy pre- and post-editing as a function of source 67
III.8 Signed distance errors 68
III.9 Inter-rater variability (DSC) as a function of source by rater and structure . . 70
III.10 Distance error as a function of source 71
III.11 True positive rate as a function of source 72
III.12 Scatter plots of DSC pre- and post-editing 74

IV.1 Average dose-volume histograms 93
IV.2 Average dose-volume histogram showing upper 25% of dose 94
IV.3 Maximum dose differences from peers 97
IV.4 Volume dose differences from peers: brainstem 98
IV.5 Maximum dose reporting differences against DSC 100

CHAPTER I

INTRODUCTION

The following dissertation presents an investigation of image segmentation with a focus on the intracranial organs critical to radiation therapy of the brain. Almost all forms of modern radiation therapy planning now rely on three dimensional imaging and subsequent segmentation, or partitioning, of the the images into important anatomical regions. These regions have traditionally been segmented manually and sometimes painstakingly on a slice-by-slice basis by medical professionals, often radiation oncologists. Over the past decade algorithms have been developed and quickly implemented clinically to segment some of these regions. The scope of algorithm development and rapidity of clinical implementation have generally far exceeded evaluative work to assess the potential impact of these algorithms.

A central theme of the three studies that comprise the bulk of this dissertation is that of a behavioral focus. This perhaps requires some qualification as such terminology is encountered more often in the social and cognitive than the physical sciences. Segmentation in the context of radiation therapy has been a process of human perception. Even the automated methods rely on atlases and models that are derived from initial conditions provided by humans. As such this work has been motivated by the desire to characterize the impact of automatic segmentation in the context of human decision making and interaction therewith.

In the remainder of this chapter, we introduce several concepts and operational definitions important to the research studies that follow. Chapter II presents a study assessing the variability and accuracy of the automatic system and human raters *de novo* (from scratch), chapter III gauges the impact of segmentation editing on quality and efficiency, and in chapter IV we investigate the dosimetric implications of segmentation variability in the context of inverse-optimized radiation therapy planning. These studies have taken place at the intersection of medical image processing, medical physics, and clinical radiation oncology. Throughout this work we use graphical methods as a tool of relating key statistical information, though when necessary we resort to more formal statistical tests. In chapter V, we discuss the main conclusions and contributions of the present work and possible directions of future work.

The format is that of a collection of papers, of which the first two (Deeley et al., 2011, 2013) have been published in *Physics in Medicine and Biology* and the third is in preparation for submission. A natural consequence of this format is a degree of overlap between the dissertation

and individual paper (chapter) introductions, though the former develops the requisite topics more broadly. For clarity I use the active voice where possible and for consistency with the published work contained herein, generally employ the first person plural. Though this work has been very much a personal odyssey, and while the design, methods, and conclusions are my own, these have resulted from interaction with numerous colleagues and coauthors.

I.1 The reliance of modern radiation therapy on image segmentation

Approximately 41 percent of Americans born today as predicted by the National Cancer Institute statistics review from 1975 to 2010 (Jemal et al., 2013), or nearly one in two men and women, will be diagnosed with cancer during their lifetime, and nearly two-thirds of these patients will receive radiation therapy. The use of radiation to treat disease has always been driven by innovation, as evidenced by the reports of its clinical use coming by Freund (Božica and Bojana, 2010) in 1896 to treat hairy nevi (moles) and rival claims to the first treatment of cancer by Grubbe, Despeignes, Williams, and Voigt in 1897 (Hall, 1994), all within two years of the discovery of x-rays by Roentgen. Soon thereafter came the discovery of radioactivity by Becquerel and Marie and Pierre Curie, followed in the mid-century by the translation of a physics research tool, the particle accelerator, into the medical linear accelerator which today is used as the primary modality for treatment of cancer with radiation.

Radiation therapy of the late twentieth and early twenty-first century has been heavily influenced by advances in two areas in particular: incorporation of medical imaging into treatment planning and the ability to modulate the intensity of the radiation beam. The availability of computed tomography (CT) images led to new treatment planning systems that could use this vastly better geometric information as well as the inherent density information to improve disease localization and dose calculation accuracy, leading to what is now known as three-dimensional treatment planning (Aird and Conway, 2002; Driver et al., 2004). With the incorporation of CT images initially and later magnetic resonance (MR) and positron emission tomography (PET) and other physiological imaging, there existed a need to segment individual anatomical volumes within the imaging space. These segmentations, also referred to as contours, have become a vital component of all definitive radiation therapy. Their use is at least two-fold: they provide for quantitative assessment of the dose distribution with regard to the targeted areas as well as the normal tissues (also known as the *organs-at-risk* or *critical* structures), and they also can be used to generate treatment apertures conforming geometrically to the targets while avoiding the the critical structures. It is the combination of the two uses that come together in inverse-treatment planning, principally intensity-modulated radiation therapy (IMRT), wherein plans are generated via optimization algorithms that accept as an inputs

the target and critical organ segmentations along with a set of dosimetric goals and relative priorities.

Nearly simultaneous to the broader incorporation of imaging, techniques for radiation beam collimation, first binary and then multileaf (MLC), were being improved dramatically. Whereas prior techniques required use of manually cast heavy metal alloys to shape the beam, the binary collimators and MLCs could be motor-driven and electronically controlled and monitored to produce dramatically more apertures than previously possible. It is the advances in collimation along with the incorporation of imaging that spurred IMRT, and in turn, the need for segmentation.

Other, even more recent developments are increasing segmentation workload. In the past decade, linac-integrated kilo-voltage cone-beam CT (CBCT) (Jaffray et al., 1999, 2002) and to a lesser extent conventional CT-on-rails (Cheng et al., 2003; Shiu et al., 2003), mega-voltage tomographic imaging (Mosleh-Shirazi et al., 1998; Ford et al., 2002), and 3D ultrasound (Bouchet et al., 2001; Tome et al., 2002; Molloy et al., 2011) have been become common place. There is also much research on-going to integrate magnetic resonance imaging (Lagendijk et al., 2008; Fallone et al., 2009) within linear accelerator platforms. This new imaging capability is used both to guide the alignment of the patient for daily treatment and to provide information about soft tissue changes over the treatment course. If a patient looses or gains weight, or the tumor grows or shrinks, the treatment can be adapted accordingly (Hansen et al., 2006; Ding et al., 2006; Schwartz and Dong, 2011; Gregoire et al., 2012; Jensen et al., 2012; Peroni et al., 2012; Schwartz et al., 2013). Each of these emerging technologies produces volume images that require segmentation if they are to be used in either the planning or re-planning process. Though it is time- and labor-intensive, this has been done manually for the most part. Some anatomical sites experience potentially important changes inter- and even intra-fraction. For example, with the male pelvis the bladder, rectum, and prostate can change significantly between and potentially during treatments (van Vulpen et al., 2008). If treatment adaptation is to be extended to its logical end, that is, daily online adaptation, high quality automatic segmentation must be a prerequisite.

I.2 Image registration

Image registration is the determination of a mapping between points in a view of an object to corresponding points in another view of that object or a different object. In other words, it is the process of aligning images so that corresponding features can be easily related (Hanal et al., 2001). In the past 20 years image registration as a field of study has emerged from a position of obscurity to be a major contributor to image processing research. In their

editorial on the rise of image registration, Pluim and Fitzpatrick (Pluim and Fitzpatrick, 2003) found that yearly publications increased from 10 in 1985 to approximately 140 in 2002 (PubMed search terms "image AND registration"). A current search produced 846 publications for the year 2012. Likewise, those authors found 34 papers in 2002 for non-rigid registration, the application of which is central to our segmentation methods, and a current search results in 191 papers published in 2012 alone. These increases have paralleled the rapid rise of new imaging modalities and their integration into the clinical workflow.

The field of image registration can be parsed in terms of transformations and the image properties conserved therewith. Two categories are "rigid" and "non-rigid" registrations; the former category being singular and the latter extremely broad. Rigid registration has the stringent requirement that when registering a reference to a target image, the transformation applied conserves distance. That is, the spatial distance between points x and y in the reference image must be preserved in the reference image transformed. Purely rigid transformations are often too restricting, even for intra-subject registration between different modalities such as CT to MR. Non-rigid methods come in many flavors as they include all transformations save rigid, and the degree of conservation varies widely. The family of projective transformations, including affine, scaling, and similarity, and rigid, preserve straightness of lines and planarity of surfaces. In fact, rigid transformations are a special case of the affine class. Affine transformations preserve collinearity (parallelness) of lines with 12 degrees of freedom: 3 for translation and 9 for rotation, scaling and and shearing. The rigid transformation restricts this to 6 degrees of freedom allowing only translation and rotation.

I.2.1 The adaptive bases algorithm

The transformations central to this work are affine and curved (non-projective) non-rigid. Our goal is to register a source image volume $S(x)$ to a target (patient) volume $T(x)$. We initialize the non-rigid registration with a mutual information-based global and then local affine registration. The local region of interest is determined by the global registration and a predefined bounded region in the atlas images. There are a number of methods for registering the volumes non-rigidly, as evidenced by the surge in non-rigid publications in recent years. The adaptive bases algorithm (ABA) (Rohde et al., 2003) solves an optimization problem in which the source image is best matched to the target image. This method was developed at Vanderbilt and is expressed mathematically (for images in 3-D) as

$$\arg\max_{x'} \; F(S(x'), T(x), x')$$

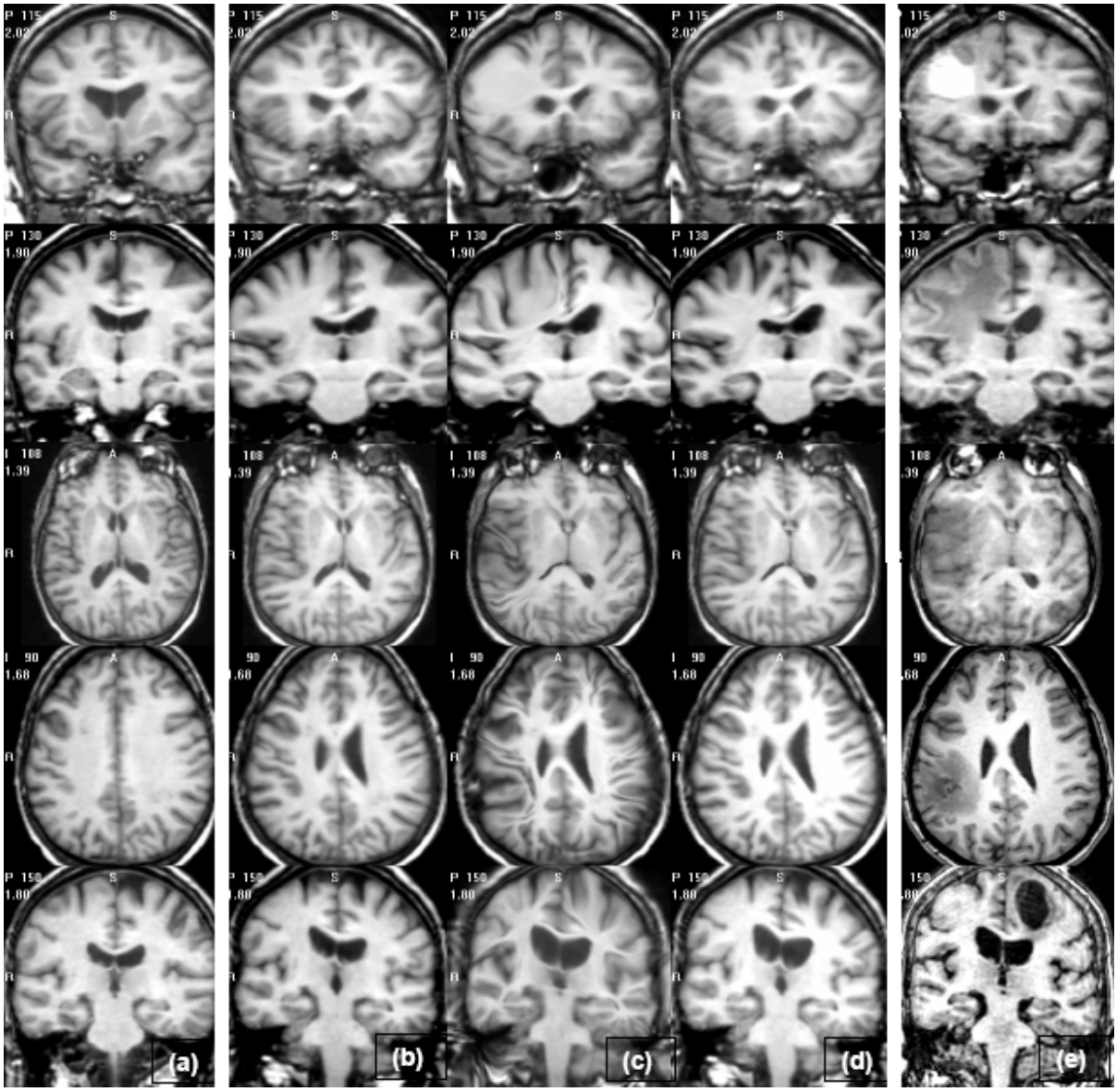

Figure I.1: Registration of source (atlas) and target (patient) MR images. Each row represents a different inter-subject registration set; source images in column (a) are transformed to patient images in column (e). The results of registering (a) to (e) are contained in the intervening columns; (b), (c), and (d) have been transformed non-rigidly using ABA with varying elasticities, stiff, elastic and spatially varying (mixed), respectively. [Images courtesy of the Vanderbilt Medical Image Processing lab/Natalie Zhaoying Han]

where $x' = x + v(x)$ and F is an intensity-based similarity measure (normalized mutual information (Studholme et al., 1999) in our case), x is a coordinate vector, and $v(x) = \{v_x(x), v_y(x), v_z(x)\}$, a deformation field that transforms image $S(x)$. The deformation field $v(x)$ is the final result of the registration.

Several aspects of ABA are beneficial in venues such as ours where $T(x)$ may contain large pathological features not present in $S(x)$. First, it reduces the complexity of the optimization problem by using compactly supported radial basis functions (Wu, 1995) on an irregular grid. Some other methods such as spline-based have typically modeled the deformation field on a regularized, or spatially invariant, grid requiring a large number of elements. ABA approaches the problem by building the deformation field incrementally over a number of scales and resolutions. Scale refers to the size of the basis functions used to model the transformation and resolution refers to the resolution of the image. The process begins at low resolution with perhaps only a few basis functions of large scale. As the algorithm moves from one level of resolution and scale to another, the basis functions are first temporarily placed on a regular grid. The areas of misregistration are then determined by computing the gradient of the cost function. The idea is that if the gradient at a specific location is large, then the cost function is not at a minimum and the registration of this region could be improved. A small gradient indicates a local extremum in which case either the images are reasonably well registered or, alternatively, they are not well registered but will most likely not benefit significantly from further optimization. The final deformation field $v(x)$ is a linear combination of a set of irregularly spaced basis functions:

$$v(x) = \sum_{i=1}^{N} c_i \Phi(x - x_i).$$

where c_i is the coefficient of each basis function, $\Phi(x)$.

An issue shared among curved transformations is the preservation of topological correctness; that is, the absence of nonphysical tearing or folding of the source image. ABA constrains the coefficients to a predetermined upper limit at each level of optimization that forces the deformation to build in a topologically sensible way. Lastly, in situations such as our patients with large brain tumors, a stiffness map may be specified in the atlas image to allow a spatially varying degree of elasticity. We largely circumvent the need for this in tackling segmentation of structures locally. That is, for our purposes, we do not need to apply ABA globally. Figure I.1 presents the results of registering several non-pathologic source (atlas) images to target (patient) images.

I.3 Medical image segmentation

The objects on which measurements are made throughout this work are known as segmentations. It is a concept encountered frequently in daily life and image processing research. In its simplest form segmentation can be defined as a process of classification of an object into categories reflective of intent. Image segmentation is a rich area of investigation extending far beyond the medical applications considered in this work, to fields such as facial recognition, remote sensing, studies of perception, and pattern analysis (Martin et al., 2001) to name a few. Goshtasby (Goshtasby, 2005) suggests it may be the most studied area of image analysis.

In this work we consider image segmentation as the classification of 3D images via voxel ownership into volumes of interest. These volumes of interest are commonly referred to as "labels" in the image processing literature, but we will generally use the term "segmentation". A voxel may be marked as belonging wholly to a specific segmentation or not, or it may be given a partial volume. In the former case, the segmentation is characterized via a binary classifier such that each voxel is assigned $\{0, 1\}$, whereas a partial volume segmentation may contain voxels of values on the continuous range $[0, 1]$ (Crum et al., 2006).

In our work we consider only binary segmentations. We are concerned with segmentation of the normal tissues of the brain, particularly the brainstem, optic chiasm, eyes and optic nerves. These segmentations arise from two methods: manual segmentations produced by human raters (physicians) and automatic segmentations from our computer algorithms. The manual raters view fused CT and MR data in a clinical treatment planning system. Using a mouse and various software tools available, through a series of mouse clicks or free-hand tracing of the mouse, a smooth contour overlays the image. In our situation the human raters could not form contours in arbitrary imaging planes; rather, only the native axial plane of the CT images was available. The resulting contours are exported from the treatment planning system as a series of ordered coordinate sets $\{x_i, y_i, slice_n\}$ representing closed contour loops in CT-space.

I.4 Challenges of and approaches to segmentation in the brain

Anatomical sites differ widely, which presents a challenge in designing accurate and robust methods for segmentation. A primary consideration is whether the anatomy of interest is defined explicitly or implicitly. Explicit anatomy is that which is well-defined by gross structures. The bladder, for instance, is a very well-defined organ enclosed by a membrane separating it from surround tissue. Other anatomy is defined implicitly, such as the lymph nodes in the neck. While the nodes are individually defined, the chains of nodes are generally defined for the purpose of radiation therapy as broad regions including intervening tissues,

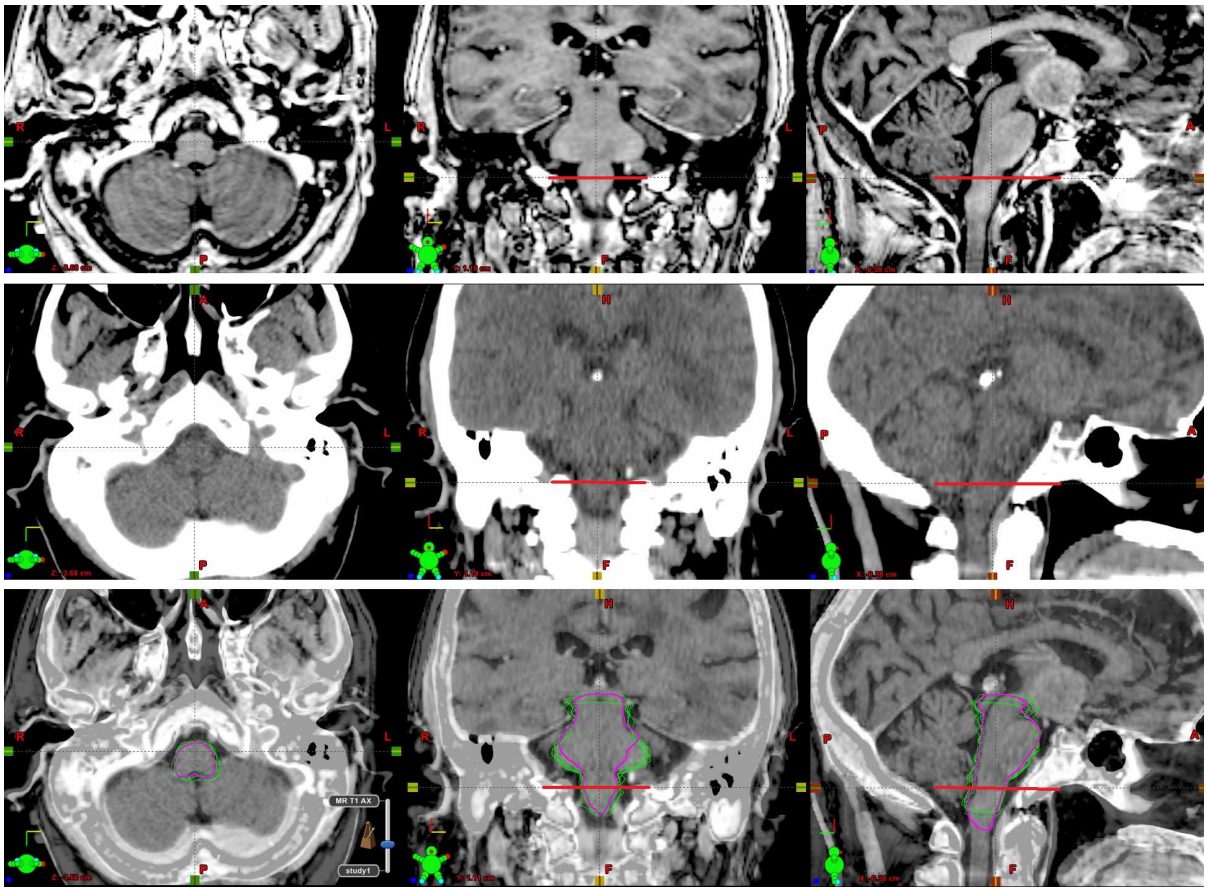

Figure I.2: Brainstem: medulla. MR imaging sections [top row], CT [middle row], and fused image with several expert (green) and the automatic (purple) segmentations [bottom row]. The sagittal MR image shows a tumor just anterior and superior, nearly abutting the brainstem.

blood vessels, and muscle. These situations pose different challenges for both human and computer algorithms as borders are not always defined at an area of contrast either *in vivo* or via imaging. The brain is a good area to begin such a study as ours because generally good imaging is available and the anatomy of interest is often explicitly defined. That is not to suggest the these structures, such as the chiasm, are always easily identified. In fact, there are areas in the brain of which borders are not well-defined explicitly, imaging contrast is limited, or both. With the following we will discuss the anatomy pertinent to this work as well as our solutions to their segmentation.

I.4.1 Atlas-based segmentation of the brainstem and eyes

Brainstem

Deep within the brain surrounded by the cerebral hemispheres and the cerebellum lie the structures that comprise the brainstem. It is composed of intermixing gray matter areas

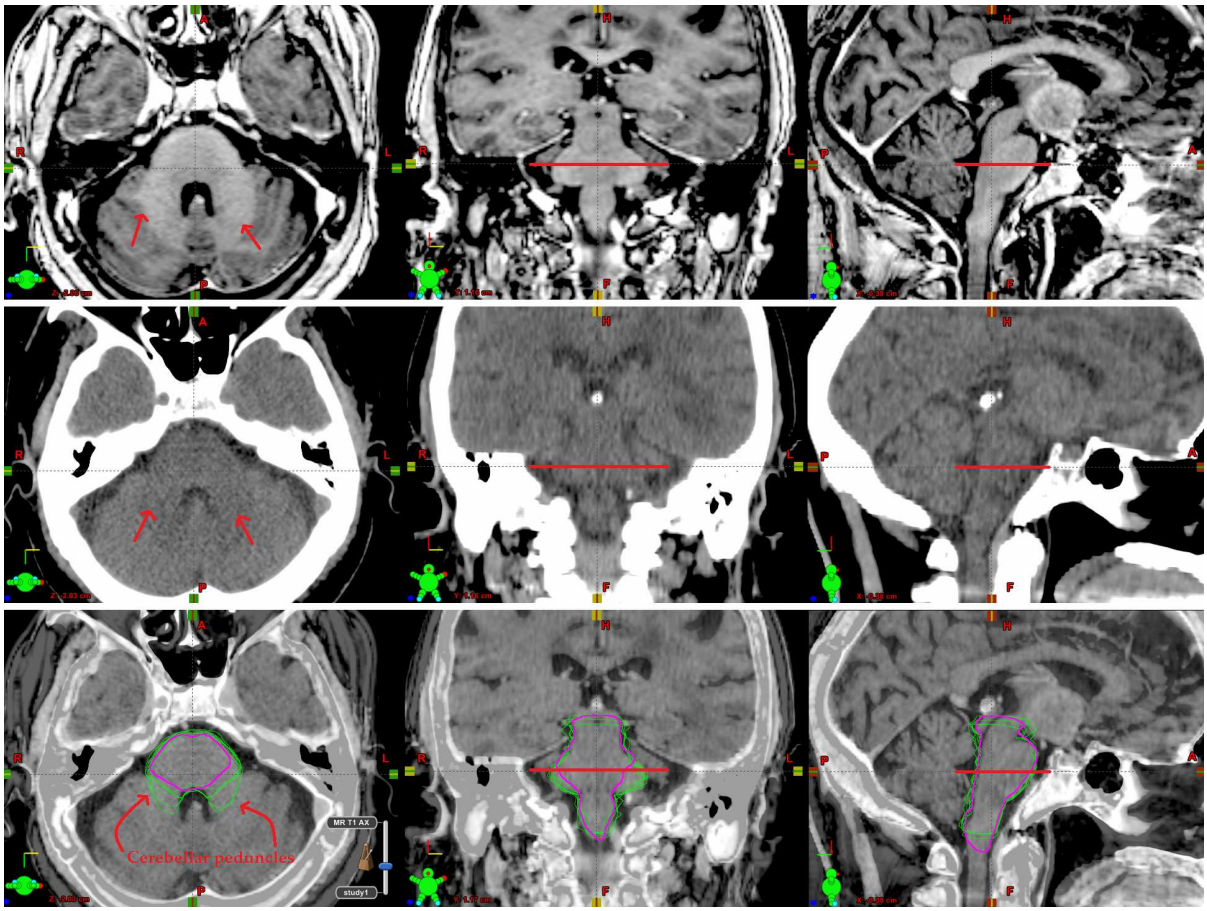

Figure I.3: Brainstem: pons. MR [top row], CT [middle row] and fusion [bottom row] images showing the pons. The area of the cerebellar peduncles where experts tend to exhibit variability, is show in the axial images [left column].

(nuclei) and white matter tracts which serve to connect the motor and sensory controls of the brain to the rest of the body. Beginning inferiorly around the level of the foramen magnum, the spinal cord transitions gradually into the medulla oblongata (figure I.2), which expands and extends superiorly until reaching an area of transverse fibers known as the pons (figure I.3). The pons connects the the cerebral hemispheres to their contralateral cerebellar hemispheres. Centrally, it is separated from the cerebellum by the fourth ventricle. Above the pons lies the midbrain (figure I.4), which is sloped such that the dorsal surface is longer than the ventral surface. On the one hand, when viewed as a series of transverse slices from the cord moving superior to the medulla and the pons, the brainstem begins as a well-contrasted organ on T1 MR and less so in CT. On the other hand, it is a complex structure whose axis changes orientation and is not contiguous in that one area does not flow directly into the next.

The T1 MR images in our work provide for universally better identification of brainstem boundaries than CT. However, there are several areas that present a challenge for seg-

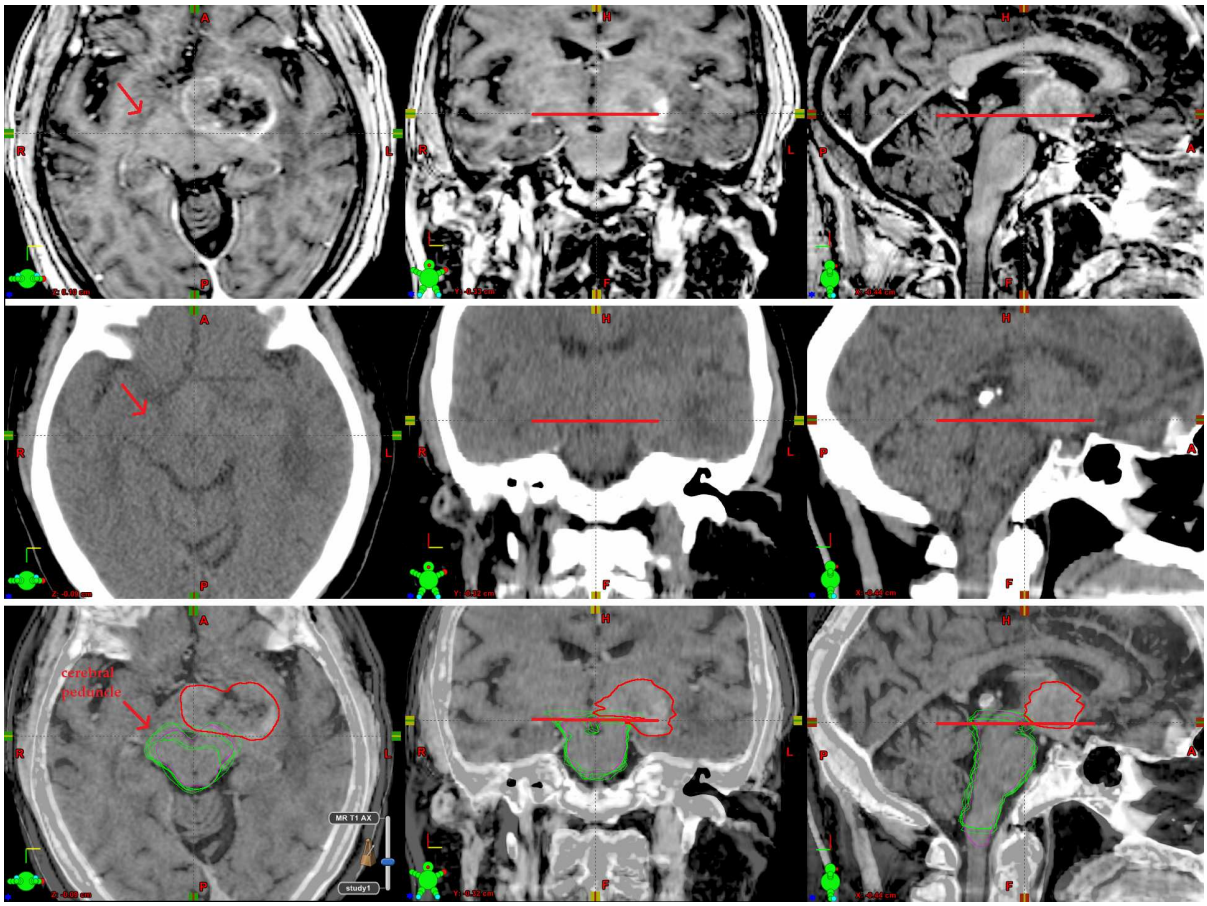

Figure I.4: Brainstem: midbrain. MR [top row], CT [middle row] and fusion [bottom row] images showing the midbrain and cerebral peduncles. The area of the cerebral peduncle is is shown anterior and right; the left peduncle has been invaded by a tumor [left column]. This area presents a challenge for manual segmentation.

mentation. There are bilateral regions of the posterior pons where the brainstem feeds into the cerebellum at the cerebellar peduncles (figure I.3). The same transition occurs between the cerebrum to the brainstem in the midbrain at the cerebral peduncles (figure I.4). These fiber bundles enter the brainstem, becoming part of it before moving on to the cerebellum. Here there is a lack of both imaging contrast and a well-defined explicit boundary. An analogy is that of a tributary joining a main stem river. Its name changes as it joins the larger flow. Where does the name change? Along a line perpendicular to the tributary where it meets the main stem, or the line parallel to the main stem at their confluence? This may seem far removed from brain anatomy, but the questions are sometimes not so different when a binary decision is required. An even more obvious example is at the border of the brainstem and spinal cord. Convention dictates the border occurs at the level of the foramen magnum, but there is no real anatomical difference slightly superior and inferior of this landmark. Thus, human raters must recall this implicit knowledge to mark the boundary at the landmark.

Our group at Vanderbilt developed an atlas-based registration driven approach to segment the brainstem and eyes. It is discussed in more detail in section II.2.2. We begin with a carefully defined consensus atlas from our group of raters. In short, we use a series of affine transformations first globally, then locally, followed by local non-rigid registration using the adaptive bases algorithm. We apply the combined transformations to the atlas delineations to produce a segmentation of the structure of interest.

Eyes

The eye is a deceptively complex organ. It includes the lacrimal gland, cornea, iris, conjuctiva, lens, blood supply, sclera, choroid and retina, to name a few of the major parts. These structures reside in the cranial cavity of the orbit, which is articulated by a number of bones. In radiation therapy we are most usually most concerned with damage to the retina and the lens of the eye. The globe, or eye ball, consists of three layers from outer to inner: the sclera, choroid, and retina. The sclera is composed of connective tissue and is continuous anteriorly with the transparent cornea. The cornea has greater curvature than the sclera and thus protrudes such that the globe is largest along the anteroposterior axis. The inner chamber is filled with vitreous body, a gelatinous transparent mass which contrasts well in CT against the outer layers. In standard radiation therapy rather than identify the retina, the entire globe is usually delineated with the lens as a separate overlaying structure. Much of the posterior hemisphere of the globe is easily identified by CT (I.5). Human experts, however, vary in their delineations of the external surface of the globe: whether at the inner surface of the retina, on the external surface of the sclera, or approximately between the two (choroid). These layers are typically indistinguishable. The posterior hemisphere is surrounded by a large amount of

orbital fat of lower attenuation (Weber and Sabates, 1996), separating it from muscle and bone. Moving anterior, contrast diminishes as the superior, inferior, lateral, and medial rectus muscles (of a more similar attenuation to the globe than the orbital fat) approach the sclero-corneal interface. In our studies we also utilize thin section T1-weighted MR images, which can be well suited to the same posterior aspects as seen in CT, but are especially challenged at the anterior aspects of the globe as a result of both motion artifact and low signal intensity of the lens and cornea.

I.4.2 Model-based segmentation of the optic chiasm and optic nerves

The optic nerve is a tract of brain connecting the eye (retina) to the visual cortex. The retina exits the eye at the optic nerve head and takes a slightly sinuous path to exit the orbit through the optic canal (Hollinshead, 1974). The left and right optic nerve meet and join to form the optic chiasm, an X-shaped structure exterior to the pituitary stalk, and then continue in the optic tracks to the mid brain. Unlike the other cranial nerves, the optic nerves are encased in all three layers of meninges comprising the optic nerve sheath. The nerve itself is bathed in a thin layer of CSF. The blood vessels of the retina and optic pathway are contained within the nerve anteriorly and pierce the sheath posteriorly in the orbital space (Harnsberger et al., 2006). For the purpose of radiation therapy the entire sheath is considered as the operative structure.

A number of challenges exist in segmenting the nerves and chiasm. First, in practice radiation oncologists typically segment the complex as three distinct structures: the right and left optic nerves and the chiasm. Typically the optic nerve is operationally defined as the portion from the nerve head, including the intraorbital segment and ending somewhere past the bony canal distal to the chiasm. The chiasm is defined operationally as two segments from the proximal end of each optic nerve, the intersection of these, and some short 5-10 mm length of optic tracts. This partitioning of a contiguous tubal structure, irrespective of imaging limitations, may contribute to inter-rater variance, especially if raters segment the structures individually without visualizing the other segments simultaneously. In other words, this may lead to overlapping segments and group ambiguity at boundaries. Additionally, most treatment planning systems until just recently, including the one utilized in our study, are not well suited to segmentation of these tubular structures. These systems often offer the user access to orthogonal planes of the 3D volume, and may even permit arbitrary planes, but generally require contouring in the axial plane of the primary CT image. As the visual pathway is somewhat sinuous in multiple planes it is typically contained in parts of several slices. This may lead to slice discontinuities. There are also imaging challenges. First the pathway is thin

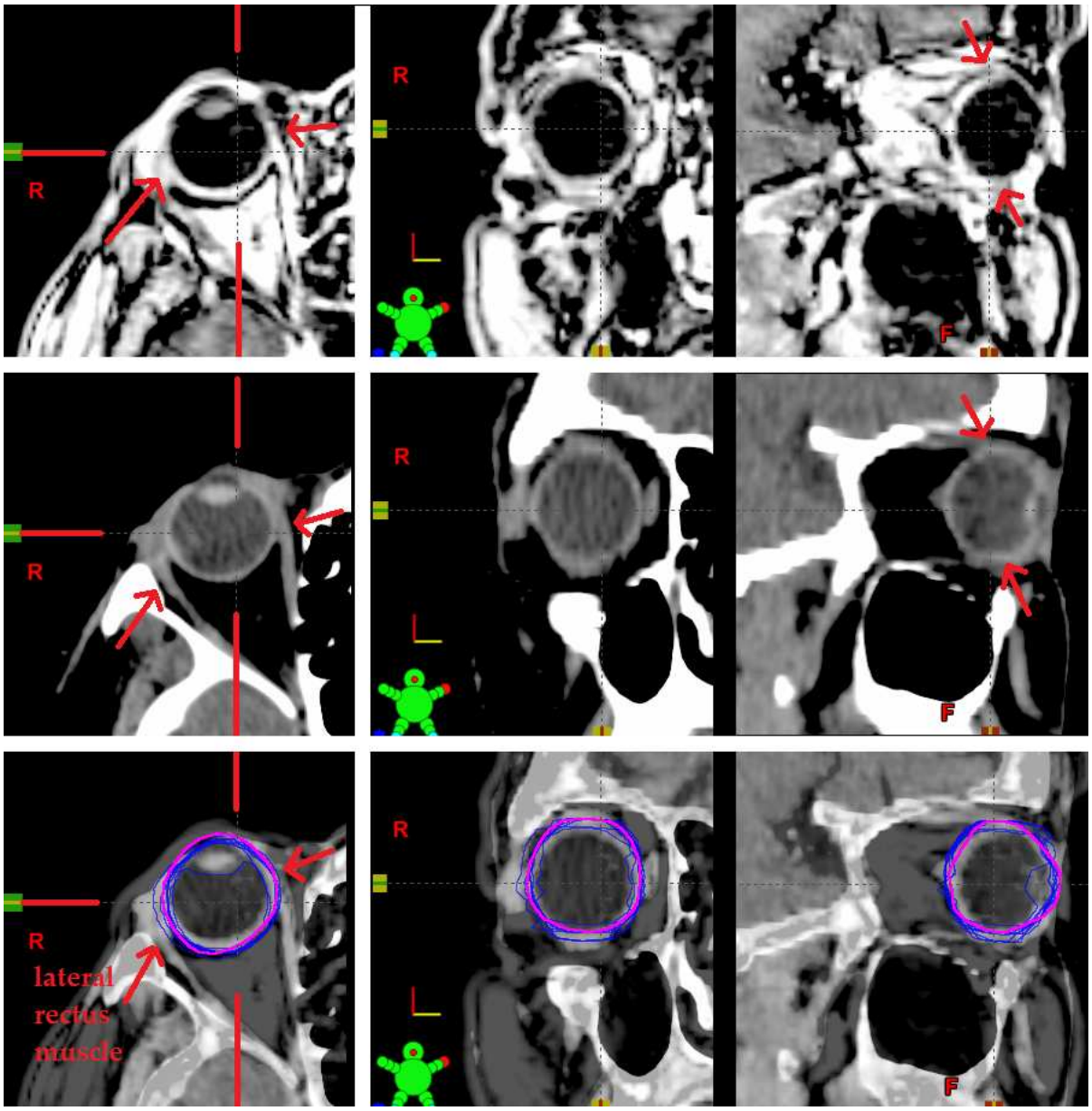

Figure I.5: Eye and recti muscles. MR [top row], CT [middle row] and fusion [bottom row] images showing the globle with views through the recti muscles. The area where the lateral rectus, the lacrimal gland, and the globe meet presents a challenge for both manual and automatic segmentation.

and tubular (all edges are curved surfaces) and therefore susceptible to partial volume effects in both MR and CT imaging. Second MR acquisition times result in motion artifact of the intraorbital nerves as a result of unavoidable eye movement. Third, signal intensity of the portions in or near the optic canal diminishes substantially in MR. CT resolves the intraorbital segments well but contrast is lost moving posteriorly through the canal and not substantially regained (I.6).

The CT and MR images complement one another in resolving the optic pathway, though both remain challenged in the areas surrounded by bone. The atlas-based methods we have used on the eyes and brainstem have proved ineffective for the visual pathway (D'Haese et al., 2003; Isambert et al., 2008).

To this end, others in our group at Vanderbilt (Noble and Dawant, 2011) have developed a model-based method that incorporates both CT and MR and localized the left and right visual pathways as contiguous tubular structures and computes the chiasm as their intersection. This has the advantage of producing non-overlapping structures, and it avoids the most challenging task of explicitly finding the optic chiasm. This approach combines the techniques of optical path finding commonly used in image-guided surgical intervention with model-based methods that incorporate *a priori* information about the area of interest.

Figure I.8 presents 3D renderings of the eye, optic nerve, chiasm and brainstem for a patient from our *de novo* study presented in I. The automatic segmentations are presented in the upper left; the rendering to the right is from a high performing expert on the patient in question. The bottom rendering is from a low performing expert. Note the spurious segment of optic nerve and the gap between the nerve end and chiasm. Also note that this rater appears to have segmented the optic nerve at the level of the pituitary (we leave the other expert chiasm in place as a reference).

I.5 Evaluative framework

A primary goal of this work is to determine the clinical acceptability of the automatic segmentation methods we have developed. As noted in the opening pages of this dissertation, we approach this from a behavioral perspective by measuring the output from a group of experts in clinically realistic situations. One distinction between our work and the relatively few other works (Chao et al., 2007; Stapleford et al., 2010) that have employed multiple raters in a radiotherapy setting is a focus on the individual as well as the group. We consider the automatic system as a potential surrogate to the physicians, and to do that well we need to understand the performance of individuals as well as the whole group. Stylistically the framework is of an inter-rater reliability, or more precisely a method-comparison study (Ludbrook, 2002; Bland

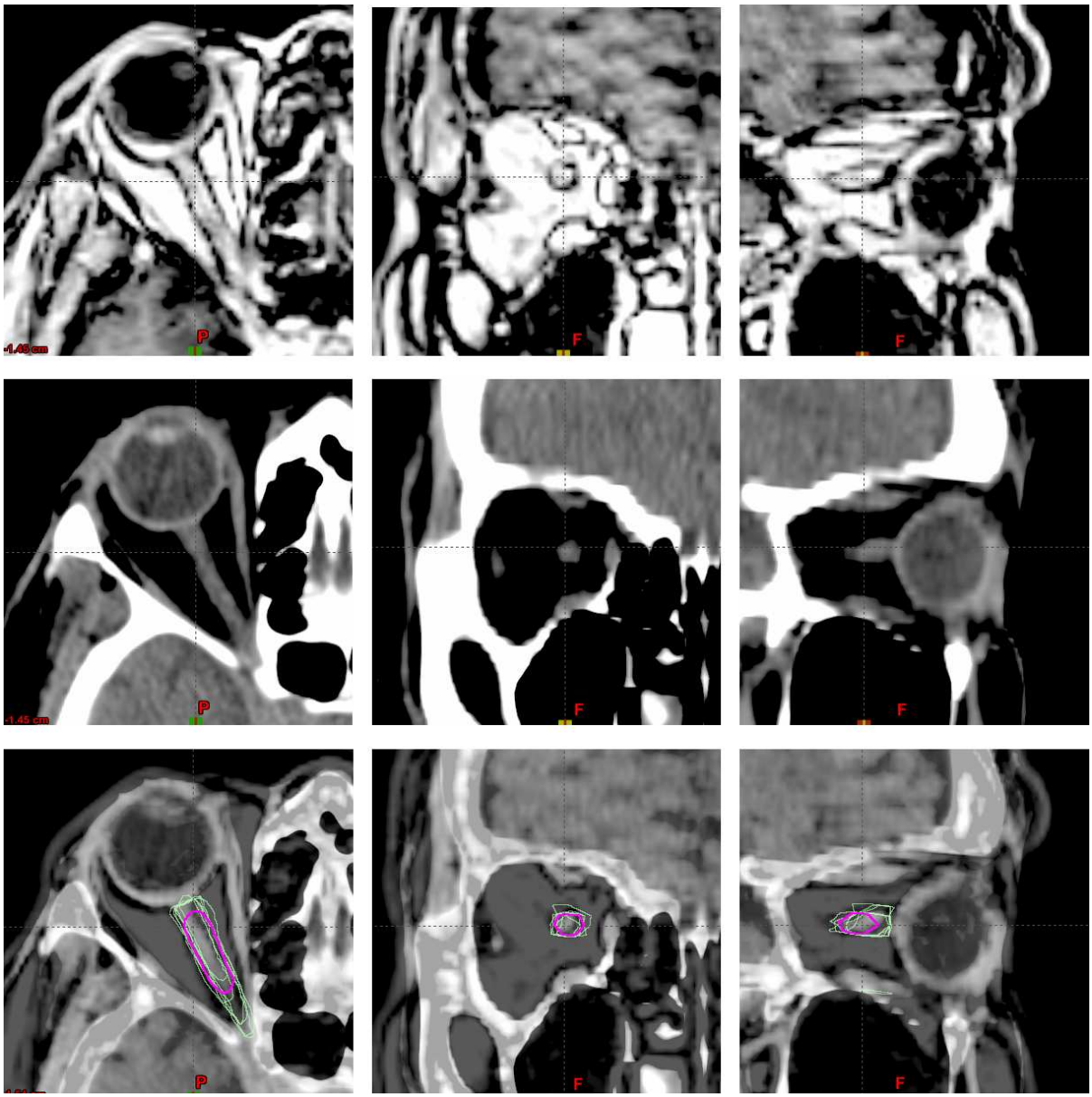

Figure I.6: Optic nerve. The right intraorbital optic nerve is shown in MR [top row], CT [middle row], and fused [bottom row] images. Expert segmentations are presented in green and the automatic in purple.

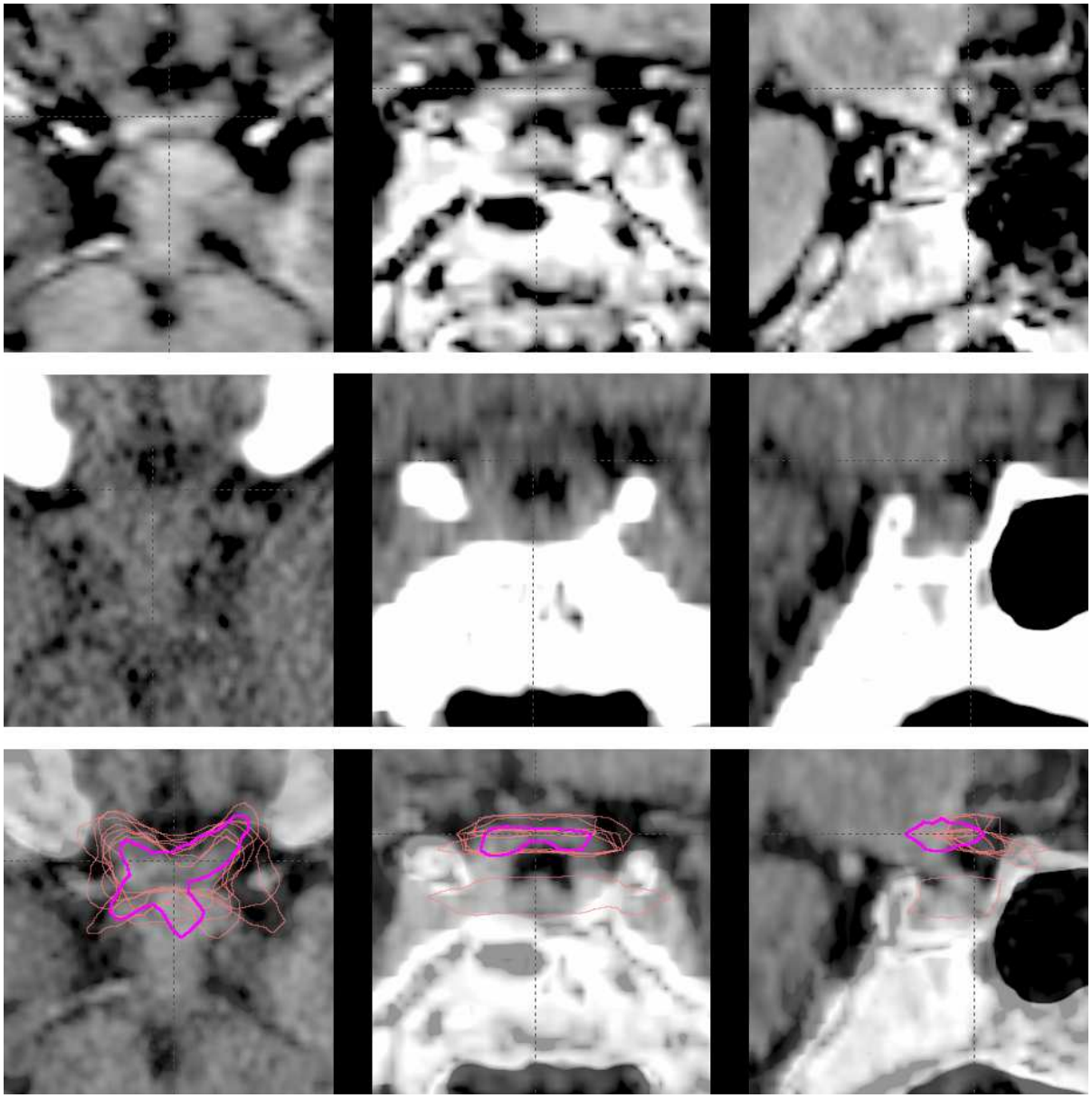

Figure I.7: Optic chiasm. The optic chiasm is shown in MR [top row], CT [middle row] and fused [bottom row] images. The expert contours are shown in pink while the automatic is purple.

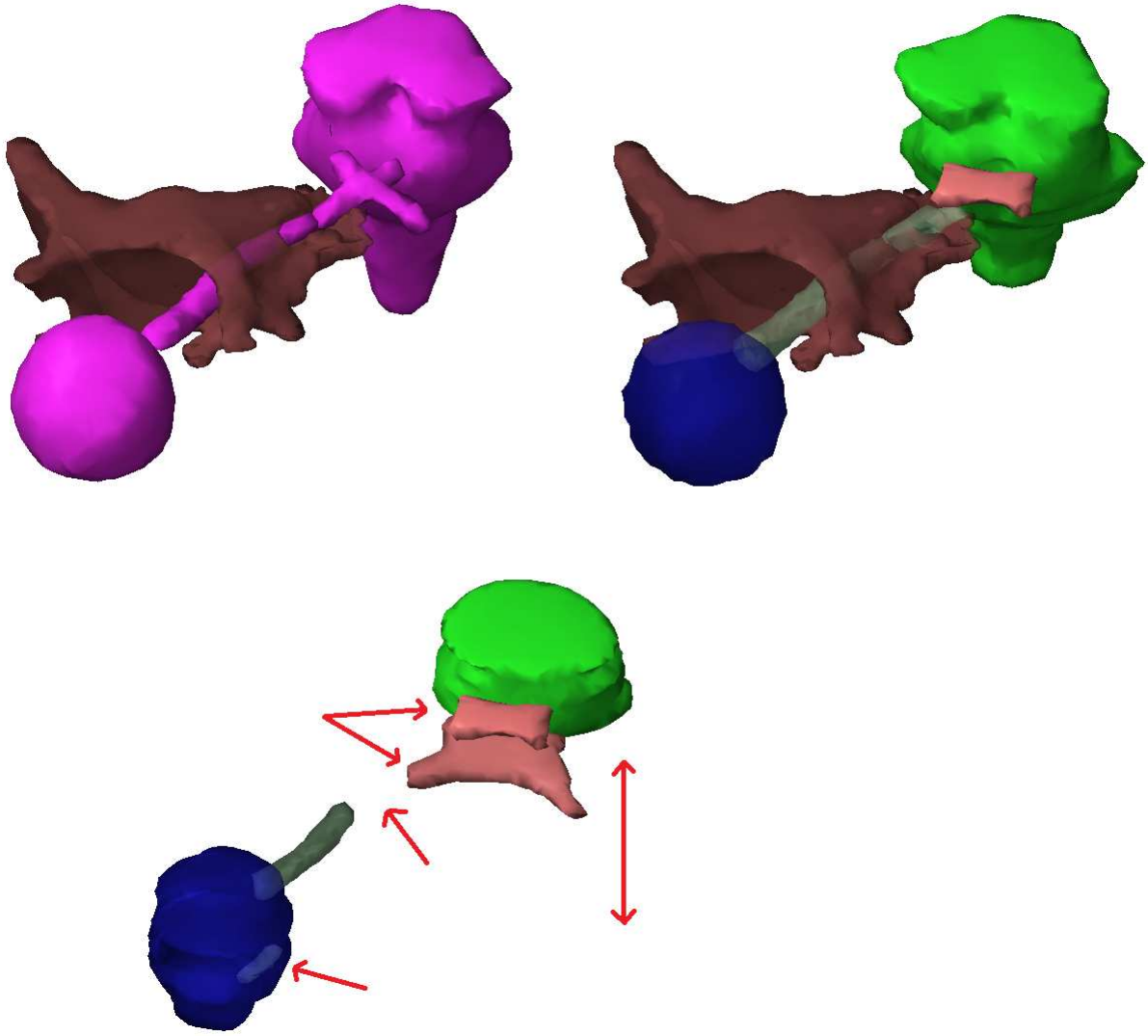

Figure I.8: Three-dimensional rendering of segmentations. The automatic segmentations are shown in the upper left. The orbital bony anatomy has been included to illustrate the path of the nerve. The upper right presents the segmentations from one of the higher performing experts for this patient. The lower rendering is from a lower performing expert for this patient. Note the extra segment of optic nerve behind the transparent inferior aspect of the globe. The chiasm from the other expert has been kept in this panel to illustrate the superior/inferior disagreement. Note also the brainstem is truncated compared to the others.

and Altman, 1999). No assumptions are made other than by virtue of their expert status, the physicians' segmentations are representations of that which would be used clinically to develop treatment plans. We have made efforts to ensure a clinically realistic environment for data gathering, as further described in chapter I. However, a potential pitfall that tempers our inferences slightly is that the experts may not produce segmentations representative of the whole population of professionals who undertake such tasks. Our experts are comprised of three senior attending oncologists and an attending radiologist (P_1-P_4), and four senior radiation oncology residents (J_1-J_4) from a single institution. This work is motivated by the following observations.

First, evaluative studies should reflect both the native form of the segmentations and their interaction along the path to and final impact on the end-use.

Second, medical image segmentation is a problem plagued by lack of a well-defined ground truth. The ground truth is the truest representation of the object of interest possible and may also be known as the gold or reference standard. We were aware of this problem in early design, and in fact, the use of multiple raters arose from this concern. However, in recent years methods for ground truth estimation from a cohort of experts have been developed. There is merit in pursuing both lines of investigation: comparisons between individuals and comparisons with ground truth estimates. The former is used primarily to gauge variability and the latter accuracy. We calculated ground truths via two methods, using the simulateneous truth and performance level estimation (STAPLE) algorithm (Warfield et al., 2004) and our own novel implementation of the concept of probability maps. The STAPLE algorithm uses expectation-maximization to provide a probabilistic estimate of the underlying ground truth and is designed to be robust to outliers within the input segmentation group. We as well as others (Biancardi et al., 2010) have noted, however, that STAPLE may by overly influenced by volumetrically larger segmentations within the input cohort. To combat this we developed an additional method. This was spurred by the work of Meyer and colleagues (Meyer et al., 2006) in an evaluation of lung nodule annotation by radiologists. They summed radiologist segmentations and normalized to the number of raters to produce what they termed probability maps (p-maps), noting that the median of the p-maps appeared to be a good segmentation of the lesions. We calculated a ground truth from the p-maps using the mean after Gaussian smoothing. This is similar to voting rule, but rather than threshold the p-maps at a predetermined level, we use the mean. By doing so, the threshold changes as the rater-decision making model changes; in other words, raters rate differently in different situations such as the brainstem versus the chiasm. Using a statistic such as the mean adjusts for the change. The methods and rationale of ground truth estimation are further discussed in section II.2.3.

Third, assessments should be multidimensional. Segmentations are not easily quantifi-

able via a single summary measure such as, for example, serum levels of a drug might be. They are a much higher level data structure, more similar to measuring the molecular distribution of a drug over the entire body. In geometric comparisons we use several complementary metrics which are cross-study compatible. We calculate volume (we refer to this as *nominal* volume to disambiguate from the other uses of the term) as a stand-alone summary measure. The Dice similarity coefficient (DSC) (Dice, 1945) measures spatial overlap between two segmentations normalized to their mean volume. It is derived as a special case of the kappa statistic (Cohen, 1968), a statistical measure of inter-rater agreement, as worked out by Zijdenbos and colleagues (Zijdenbos et al., 1994). DSC offers the advantages of a simple means of pairwise comparison, size and location sensitivity, and a finite range [0,1]. What level of DSC constitutes satisfactory agreement is unclear, both statistically and from a segmentation standpoint. Our work is generally invariant to this as it is clear that higher DSC represents better agreement, and we are more interested in comparisons of distributions than absolute agreement. The quality of DSC as a measure of similarity, however, is likely not universal over different types of structures. It is less sensitive to differences for structures such as the brainstem where there is a relatively large volume of agreement compared to small though potentially important regions of disagreement. This underscores the need for more than a single metric. Lastly, we use distance-based metrics to gauge differences between edges. Distance-based metrics are generally directional, that is, the distance A $\rightarrow$ B does not equal A $\leftarrow$ B. Often the bidirectional mean is used. We calculate in only one direction, however, as we are most concerned specifically with edge difference in this direction (from ground truth to test segmentation). A drawback of this is that a segmentation may have several slices entirely missing and yet return very small distance errors. We overcome this with what information is provided by nominal volume and DSC. This yields important information about the quality at edges where raters in fact decided to delineate as opposed to where they decided not to delineate. From this information we can calculate what we term the true positive rate at a specified distance. The true positive rate is simply the proportion of contour points falling within a shell of specified thickness about the ground truth estimate. For example, a rater with a high true positive rate but low DSC and small volume may focus on specific areas of a structure with high accuracy while completely omitting another area of the structure.

I.6 Goals and contributions of the work

For automated segmentation methods to be clinically useful, they need to improve efficiency and at minimum maintain variability and accuracy compared to clinicians. In other words, the system must serve as a robust surrogate to the human actors. The primary goal of

this work is to determine whether our automated system for segmentation of intracranial organs at risk satisfies these requirements. We do so via a multi-rater behavioral study that seeks to 1) assess geometrically the automated segmentations in the context of inter-expert variability and accuracy *de novo*, 2) gauge the impact of segmentation editing, and 3) measure the sensitivity of the end-use, that is, radiation dosimetry, to segmentation differences. In doing so we gain insight not only into the quality and utility of the automated methods but also new information regarding the accuracy and variability of experts and impacts thereof on dosimetric outcome.

A secondary goal of this work is to develop a tool for future investigation. We aim to develop a framework that can be applied to other anatomical sites, specifically within radiation therapy. Our framework utilizes a combination of multiple complementary metrics both on the native segmentations and at their end-use, a behavioral approach with multiple expert raters, and a novel method of ground truth estimation from the cohort of experts.

CHAPTER II

CHARACTERIZATION OF SEGMENTATION VARIANCE

Phys. Med. Biol. 56 (2011) 4557-4577

COMPARISON OF MANUAL AND AUTOMATIC SEGMENTATION METHODS FOR BRAIN STRUCTURES IN THE PRESENCE OF SPACE-OCCUPYING LESIONS: A MULTI-EXPERT STUDY

M A Deeley[1], A Chen[2], R Datteri[2], J Noble[2], A Cmelak[1], EF Donnelly[3], A Malcolm[1], L Moretti[1,5], J Jaboin[1], K Niermann[1], Eddy S Yang[1], David S Yu[1], F Ye[4], T Koyama[4], G X Ding[1] and B M Dawant[2]

[1] Department of Radiation Oncology, Vanderbilt University, Nashville, TN
[2] Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN
[3] Department of Radiology and Radiological Sciences, Vanderbilt University, Nashville, TN
[4] Department of Biostatistics, Vanderbilt University, Nashville, TN
[5] Department of Radiation Oncology, Institut Jules Bordet, Université Libre de Bruxelles, Brussels, Belgium

Abstract

The purpose of this work was to characterize expert variation in segmentation of intracranial structures pertinent to radiation therapy, and to assess a registration-driven atlas-based segmentation algorithm in that context. Eight experts were recruited to segment the brainstem, optic chiasm, optic nerves, and eyes, of 20 patients who underwent therapy for large space-occupying tumors. Performance variability was assessed through three geometric measures: volume, Dice similarity coefficient, and Euclidean distance. In addition, two simulated ground truth segmentations were calculated via the simultaneous truth and performance level estimation (STAPLE) algorithm and a novel application of probability maps. The experts and automatic system were found to generate structures of similar volume, though the experts exhibited higher variation with respect to tubular structures. No difference was found between the mean Dice coefficient (DSC) of the automatic and expert delineations as a group at a 5% significance level over all cases and organs. The larger structures of the brainstem and eyes exhibited mean DSC of approximately 0.8-0.9, whereas the tubular chiasm and nerves were lower, approximately 0.4-0.5. Similarly low DSC have been reported previously without the context of several experts and patient volumes. This study, however, provides evidence that experts are similarly challenged. The average maximum distances (maximum inside, maximum outside) from a simulated ground truth ranged from (-4.3, +5.4) mm for the automatic system to (-3.9, +7.5) mm for the experts considered as a group. Over all the structures in a rank of true positive rates at a 2 mm threshold from the simulated ground truth, the automatic system ranked second of the nine raters. This work underscores the need for large scale studies utilizing statistically robust numbers of patients and experts in evaluating quality of automatic algorithms.

II.1 Introduction

Three-dimensional imaging advances have revolutionized the treatment planning process in external beam radiation therapy. They provide physical information by which to calculate dose and specify external geometry, and as highly conformal treatments have become prevalent, they provide increasingly important information regarding patient anatomy both diseased and at risk. As a result image segmentation has become a central part and often rate limiting step in the planning process. Radiation oncologists must make judgments incorporating implicit and explicit anatomic, histologic and physiologic information in the presence of varying image quality to partition an image volume into normal and diseased tissue. This is a time consuming process that must occur before designing fields or calculating dose and thus can be a significant contributor to the overall efficiency of the process. The need for segmentation is only expected to increase in the future as additional conformal and adaptive techniques are implemented (Mell et al., 2003, 2005).

Until recently segmentation of all but the simplest structures was accomplished manually. Of late, however, a number of semi- and fully-automated methods have been developed to segment normal tissues in a radiotherapy clinical context (Gorthi et al., 2009; Malsch et al., 2006; Lu et al., 2004, 2006; Xie et al., 2008; Reed et al., 2008; Zhang et al., 2007; Pasquier et al., 2007; Isambert et al., 2008). Evaluation of these methods has been a persistent challenge as medical image segmentation unfortunately lacks a known ground truth, or gold standard, in its real world application. Phantoms provide an easily identifiable ground truth but are an unrealistic surrogate for patient imaging. The same can be said for synthetic images and cadaver sections. As noted by Warfield et al., the accuracy of a reference standard and the degree to which it reflects the clinical concerns are often inversely related. Accordingly, a single manual rater provides realistic data but can suffer from intra- and inter-rater variance. Recognizing the need for a useful reference standard, Warfield and colleagues introduced a method known as the simultaneous truth and performance level estimation (STAPLE) algorithm (Warfield et al., 2004) to simulate a ground truth from a cohort of manual segmentations.

In addition to the absence of a known ground truth, evaluation methods have also lacked consensus as to comparison metrics. The choice of comparison metrics is quite important, as each yields different information and must be considered in the appropriate context. Generally, these measures fall into one of two categories: volume-based or distance-based. Measurement of nominal segmentation volume is a simple measure that does not require a reference standard for calculation, which makes it computationally inexpensive and allows for easy cross-study comparison with minimal background information. Spatial overlap measures such as the Dice similarity coefficient (DSC) (Dice, 1945) and related Jaccard coefficient (Jaccard, 1908)

have been most broadly adopted in the literature in recent years. While these yield a good sense of volume overlap of two segmentations, they provide little in terms of the scale of mismatch (Crum et al., 2006). Specificity and sensitivity are also commonly applied. Specificity, however, is plagued by its dependence on the number of true negatives; that is, the number of voxels in the image space not contained within the segmentation. This value may change quite considerably between studies simply as a function of image or region of interest size. A weakness of volume, DSC, and specificity and sensitivity, is that they are fairly insensitive to edge differences when those differences have a small impact on overall volume. For example, two segmentations with large total volume may show a high degree of spatial overlap while exhibiting clinically relevant differences at their edges. Distance measures, however, such as the Hausdorf and Euclidean, or surface normal, distances offer yet another means of comparison by providing information regarding the differences in edges of two segmentations. The distance calculations generally result in a vector of distances that may be summarized as mean or median, or may be used in further statistical analyses. Thus, our experience has been that a combination of several volume and distance measures is required to gain a deep perspective of the dataset.

Our work is motivated by the observation that medical image segmentation is inherently a problem lacking a known ground truth. Accordingly, clinical evaluation studies should be behavioural in nature, employing a number of raters and patient volumes such as to provide good statistical power in the targeted clinical context. We designed a study to quantify variation amongst physicians in segmenting organs at risk in the brain and to assess our automated system in this context. Several other multiple observer studies have focused on evaluating automatic or semi-automatic systems within the brain (Bondiau et al., 2005; Isambert et al., 2008; Babalola et al., 2009) and head and neck (Chao et al., 2008; Stapleford et al., 2010), but we know of no other study as comprehensive in terms of patient numbers, expert raters, and organs segmented. In addition, to be as clinically relevant as possible, we chose to conduct the study on volumes with large space-occupying lesions. We chose this anatomical site for the wealth of matched computed tomography (CT) and magnetic resonance (MR) imaging available, the clinical relevance to intensity-modulated radiation therapy (IMRT), as well the ubiquity in physician training in intracranial anatomy. We tested the hypothesis that the automatic system would produce segmentations that could serve as surrogates to the manual physician segmentations. An ancillary goal of this work was to collect a large and statistically robust dataset, which is useful for evaluating not only our algorithms but also those being developed by other groups. The recent release of several commercial radiotherapy segmentation systems underscores the need for a strong multi-rater data set for evaluation.

II.2 Methods

II.2.1 Study design

We selected 20 patients that had been previously treated in our department with IMRT for high grade gliomas. We chose difficult cases with large space-occupying tumors, often close to the critical structures, which would present a challenge for the non-rigid registration-based segmentation algorithm we use as well as yield pertinent dosimetry for the next phase of analysis. The mean gross tumor volume (GTV) and clinical tumor volumes (CTV) were 49 and 199 cm^3, respectively. As a point of reference, these volumes roughly translate into a mean spherical equivalent of 4 and 7 cm in diameter. Each patient underwent stereotactic biopsy for which high resolution T1 MRs were acquired under 1.5T (N=10) or 3T (N=10) magnetic fields and reconstructed into image volumes of voxel size approximately 1x1x1.2 mm^3. A helical CT of dimensions approximately 0.6-0.7 mm in the axial plane and either 2 mm (N=14) or 3 mm (N=6) in slice thickness was acquired for treatment planning.

Eight physicians were enlisted in this study as expert raters: 4 junior physicians (J_1-J_4) and 4 senior physicians (P_1-P_4). The senior physicians were comprised of 3 radiation oncologists and a diagnostic radiologist, while the junior physicians were PGY5 radiation oncology residents. Before initiating the study, we reviewed images and our atlas delineations with them as a group to set general anatomical guidelines. One important guideline reiterated throughout the process was to set the inferior border of the brainstem at the foramen magnum, as the brainstem lacks a physical boundary with the the spinal cord. Another concern was where the brainstem meets the cerebellum in the lower pons. Here there is no significant contrasted boundary, so we developed an implicit rule whereby the experts should begin the contour anteriorly at the basilar sulcus of the pons, extend laterally to include the middle cerebellar peduncles, and continue posteriorly and medially toward the median sulcus of the fourth ventricle making an angle of approximately 45 degrees to the anterior-posterior axis.

The patient volumes were anonymized and loaded into a commercial treatment planning system (Eclipse version 8.5, Varian Medical Systems, Palo Alto, CA). This workstation was identical to the clinical systems in our department while reserved for research only. Computed tomography and MR images were registered within the planning system and fused. Each physician was given the opportunity to change window and level settings to his liking and received instructions to use all imaging information available to them to the point at which each felt confident delineating a critical structure. They were asked to delineate the brainstem, optic chiasm, optic nerves, and eyes. An in-house graphical user interface was constructed to inform the physicians where they stood in the task queue and to provide a mechanism to record time.

The timing mechanism allowed the rater to pause momentarily or leave the system entirely and return later. Each rater was blinded to the work of the others. The delineations were collected over approximately one year.

Each physician was given all of the tools afforded by the clinical treatment planning system for contouring. A "paintbrush" tool produces an opaque segmentation as the expert traces out the structure. A "pencil" tool is similar without producing the opacity and can be used in a continuous or stepwise mode. There was also an "eraser" tool and the ability to stretch and deform contours after delineating. Three orthogonal views were present on screen at all times, though only axial were available for contouring. This is a limitation of the clinical software. We advised the experts to use the same tools they would use clinically and with which each was comfortable. We also advised them to inspect the final product of their work before completely the task. Lastly, above all we instructed the experts to perform these tasks in the context of real world clinical relevance.

The final result of each contouring session was a set of points in DICOMRT standard format that were stored at sub-voxel resolution.

II.2.2 Automatic segmentation

Two methods were utilized for the automatic segmentations in this study. The first method utilizes atlas-based registration (Crum et al., 2004) to segment the eyes and the brainstem, while the second method utilizes a general technique we have developed for the segmentation of tubular organs, which we call the atlas-navigated optimal medial axis and deformable model algorithm (NOMAD) (Noble et al., 2008; Noble and Dawant, 2009).

We first manually delineated the brainstem and the eyes in an atlas image. Then, a global affine registration was computed and used to register the atlas image (panel II.1a, bottom row) onto the target image (panel II.1a, top row) that we want to segment. A predefined bounding box around each organ is extracted from both the atlas and target image after the global affine registration (panel II.1b). Another affine registration is performed locally between the extracted boxes of the atlas and target images, again resulting in a transformation that is used to project the atlas onto the target image. This second affine registration is performed to limit the registration on a local area within the image. The size of the boxes is determined by the size and shape of the organ of interest within the atlas image, with an arbitrary amount of padding to aid in the local affine registration. This registration utilizes the normalized mutual information (NMI) (Studholme et al., 1999) as the similarity measure. Lastly, local non-rigid registration is then performed between the results of the local affine registration and the atlas image. The manual contours drawn on the atlas are then projected onto the target image

utilizing the deformation fields that were the result of the three registrations (panel II.1c).

The non-rigid registration approach is an algorithm we termed the adaptive bases algorithm (ABA) (Rohde et al., 2003). This algorithm uses normalized mutual information (Studholme et al., 1999) as the similarity measure and models the deformation field that registers the two images as a linear combination of radial basis functions (Wu, 1995) with finite support.

Both the forward and the backward transformations are computed simultaneously, and the transformations are constrained to be inverses of each other using the method proposed by Burr (Burr, 1981). Although this cannot be proven analytically, experience has shown that the inverse consistency error (Christensen and Johnson, 2001) achieved with this approach is below the voxels' dimension. In our experience, enforcing inverse consistency improves the smoothness and regularity of the transformations.

In this work, we segment the optic nerves by applying the NOMAD algorithm. The NOMAD algorithm first computes the medial axis of the structure as the optimal path with respect to a cost function based on image and shape features. The medial axis is then expanded into the full structure using a level-set algorithm. Unlike other methods (Feng et al., 2004; Yim et al., 2001), NOMAD uses a statistical model and image registration to provide the above segmentation framework with a priori, spatially varying intensity and shape information, thus accounting for unique local structure features. The statistical models were trained on volumes not included in this study.

In order to compensate for the lack and changing contrast of the structures, we take advantage of both the CT and MRI to build the models used by the algorithm. To ensure that the intensity information will consist of the best possible contrast, we rely solely on the CT in the region of the optic nerves, and solely on the MR in the region of the optic tracts and chiasm. The model consists of the set of points that compose the center line of the structure and their associated expected values for intensity and shape features extracted from the rigidly aligned MRs and CTs. Once the models are built, new sets of images can be segmented.

II.2.3 Calculation of simulated ground truths

We calculated two simulated ground truths for comparison to individual raters (P_1-J_4) and our automatically generated segmentations, A_1.

First, we used the STAPLE algorithm (Warfield et al., 2004) to calculate a consensus estimate from the physician segmentations. The STAPLE algorithm uses expectation-maximization to provide a probabilistic estimate of the underlying ground truth. It is designed to be robust to outliers within the input segmentation group. A second simulated ground truth

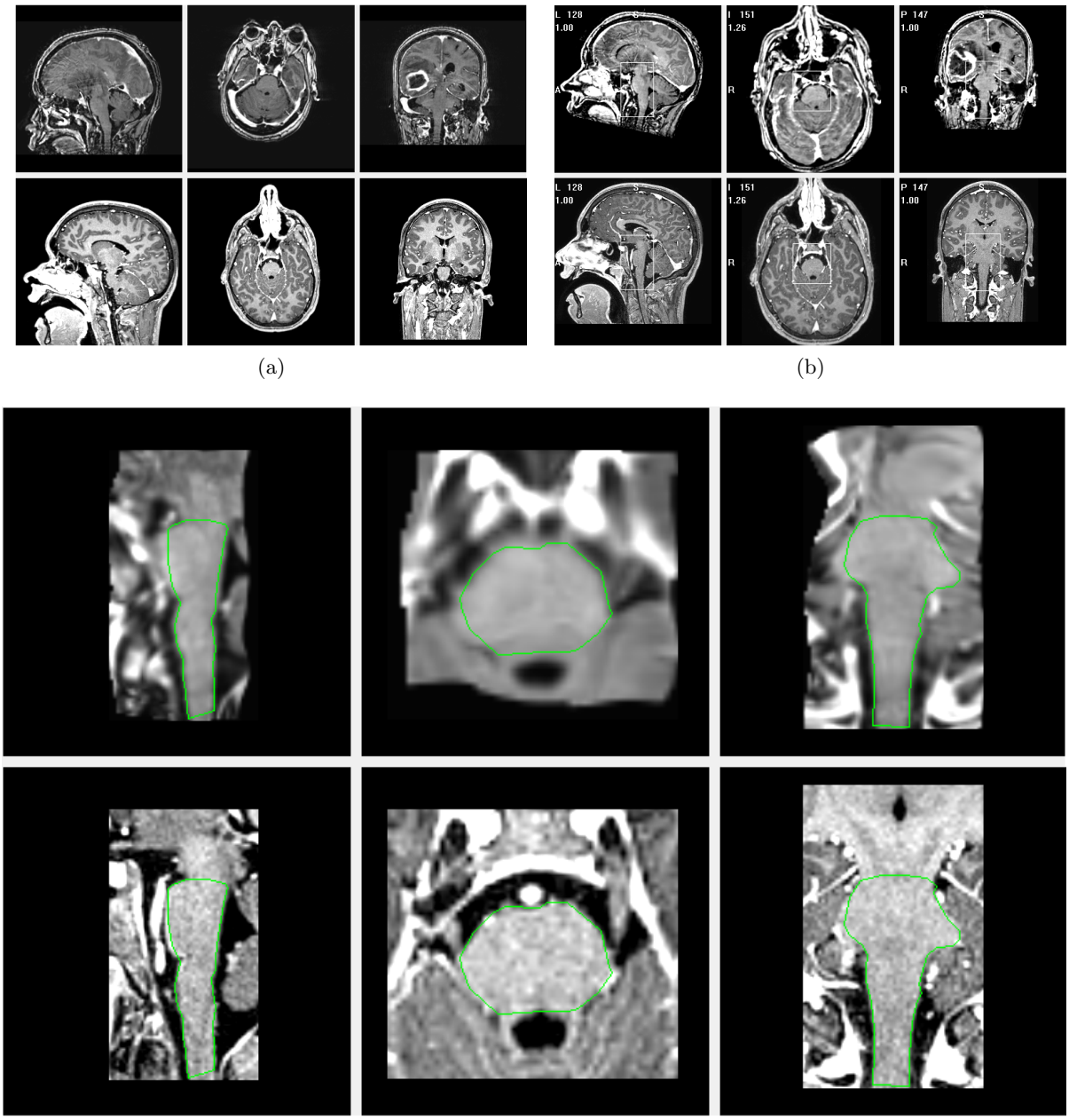

Figure II.1: Atlas-based segmentation process for the brainstem and eyes. Panel (a): Orthogonal slices of a patient (top row) with a large right sided lesion and the atlas (bottom row) before registration. Panel (b): Volumes are then globally, affinely registered, and a bounded atlas region (white box) is projected onto the patient. Panel (c): Local affine and local non-rigid registration are performed on the bounded region where the top row represents the final product of the patient brainstem deformed to the atlas. The green contour drawn on the atlas and fused with the final registration result for the patient demonstrates the correspondence that has been achieved between the two images.

was calculated through the creation of probability maps, termed p-maps (Meyer et al., 2006). A separate probability map was created for each rater across critical organ structures and patients to remove potential bias explicitly. The p-maps were created by summing the binary masks of each rater for a particular organ, omitting the rater for which the p-map will be used in comparison. For example, the p-map for rater P1 would be formed by summing the 7 binary masks of raters P_2-J_4. The 3D array is then normalized to the number of raters included and smoothed using a 3x3 pixel Gaussian kernel applied in-plane with a standard deviation of 0.65 pixel width. The smoothing increases correlation between adjacent voxels, but it also improves the validity of later statistical tests that rely on assumptions of normality. We chose the filter parameters heuristically as a balance between reduction in gross quantization and an increase in spatial correlation between voxels. Additionally, we removed rater P_2 from the p-maps, as an initial statistical analysis showed this rater produced several outliers within the complete dataset. The ground truth estimate was then created by thresholding the p-map at a desired level to form a binary mask. The choice of threshold level presents a challenge in using p-maps for ground truth estimation. A common interpretation is to choose a static, fixed value. For example, 0.5 would represent majority vote in which at least half of the raters agree. We chose to threshold at the mean value of the distributions, thus yielding a threshold specific to each p-map. That is, each voxel with a value greater than or equal to the mean of that p-map was included in the ground truth segmentation. While the mechanics of p-map creation and thresholding are identical for a static level, our method recognizes that rater consensus may vary considerably between structure and even between cases within structure. Another way to think about this is that the level of spatial independence within p-maps, an assumption violated for both STAPLE and p-map methods, varies over structures and cases. Choosing a static level such as simple majority vote, 0.5, assumes that value to be most representative of the group preferences over all structures and cases. However, in calculating a measure of central tendency we treat each scored voxel as a sampling distribution, and we take the mean of these sampling distributions as an appropriate level of consensus, thereby adjusting the level in response to the nature of the data.

We chose the STAPLE method as it is designed to produce a probabilistic ground truth estimate robust to deviations in rater performance. Use of STAPLE has become prevalent in segmentation evaluation work, and thus its inclusion herein should facilitate comparison with current and future work. While STAPLE was easily applicable to our imaging data, we also calculated the p-map-derived ground truth for its computational simplicity and the statistical value of the p-maps in future studies. We will refer to these simulated ground truths as STAPLE and PMAP_{mean}.

II.2.4 Comparison metrics

The data obtained in this study are most basically three-dimensional coordinate sets. To make judgments and draw conclusions about these data, we compare them using several metrics sensitive to different aspects of geometry. For this study we calculated two volumetric measures and one distance measure: volume, Dice similarity coefficient, and Euclidean distance from a simulated ground truth. The volume is calculated quite straightforwardly as the sum of the voxels contained within the binary mask of a segmentation multiplied by the voxel dimensions, which in our case were in CT space. The Dice similarity coefficient (DSC) has been used broadly in the field of segmentation as a measure of spatial overlap (Dice, 1945; Jaccard, 1908; Zijdenbos et al., 1994). The volumetric DSC is defined in equation II.1 as the intersection of two masks normalized to their mean volume, where A and B are the masks and N is an operator yielding the number of voxels.

$$\text{DSC(A,B)} = \frac{N(A \bigcap B)}{\frac{1}{2}(N(A) + N(B))} \tag{II.1}$$

Its range is [0,1] where zero indicates no overlap and 1 indicates exact overlap. Measures such as volume and DSC can be insensitive to differences in edges if these differences lead to an overall small volumetric effect in relation to the total volume. The relative sensitivity of DSC to edge differences is a function of shape, or more explicity the number of edge voxels in comparison to the number of inner voxels. For example, DSC will be more sensitive to edge variation in thin tubular structures such as the optic chaism and nerves than in the brainstem and eyes, where the majority of voxels are not at the edges. Edge variation, however, could be quite important in a radiotherapy inverse planning context.

To gain information about differences at the edges of segmentations, we used the three-dimensional coordinates obtained from individual physician and automatic segmentations. We used these points to sample a distance map. The Euclidean distance map, or transform, is a pregenerated 3D array in which each voxel contains the value of the straight-line, or surface-normal, distance to the nearest non-zero voxel. We used PMAP_{mean} as the source from which to calculate the distance maps. To determine the distribution of distances for an individual segmentation, the appropriate distance map was sampled at the contour points of the segmentation. This method yields a distance from each point drawn by a physician or the automatic system to the simulated ground truth. The distances were signed such that a rater's contour point lying inside the boundary of the ground truth was scored negative and outside scored positive. There are several ways to utilize distances. Often only the absolute distance from the ground truth is considered where direction is unimportant. In the context of radiotherapy we

feel it is important to know whether a rater segments consistently small or consistently large as compared to the ground truth estimate. From this signed distribution of distances one can then do a number of things. We chose to generate boxplots of the distributions to get a sense of overall variability and understanding of whether there were instances of systematically positive or negative distances. We further used the absolute values of these distances to calculate true positive rates as a gauge of overall quality of segmentation.

It is important to recognize that this calculation provides information about where a rater made the decision to segment. It says nothing about where the rater decided to not segment. For example, we can imagine a simulated ground truth that extends for several axial slices of a CT image. A rater in question may draw an exact match to the simulated ground truth but on one slice only. The resulting distance distribution for this rater would be a vector of zeros, indicating that in every place the rater made a decision to draw a line, that decision was correct. The distance distribution says nothing regarding the failure of the rater to segment the other slices.

We chose two volume based measures coupled with the distance measure to provide more complete information about how segmentations differ. Alone each measure has a weakness. Volume and DSC tend to integrate edge differences that are small relative to the overall size of the segmentation. Meanwhile, the distance measure captures information only in the context of edges that were drawn, ignoring areas that a rater opted not to segment.

II.3 Results

Figures II.2 and II.3 present manual and automatic segmentations from a subject chosen randomly from the 20 patients used in this study. The eight physician-segmentations for the brainstem, chiasm, eyes, and optic nerves, can be seen in multiple colours, while the automatically generated segmentations are purple for each structure. The tumor volume is shown in red on the coronal slice. Figure II.3 similarly presents axial contours of the brainstem and eyes, illustrating the variation that can be seen qualitatively between experts. We found this area of the brainstem at the cerebellar peduncles to be a consistent source of variation amongst the experts. The results we present here are an attempt to quantify the variation geometrically such that we can make judgments about the expert and automatic segmentations, as well as the interaction of the two.

We calculated several quantitative measures to make comparisons between segmentations: volume, Dice similarity coefficient (DSC) and Euclidean distance from a simulated ground truth. Figures II.4–II.7 use the boxplot to represent the results of these calculations. The boxplot presents the range of the distribution with a thin vertical line through the box.

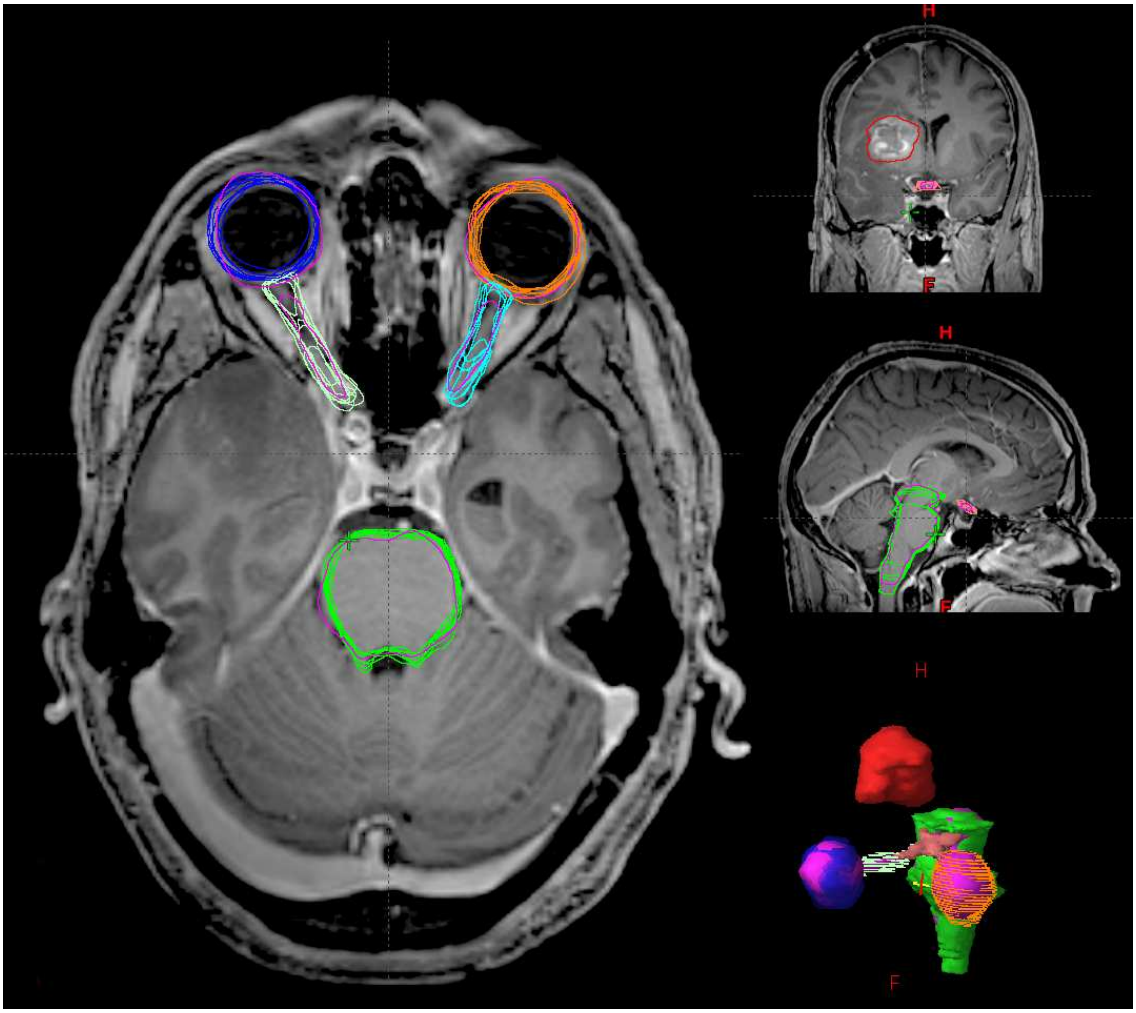

Figure II.2: A randomly chosen patient from the 20 cases used in this study. Eight physician raters segmented the brainstem, optic chiasm, eyes, and optic nerves using a fused CT/MR image set. The automatically generated segmentations are shown in purple. The large red contour in the right parietal is the gross tumor volume.

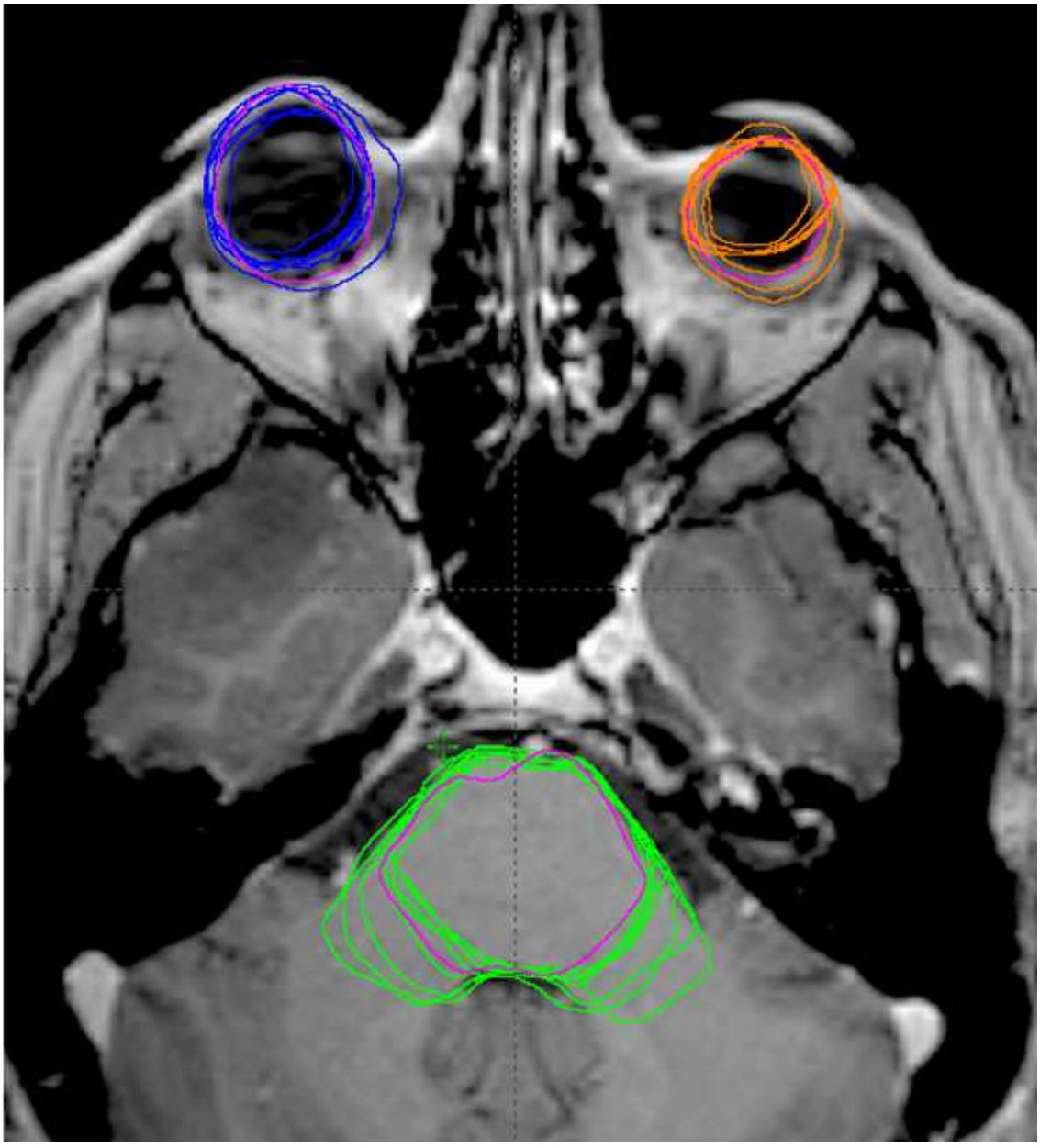

Figure II.3: Axial slice showing an area of high physician variability within the brainstem. In this area of the cerebellar peduncles there is little anatomical contrast, such that the physicians rely primarily on implicit knowledge. The automatic contour is represented in purple.

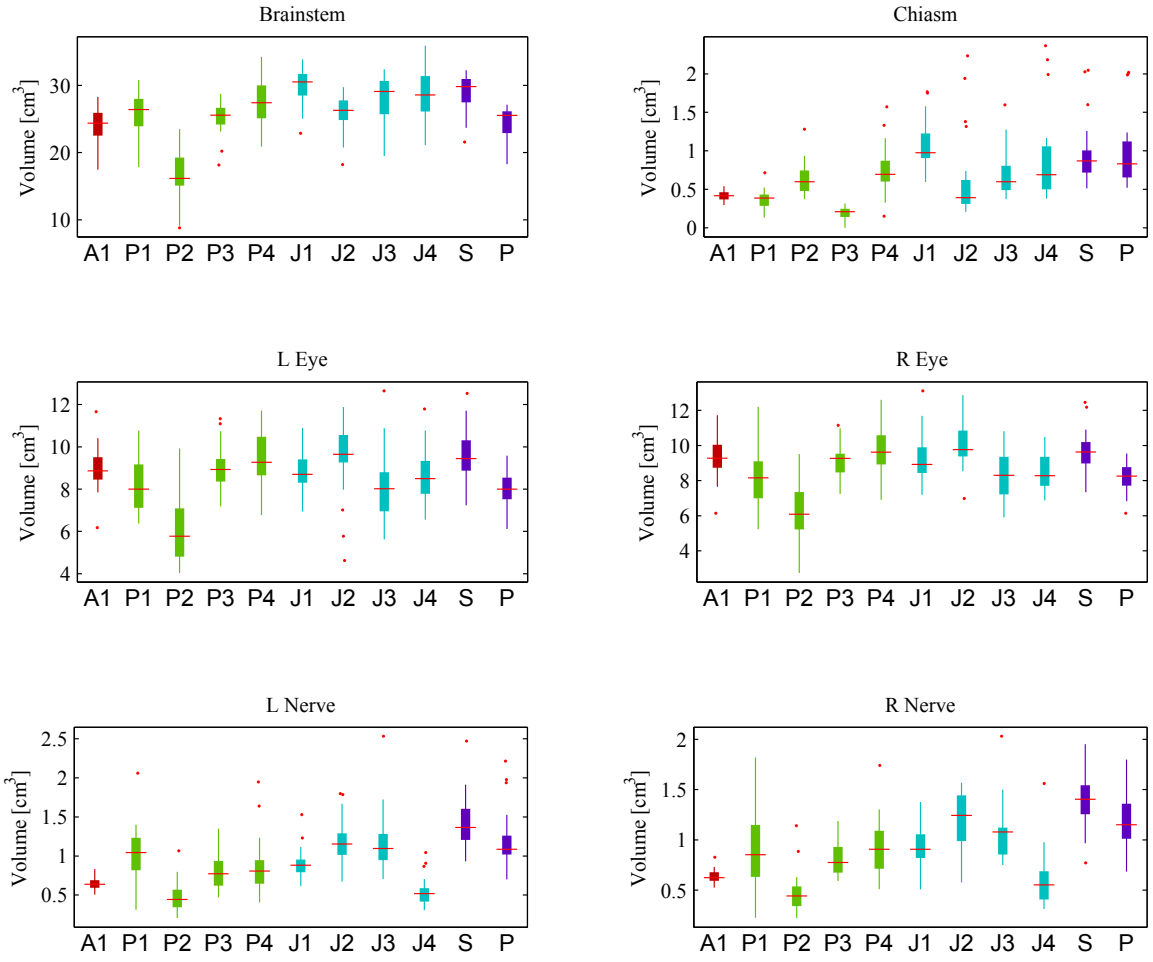

Figure II.4: Volume [cm^3] for the automatic (A_1), senior physician (P_1-P_4), junior physician (J_1-J_4), and simulated ground truth, STAPLE (S) and $PMAP_{mean}$ (P) segmentations. The horizontal line through each box indicates the median of the volume distribution while the rectangular box represents the interquartile range. Small dots are outliers for the distribution.

Dots above and below this line represent statistical outliers, or values outside 1.5 times the interquartile range. The thicker vertical line, the box, is bounded by the 25th and 75th percentiles of the distribution, and the median is shown via a short horizontal line. In these plots the automatic results are represented in the far left column labeled A_1 on the abscissa, followed by the senior physicians, P_1-P_4, and the junior physicians, J_1-J_4. In addition, when appropriate the two simulated ground truths are included and labeled S and P for STAPLE and $PMAP_{mean}$, respectively.

II.3.1 Volume

Figure II.4 plots the volume distributions of the automatic, expert, and simulated ground truth segmentations for each of the six organs investigated. First, we note that physician P_2 segmented smaller structures than the others except in the case of the optic chiasm. The brainstem as segmented by P_2 was on average 40% smaller than the other physicians' segmentations with twice the coefficient of variation, 24%. The mean volumes [and 95% confidence intervals] across all physician segmentations were 25.88 [25.08, 26.70], 0.66 [0.60, 0.74], 8.5 [8.20, 8.73], 8.69 [8.40, 8.96], 0.88 [0.81, 0.94], and 0.87 [0.82, 0.92] cm^3 for the brainstem, optic chiasm, left and right eyes, and left and right optic nerves, while the automatic volumes were 23.99 [22.82, 24.87], 0.41 [0.39, 0.45], 9.00 [8.53, 9.42], 9.26 [8.65, 9.71], 0.64 [0.61, 0.68], and 0.63 [0.61, 0.67] cm^3, respectively. The junior physicians as a group segmented larger structures than the senior physicians as a group. Although there were small differences in volume significant at the 5% level between the automatic structures and the physicians as a group, this difference disappears at an individual level. That is, the distribution of the automatic volumes falls within the variation of the individual physicians. It is clear, however, for the smaller tubular structures of the optic nerves and chiasm, the automatic structures were closer in volume to the smallest of physician segmentations. Additionally, the coefficient of variation of the automatic structures, 11-16%, was consistent across all organ structures. The individual physicians produced similar variation to the automatic system for the brainstem and eyes. For tubular structures, however, the physicians displayed more variation than the automatic segmentations, with coefficients of variation over the 20 patient cases ranging from 21-93% of mean structure volume.

Volumes for the two simulated ground truth segmentations were also calculated and can be seen in Figure II.4. STAPLE consistently produced segmentations with larger volumes than the p-map derived method.

II.3.2 Dice similarity coefficient

The Dice similarity coefficient (DSC), a measure of volumetric overlap, was calculated and plotted in Figures II.5 and II.6. Each boxplot contains several columns representing distributions of non-redundant pairwise DSC comparisons for each of the raters. Figure II.5 assesses inter-rater performance and variance. In the first column, A_1, represents the distribution of DSC between automatic segmentations and individual physician segmentations. Columns P_1-J_4 represent inter-physician comparisons: P_1-P_4 senior and J_1-J_4 junior physicians. Each distribution in these columns represents pair-wise comparisons of the expert in question to each of the other experts. The automatic segmentations are included only in the first column. In this way we are able to gauge automatic performance in the context of all experts as well as inter-expert performance. Table II.1 provides the mean DSC and 95% confidence intervals. The distributions of DSC are often skewed and depart from assumptions of normality required for statistical inference. To avoid making assumptions of the underlying population or transforming the data (Zou et al., 2004), confidence intervals were calculated via bias corrected and accelerated bootstrap (Davison and Hinkley, 1997) with 1000 replicates, about the mean DSC for each distribution plotted in figure II.5. In individual comparisons only P_2 produced segmentations with mean DSC different from the other physicians and the automatic system. Additionally, we calculated the same statistic grouping the experts as a single group and as two groups representing senior and junior physicians. At the 5% significance level across all raters, cases, and organs, no difference exists between the mean DSC of the automatic segmentations and the physicians as a single group. The junior physicians and A_1 performed better than the senior at the 5% level, but the magnitude of the difference was small.

Figure II.6 plots Dice coefficients against two ground truth estimations. The first two left most columns represent the distribution of Dice for the automatic segmentations compared to STAPLE and PMAP_{mean}, respectively. The same is plotted for the physician group in columns three and four. Lastly, the fifth column compares the two ground truth estimations. First, we note a high degree of overlap between STAPLE and PMAP_{mean}. Generally, the physician segmentations had a slightly higher spatial overlap with the ground truths than did the automatic system. However, the automatic system was more consistent, with smaller standard deviations and fewer outliers.

Figures II.5 and II.6 make essentially three types of comparisons: automatic-physician, physician-physician, and automatic- and physician-simulated ground truth. Another valuable comparison is that of individual groups to the simulated ground truths. For some structures, there was a small but significant ($p<0.05$) difference between senior and junior physicians. This difference was almost entirely a result of P_2 as a member of the senior group. Looking across

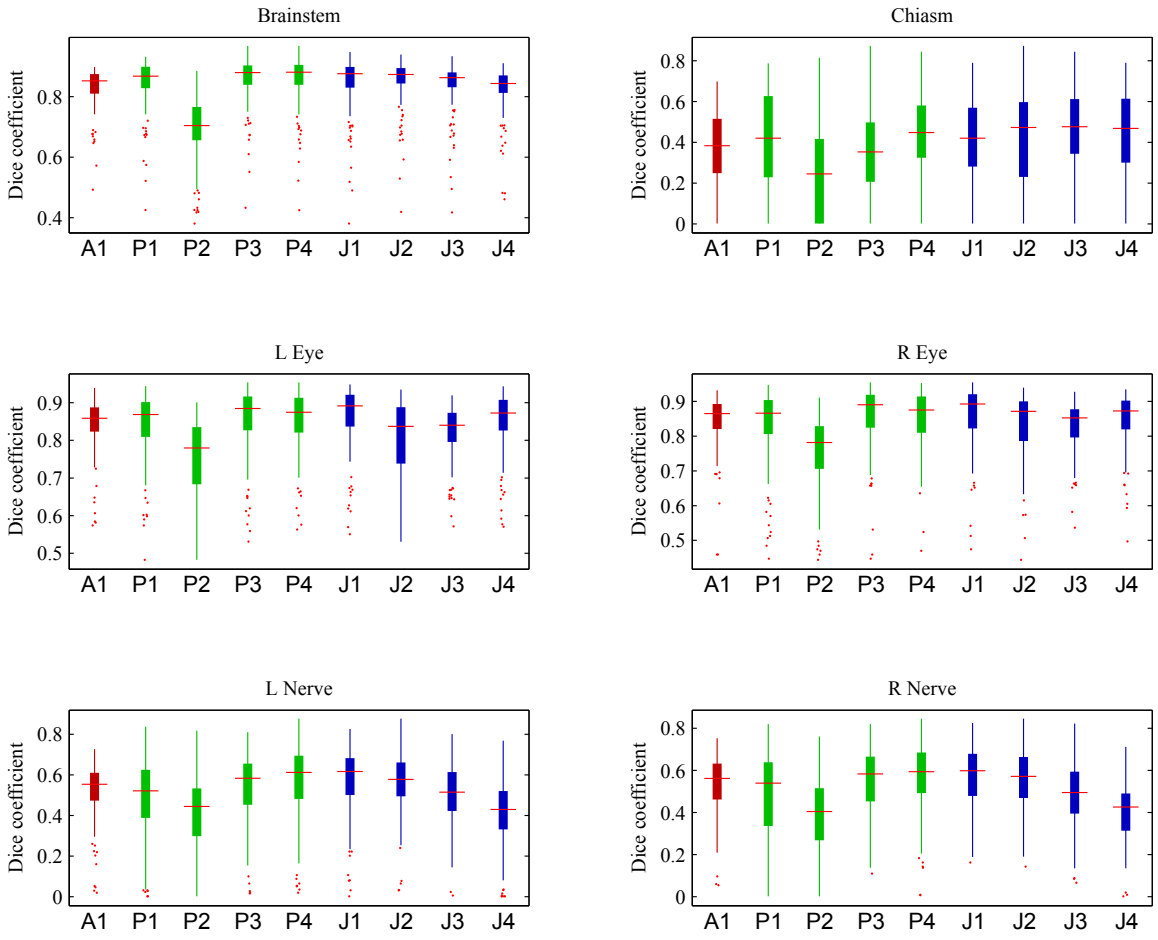

Figure II.5: Dice similarity coefficients across the 20 patients per structure to assess inter-rater performance and variance. Columns P_1-J_4 plot inter-physician comparisons: P_1-P_4 senior and J_1-J_4 junior physicians. Each distribution in these columns is comprised of pair-wise comparisons of the expert in question to each of the other experts. The automatic segmentations are included only in the first column.

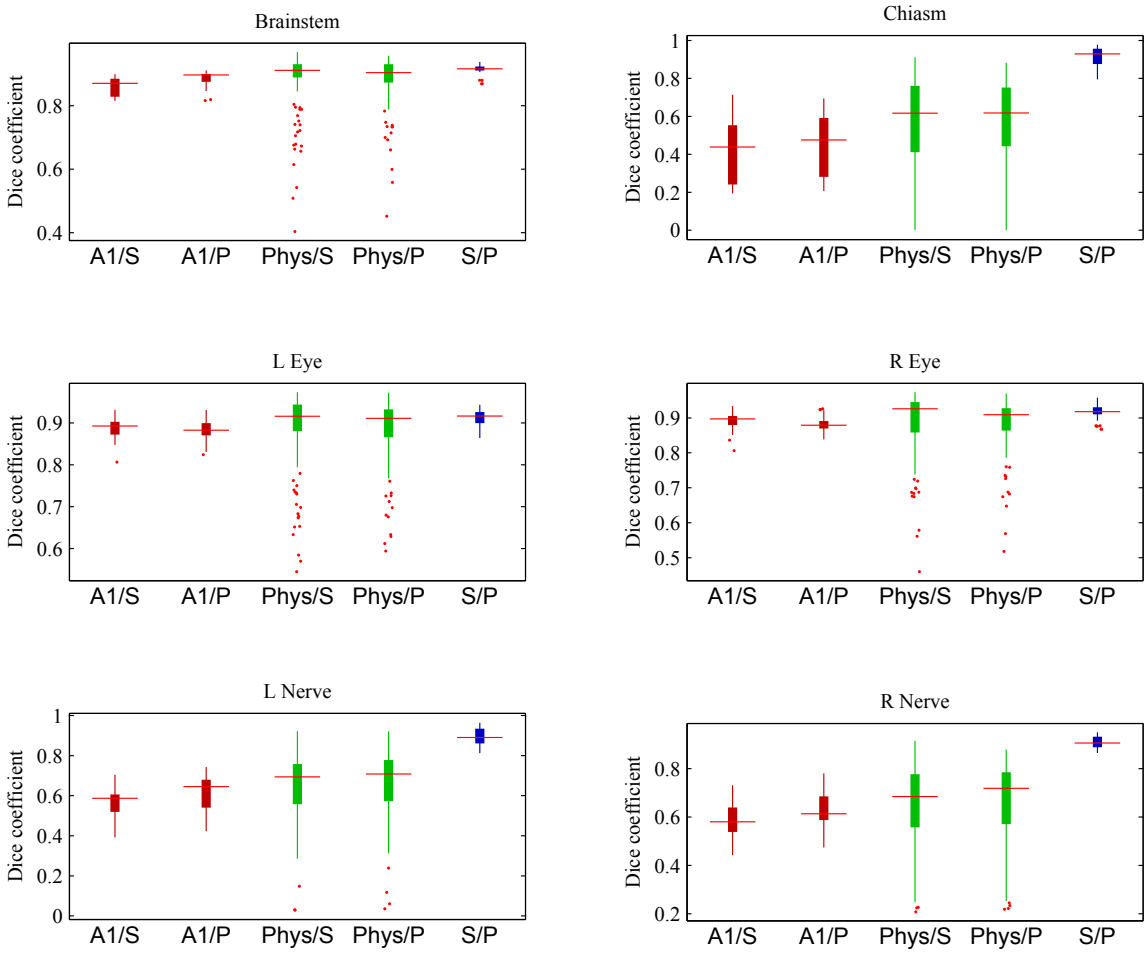

Figure II.6: Dice similarity coefficients for each rater group with respect to the simulated ground truths. The first two columns from the left compare A_1 to STAPLE (S) and $PMAP_{mean}$ (P), followed by comparison with the physician group, followed by comparison between S and P in the far right column.

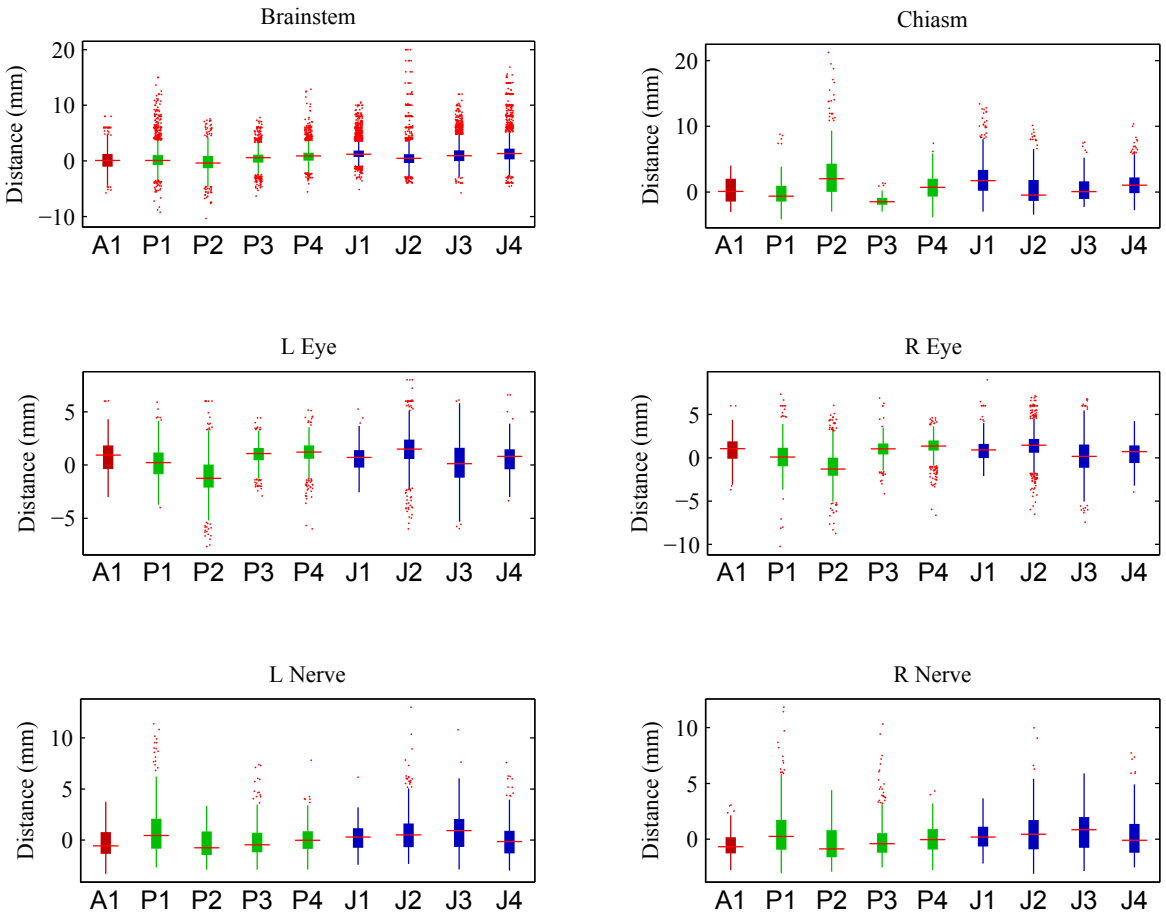

Figure II.7: Distance (mm) distributions from rater segmentations to PMAP_{mean}. Positive distances indicate a contour point lying outside the ground truth segmentation while negative distances indicate a contour point lying within the ground truth.

all the structures, the automatic segmentations produced a mean DSC against the simulated ground truths of 0.71 compared to 0.76 for the physicians. When decomposed into structures, again the biggest challenge was presented by the tubular chiasm and nerves, for both physicians and the automatic system. Whereas the mean for the brainstem and eyes was typically greater than 0.8, the chiasm and nerves were approximately 0.4 and 0.5, respectively. The tubular structures also had standard deviations on average over twice that of the brainstem and eyes.

II.3.3 Euclidean distance

Euclidean, or surface normal, distances were calculated in 3D between the segmentations and PMAP_{mean}. Signed distance maps for PMAP_{mean} were pregenerated using an implementation of the algorithm proposed by Maurer et al. (Maurer et al., 2003) and then evaluated at the contour points of the automatic and physician segmentations. In Figure II.7 the distances

Table II.1: Mean Dice similarity coefficients (DSC) with 95% confidence interval and standard deviation. Row A_1 gives the mean, 95% CI, and standard deviation on the distribution of non-redundant pairwise DSC comparisons of A_1 versus each expert rater. Similary, rows P_1-J_4 provide the statistics for physician-physician comparison (A_1 not included in these comparisons). The final three rows provide the same statistics grouping the physicians as senior, junior, or as one group including all experts.

Rater	All structures				Brainstem				Chiasm			
	Mean	Mean CI		std	Mean	Mean CI		std	Mean	Mean CI		std
A_1	0.656	0.642	0.671	0.223	0.830	0.818	0.839	0.064	0.374	0.345	0.403	0.179
P_1	0.647	0.626	0.664	0.252	0.845	0.828	0.856	0.082	0.400	0.360	0.439	0.235
P_2	0.543	0.528	0.560	0.248	0.691	0.670	0.706	0.105	0.253	0.218	0.293	0.225
P_3	0.666	0.652	0.681	0.245	0.856	0.839	0.866	0.078	0.344	0.308	0.379	0.219
P_4	0.686	0.669	0.700	0.223	0.855	0.840	0.867	0.084	0.437	0.407	0.470	0.194
J_1	0.683	0.668	0.697	0.224	0.843	0.826	0.857	0.091	0.414	0.377	0.446	0.193
J_2	0.667	0.651	0.683	0.223	0.851	0.837	0.862	0.078	0.416	0.381	0.455	0.239
J_3	0.652	0.638	0.668	0.222	0.836	0.820	0.847	0.080	0.447	0.409	0.478	0.209
J_4	0.626	0.609	0.643	0.255	0.822	0.806	0.834	0.078	0.423	0.389	0.461	0.231
All senior	0.619	0.606	0.633	0.256	0.798	0.781	0.813	0.122	0.322	0.298	0.353	0.224
All junior	0.669	0.659	0.680	0.219	0.859	0.855	0.863	0.034	0.476	0.450	0.499	0.192
All experts	0.646	0.641	0.652	0.241	0.825	0.819	0.83	0.099	0.392	0.378	0.405	0.226

Rater	Eyes				Optic nerves			
	Mean	Mean CI		std	Mean	Mean CI		std
A_1	0.843	0.831	0.853	0.070	0.523	0.501	0.544	0.138
P_1	0.836	0.820	0.851	0.095	0.482	0.450	0.515	0.193
P_2	0.754	0.734	0.770	0.102	0.403	0.377	0.429	0.162
P_3	0.855	0.838	0.868	0.090	0.543	0.517	0.567	0.158
P_4	0.851	0.835	0.862	0.083	0.560	0.525	0.587	0.178
J_1	0.860	0.845	0.872	0.085	0.560	0.532	0.586	0.161
J_2	0.818	0.800	0.834	0.099	0.549	0.523	0.572	0.149
J_3	0.824	0.812	0.836	0.073	0.490	0.466	0.514	0.154
J_4	0.849	0.834	0.861	0.077	0.407	0.381	0.431	0.150
All senior	0.810	0.799	0.819	0.110	0.486	0.468	0.503	0.195
All junior	0.842	0.834	0.848	0.076	0.497	0.484	0.511	0.152
All experts	0.831	0.827	0.835	0.094	0.499	0.493	0.507	0.174

are signed to differentiate a contour point lying inside from a point outside $PMAP_{mean}$. Table II.2 provides the minimum (furthest inside), mean, and maximum (furthest outside) distances for each structure averaged over the 20 patients. When the distance distribution is decomposed by structure, all raters had a mean distance between 0 and +2 mm except for P_2's segmentation of the chiasm, which on average was 3 mm from the simulated ground truth. The average maximum distances (inside and outside) across the 20 cases ranged from -4.3 to +5.4 mm (inside and outside) for the automatic segmentations. The same for individual physicians ranged from -5.8 to +10.8, and when physicians are considered as a group, -3.9 to +7.5 mm.

Figure II.8 plots the proportion of contour points that fall within 2 mm of the simulated ground truth as a function of rater and structure. This value can be thought of as the true positive rate, whereby any contour point drawn within a 2 mm shell of the simulated ground truth scores positive. The abscissa is partitioned by rater and structure, the ordinate is the 2 mm true positive rate, and the whiskers represent the 95% confidence interval on the proportion. This plot shows a broader variation amongst the physicians than within the automatic system. When we rank true positive rates, a senior physician, P_3, ranked the best overall and was the most consistent. The automatic system was second only to P_3 in terms of overall false positive rate and consistency.

II.3.4 Time

Segmentation time was recorded for each physician and is presented in Table II.3. The average physician time-to-segment was 14.5 minutes with a standard deviation of 6.2 minutes. These times include only the task of segmenting the organs and explicitly exclude all time required to open the software or make adjustments before delineation began.

II.4 Discussion

In this work we desired to evaluate our automated segmentations in a real-world clinical study, to test the hypothesis that automatically segmented structures could serve as a surrogate to manual delineations. Accordingly, we designed a large study and chose a cohort of 20 challenging patient cases containing large space-occupying tumors, which are generally challenging for registration algorithms (Dawant et al., 2002; Bach Cuadra et al., 2004; Bach Caudra et al., 2006). To our knowledge no other clinically evaluative study of this scale has presented data on segmentation under these circumstances. In the absence of a well-defined or well-suited ground truth, Warfield (Warfield et al., 2004) and Meyer (Meyer et al., 2006) have presented alternatives. The Simultaneous truth and performance level estimation (STAPLE) algorithm

Table II.2: Distances [mm] to $PMAP_{mean}$ simulated ground truth for each rater or rater group. The positive direction is outward looking from the ground truth while the negative direction is inward looking.

	Brainstem			Chiasm			Eyes			Nerves		
	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max
A_1	-4.30	0.15	5.39	-2.39	0.04	2.95	-2.33	0.82	3.81	-2.65	-0.40	2.41
P_1	-5.70	0.40	8.06	-2.26	-0.23	2.46	-3.35	0.14	2.90	-2.29	0.69	5.43
P_2	-5.84	-0.36	5.40	-0.65	3.13	8.46	-5.43	-1.07	2.74	-2.64	-0.41	2.56
P_3	-4.12	0.43	5.31	-2.39	-1.47	0.29	-1.72	1.04	3.19	-2.24	-0.09	5.17
P_4	-2.77	0.88	6.64	-2.25	0.62	4.16	-1.71	1.18	3.33	-2.19	0.07	2.91
J_1	-2.50	1.48	8.28	-2.03	1.89	8.86	-1.85	0.79	3.24	-2.10	0.25	2.83
J_2	-3.70	0.61	7.43	-1.99	0.13	3.56	-2.25	1.41	4.81	-2.20	0.59	5.27
J_3	-2.95	1.10	8.26	-2.15	0.50	4.70	-3.64	0.08	3.90	-2.36	0.76	4.52
J_4	-3.39	1.60	10.78	-2.06	1.24	4.79	-2.49	0.51	3.28	-2.39	0.13	3.17
All senior	-4.69	0.34	6.45	-1.86	0.55	3.92	-3.04	0.33	3.07	-2.36	0.05	3.89
All junior	-3.14	1.20	8.69	-2.06	0.94	5.48	-2.56	0.70	3.81	-2.26	0.43	3.95
All experts	-3.87	0.77	7.52	-1.97	0.73	4.66	-2.80	0.51	3.42	-2.30	0.25	3.98

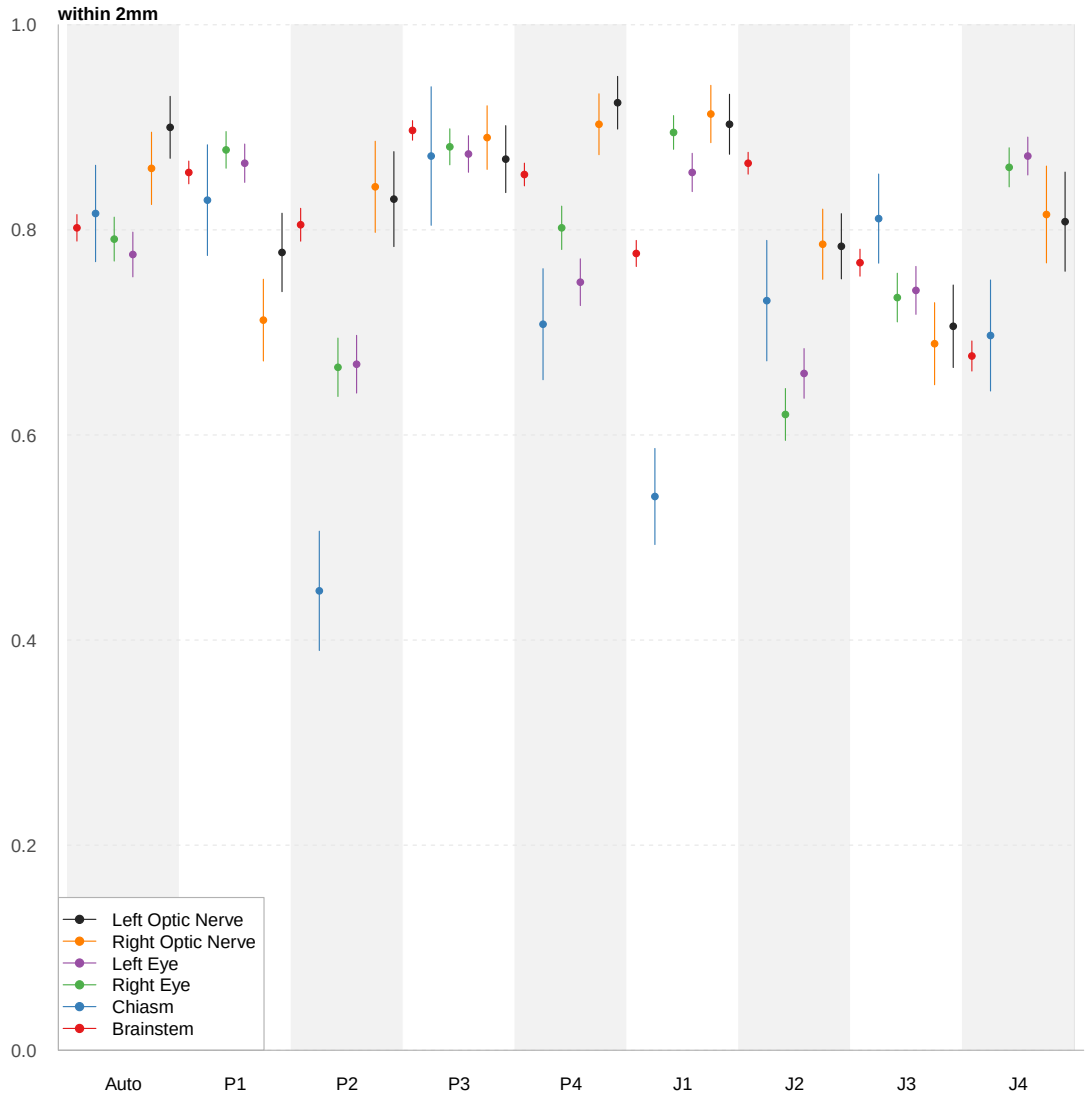

Figure II.8: True positive rate of contour points drawn within a 2 mm shell around the simulated ground truth. The abscissa is partitioned by rater and structure, the ordinate is the 2 mm true positive rate, and the whiskers represent the 95% confidence interval on the proportion.

Table II.3: Mean and standard deviation of segmentation times for the physician raters.

	Time [minutes]	
	Mean	std
P_1	9.6	2.5
P_2	14.1	4.5
P_3	19.8	3.4
P_4	21.1	6.2
J_1	18.8	3.3
J_2	14.8	4.4
J_3	6.6	1.4
J_4	11.1	3.0
All experts	14.5	6.2

produces a simulated ground truth from a cohort of expert delineations and can be compared directly with the automatic segmentations. STAPLE is a complex algorithm that has been shown to yield quality estimates of ground truth. However, early in our investigation we noted qualitatively that STAPLE could be influenced disproportionately by volumetrically larger segmentations within a group. Biancardi and colleagues (Biancardi et al., 2010) have noted a similar phenomenon. To provide an additional basis of comparison, we simulated a second ground truth using the computationally simple concept of probability maps, which is analogous to the idea of voting rule. We chose to threshold the probability maps at a variable level, the non-zero mean of each probability distribution, to form the mask. Previously, Biancardi chose to threshold at fixed levels such as 0.5 or 0.75, which tended to produce consistently large or small estimates, respectively. Thresholding the p-maps at a static, predetermined level is problematic for two reasons. First, determination of a threshold level presents a challenge. A reasonable first choice is 50% as it is the threshold for majority vote. However, with a statistically small number of raters of unknown individual variance, 50% may not be reliable depending on whether false positives or false negatives are more important. This suggests that a threshold appropriate for one cohort of experts may not be appropriate for another cohort. Likewise, the same logic applies to different organ types. We believe our results show that consensus among experts is quite dependent on organ structure. In a large structure such as the brainstem we found significant areas over which 100% of the experts agreed, but in the optic chaism and nerves such agreement was far rarer. Second, this method does not address the concern of spatial homogeneity. Both STAPLE and probability maps assume spatial independence of voxels. The STAPLE algorithm attempts to overcome deviations from this assumption through either incorporation of a priori information or using a Markov random field model. In our method we recognize that adjacent voxels are correlated, and in fact we increase that correlation through Gaussian smoothing. The

smoothing, however, helps achieve an approximately normal distribution of p-map values from which we calculate the mean probability as a threshold. Therefore, the appropriate threshold level will be unique to each cohort of expert segmentations and each structure. The end result shows that STAPLE and the probability map method produced ground truths of a high degree of spatial overlap (figure II.6). However, while a full investigation of STAPLE was beyond the scope of this work, we did find that STAPLE produced volumetrically larger segmentations than the pmap method (figure II.4).

In this work we used three principal metrics to characterize and compare segmentations: volume, the Dice similarity coefficient, and Euclidean distance calculated from a simulated ground truth. These measures offer several advantages. Volume is quite simple to calculate and stands alone, requiring no direct comparison to or use of a reference standard. The Dice similarity coefficient (DSC) is likely the most ubiquitous of metrics used in present literature. The Euclidean distances are particularly useful in the radiation therapy context, as their unit has implications to dose distributions and are well understood by the community. Volume, DSC, and Euclidean distance are invariant to image or mask size in terms of calculation, and thus do not suffer some of the pitfalls of specificity. A major goal of this work has been to provide a resource for others in algorithm assessment.

Each of the geometric measures showed the automatic segmentations to fall within the variation of the expert group, shown visually in boxplots (figures II.4–II.7). Generally, there were few statistical differences between the automatic system and the ground truth estimations or the physicians as a group and the ground truth estimations, which were evaluated via bootstrapping 95% confidence intervals. The automatic system produced less variance than the physicians as a group over all the organs, and the magnitude of variance was more consistent across organs than within the physician group. This can be seen in figure II.8, the 2 mm true positive rates.

Looking at individuals and groups within the larger physician group provides some trends. Junior physicians tend to segment volumetrically larger than their senior physician counterparts. We postulate this could result from a tendency to avoid risk of anatomically missing a portion of organ, while the more experienced physicians may be more confident in delineating a tighter border. We did not find, however, any evidence of reduced variance or higher spatial overlap in the senior physician group. In fact, one senior physician, P_2, was found to be different from the other physicians on all measures. A portion of this variance can be explained through the 2 mm true positive rates in figure II.8. The 2 mm true positive rate for P_2 is low for most structures but ranks fifth of nine for the brainstem, which was grossly different as measured by volume and DSC. Upon closer inspection we found that for

the brainstem this rater was inconsistent when marking inferior and superior extent of the organ, often not extending the slices as far in either direction as the rest of the group. This underscores the importance of choosing complementary metrics, as each examines a different aspect of geometry.

This work is not the first to evaluate automatic segmentation in the context radiation therapy organs at risk in the brain. Direct comparisons to other work are often compromised by the choice of metrics and differences in data acquisition. Bondiau and colleagues (Bondiau et al., 2005) investigated atlas-based segmentation of the brainstem using MR images of 6 patients and 7 experts. Here we compare their observations to (our observations). Inter-expert volumes varied from 16.70 to 41.26 (8.82 to 35.89) cm^3 across all cases. The mean expert delineations varied from 20.58 to 27.67 (19.66 to 29.15) cm^3, and the automatic delineations varied from 17.75 to 24.54 (17.47 to 28.28) cm^3 as a function of patient. Isambert (Isambert et al., 2008) also segmented the brainstem, optic chiasm, optic nerves and eyes, for 11 patients against a single reference standard jointly delineated by a radiation oncologist and neurosurgeon. They concluded that automatic segmentation was well suited for organs greater than approximately 7 cm^3, as they measured DSC above 0.8 for the eyes and chiasm, and concluded the small structures (DSC approximately 0.4) should be manually delineated by an expert. We noted a similarly low DSC for the chiasm in our study, though our optic nerves showed higher agreement of approximately 0.6 with respect to simulated ground truths. Though indeed spatial overlap is lower amongst the small tubular structures, we found that these structures are equally a challenge for the experts. In fact, in our study the automatically generated structures exceed the experts in some respects such as consistency, or robustness. This is seen in the variance of Dice index distribution (figure II.6) of the automatic against the simulated ground truths, which is smaller than the physicians. The automatic system also scored near the top of the expert group with respect to the 2 mm true positve rates plotted in figure II.8. Lastly, one must also consider the large variation in manual delineations for the optic nerves and chiasm reduces the accuracy of the ground truth produced with these contours.

II.4.1 Limitations and future work

We are undertaking a comprehensive clinical evaluation of our fully automatic segmentation system. Our experimental design is motivated by the following three observations. First, medical image segmentation is inherently a problem lacking a known ground truth. Accordingly, clinical evaluation studies should be behavioural in nature. Such a study requires a number of raters and patient volumes such as to provide good statistical power in the targeted clinical context. Second, realistically, the segmentation product of any automated system will require

review and most likely modification by a qualified professional. Evaluation should characterize the impact of the modification process on efficiency, individual and group rater variance, and accuracy. Third, in the radiotherapy context organ segmentations are an important variable in a complex process culminating in the delivery of radiation dose to a patient. Traditional approaches to evaluation focus on the geometric properties of the resultant segmentations. While these are certainly the first and an important part of any evaluation, much value exists in understanding the impact of an automated system with respect to dosimetry.

There were several complementary goals in this work. The first was to evaluate our automatic segmentation methods in the brain on clinically relevant organs at risk. To our knowledge, this study is the largest and most robust that has been offered to date for such organs, specifically in the presence of large space-occupying brain lesions. Second, we hope that this work will provide a framework and a basis for comparison to others implementing similar algorithms. We emphasize the importance of using multiple complementary and easily reproducible metrics, as well as experimental designs that recognize the behavioural nature of human medical image segmentation.

Lastly, there are several limitations to the current study. First, we evaluate only our own algorithm for automatic segmentation. There are now scores of segmentation methods based on a seemingly equal number of algorithms and body sites. It is difficult to make comparisons to other algorithms without making those comparisons directly within the same dataset. Second, we implemented this investigation at only a single site with physicians who have often trained and work together, and accordingly, may be systematically biased in their understating of anatomy or manual delineation in general. This is in part a result of time and logistics as these studies are time intensive and costly. We spent over a year collecting the manual segmentations for this analysis. Third, we have made considerable effort to characterize inter-physician variance but have not evaluated intra-physician variance, which could be important in parsing variance into real differences and randomness. Lastly, we have presented what we believe to be thorough though initial assessment of automatic segmentation within the context of radiation therapy. As these segmentations will undoubtedly be reviewed and modified by physicians in clinical practice, it is important to understand the impact of such a process on the workflow, consistency, and accuracy of segmentation as well as the final planned dose distribution.

CHAPTER III

IMPACT OF EDITING ON SEGMENTATION VARIANCE AND ACCURACY

Phys. Med. Biol. 58 (2013) 4071-4097

SEGMENTATION MODIFICATION IMPROVES EFFICIENCY WHILE REDUCING INTER-EXPERT VARIATION AND MAINTAINING ACCURACY FOR NORMAL BRAIN TISSUES IN THE PRESENCE OF SPACE-OCCUPYING LESIONS

M A Deeley[1], A Chen[3], R Datteri[3], J Noble[3], A Cmelak[2], E Donnelly[4], A Malcolm[2], L Moretti[5], J Jaboin[2], K Niermann[2], Eddy S Yang[2], David S Yu[2], and B M Dawant[3]

[1] Departments of Radiology and Radiation Oncology, University of Vermont, Burlington, VT
[2] Department of Radiation Oncology, Vanderbilt University, Nashville, TN
[3] Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN
[4] Department of Radiology and Radiological Sciences, Vanderbilt University, Nashville, TN
[5] Department of Radiation Oncology, Institut Jules Bordet, Université Libre de Bruxelles, Brussels, Belgium

Abstract

Image segmentation has become a vital and often rate limiting step in modern radiotherapy treatment planning. In recent years the pace and breadth of algorithm development, and even commercial ventures, have far outpaced evaluative studies. In this work we build upon our previous evaluation of a registration driven segmentation algorithm in the context of 8 expert raters and 20 patients who underwent radiotherapy for large space-occupying tumors in the brain. In this work we tested four hypotheses concerning the impact of segmentation editing in a randomized single-blinded study. We tested these hypotheses on the normal structures of the brainstem, optic chiasm, eyes and optic nerves using the Dice similarity coefficient, volume, and signed Euclidean distance error to evaluate the impact of modification on inter-rater variance and accuracy. Accuracy analyses relied on two simulated ground truth estimation methods: STAPLE and a novel implementation of probability maps. The experts were presented with automatic, their own, and their peers' segmentations from our previous study for modification. We found, independent of source, modification reduced inter-rater variance while maintaining or improving accuracy and improving efficiency with at least 60% reduction in contouring time. In areas where raters performed poorly contouring from scratch, modification of the automatic segmentations reduced the prevalence of total anatomical miss from approximately 16% to 8% of the total slices contained within the ground truth estimations. These findings suggest that contour modification could be useful for consensus building such as in developing delineation standards, and that both automated methods and even perhaps less sophisticated atlases could improve efficiency, inter-rater variance, and accuracy.

III.1 Introduction

Image segmentation is a vital step in most radiotherapy planning today. It describes a process that partitions imaging studies into discrete geometric information that can be used to plan and evaluate radiation treatment. The information usually consists of coordinate point sets or binary masks in the reference frame of the imaging study. Since the integration of x-ray computed tomography (CT) in treatment planning systems, segmentation of images has been used to optimize dose distributions by providing dose volume information of both targets and organs at risk. This is of particular importance in inversely planned therapy, such as intensity modulated radiation therapy (IMRT) and volumetric modulated arc therapy, and in situations such as stereotactic radiosurgery and other ablative methods that require high doses over a short time scale. Traditionally, images have been segmented manually in a time-consuming process that must occur before designing treatment fields or calculating dose. Our experience and that of others (Das et al., 2009) has been that segmentation is the rate-determining step in the treatment planning process.

In recent years a number of algorithms for automatic or semi-automatic segmentation have emerged, and quickly following several clinical systems have been marketed both within and as stand-alone to treatment planning systems. In the context of radiation therapy the vast majority of scholarly activity has involved algorithm development, and these algorithms have been quickly adapted to clinical systems with a relative lack of information regarding overall impact.

A potential explanation for the lack of evaluation studies involves the nature of segmentation itself. This is a problem lacking a known ground truth for comparison. Organ delineation in the human body requires decisions drawing from an aggregation of both explicit and implicit anatomic and physiologic information. Phantom studies, synthetic datasets and cadaver sections offer a more controlled but less realistic environment and hence are not well suited for gauging clinical impact. Several authors have shown previously that using a single expert rater as a gold standard is unreliable (Chao et al., 2007; Stapleford et al., 2010; Deeley et al., 2011). Isambert (Isambert et al., 2008) after noting low correlation with a single expert segmentation concluded that perhaps automatic segmentation was not well suited for small tubular structures such as the optic nerves and chiasm. Our previous study showed relatively low similarity between the automatic and expert segmentations as well. However, in the context of several experts, we found that the automatic system performed no worse than the experts. That is, the inter-rater variance amongst the experts was similar to the automatic-expert variance, indicating not that automatic systems are inadequate but that these structures are inherently difficult to segment.

Several methods (Warfield et al., 2004; Kittler et al., 1998; Meyer et al., 2006; Asman and Landman, 2012; Windridge and Kittler, 2003; Jacobs, 1995) have been proposed for estimating ground truth through a combination of expert segmentations. The method put forth by Warfield and colleagues, termed simultaneous truth and performance level estimation (STAPLE), is designed to incorporate truth priors and rater performance priors and be robust to outliers. However, truth priors are rarely known, and incorporating rater priors is problematic in clinical studies as relative rater quality generally cannot be anticipated accurately. In prior work we found that STAPLE tended to be influenced disproportionately by larger segmentations within the expert cohort. Biancardi (Biancardi et al., 2010) noted a similar phenomenon. This may be a byproduct of STAPLE depending heavily on a sometimes inaccurate truth estimate in the absence of a truth prior (Zhu et al., 2008). In our prior and current work, we rely on both STAPLE and another method, the computationally simple idea of probability maps (p-maps) (Meyer et al., 2006; Deeley et al., 2011), similar to voting rule (Kittler et al., 1998). Often the p-maps are thresholded at a predetermined level such as 0.50, where half of the raters agree. Recognizing that rater consensus may well be a function of organ type and location, we allow a moving threshold as determined by the p-map mean over the range (0,1] to be the best "vote" level for the ground truth.

Another persistent problem in the design of evaluation studies is the choice of comparison metrics. A number of volume and distance-based metrics have been used. Nominal volume (we use this terminology to disambiguate the use of volume from other meanings such as a three dimensional set of images or contours) is a useful measure that does not require pairwise calculation and is easily compared across separate studies. However, its value is that of a summary statistic. Two segmentations of different shape and location may have the same volume. Measures of spatial overlap such as the Dice (Dice, 1945) similarity coefficient and the Jaccard coefficient (Jaccard, 1908) provide pairwise comparison incorporating general shape and location information and are intuitive, but they do not provide information about whether differences are a result of over- or under-segmentation (Crum et al., 2006; Popovic et al., 2007). Additionally, volume and overlap measures by nature deemphasize central-peripheral (Meyer et al., 2006), or edge, deviations when they are small in comparison to overall volume. Distance measures, such as the Hausdorff and Euclidean distances, fill the gap by adding detailed information about edges.

We believe evaluation studies should be behavioural in nature, bringing together clinically relevant disease sites and imaging studies as well as enough raters and cases to provide robust statistical analysis. In our previous study we collected manual segmentations from a group of eight expert raters over 20 challenging cases. The experts delineated the brainstem, optic

chiasm, eyes, and optic nerves in the presence of large space-occupying lesions. We also used our algorithms to segment these organs automatically for each case. We tested the hypothesis that the automatic system would produce segmentations that could serve as surrogates to the manual physician segmentations, and we evaluated inter-rater variance and accuracy through simulated ground truths using STAPLE and our own application of the concept of thresholded probability maps. The results of this study, to which we will refer as the *de novo* study, have been published previously (Deeley et al., 2011). In summary, we found that differences in raters could be large and that at least one rater was often markedly different from the group. We also found that the automatic system performed well against the group of experts and, indeed, could serve as a surrogate.

Realistically, we contend that no automated system will completely replace expert segmentations in radiation therapy planning in the near future. However, automatic segmentation will and indeed already is offering a starting point to clinicians. From this starting point the clinicians will have to make judgments about the quality of the initial segmentations and make edits accordingly. Our *de novo* study provided information about expert delineation when starting from a blank slate but did not evaluate editing of pre-existing delineations. Building on the work of Chao (Chao et al., 2007) and Stapleford (Stapleford et al., 2010) in the present work we have undertaken a single-blind, randomized study presenting the same eight raters from the *de novo* study contours for editing. We tested four general hypotheses. First, editing the automatically generated contours (A_1) reduces inter-rater variance. Second, editing A_1 either increases or maintains accuracy. Third, editing A_1 salvages the results of low performing raters in the *de novo* study. In other words, raters who were low performers will produce better performing contours when they use A_1 as a starting point. Fourth, contour editing in general (independent of segmentation source) reduces inter-rater variation while maintaining or improving accuracy. Much of the methodology in terms of ground truth estimation and metrics was covered at depth in our prior work. Our attempt here is to refer the reader to the prior work as much as possible while maintaining clarity.

III.2 Methods

III.2.1 Study design

In this study we utilized imaging volumes from the same 20 patients used in our *de novo* segmentation study, as well as the same eight expert raters. Extensive descriptions of those imaging volumes, raters, delineation guidelines, and technical considerations are given in that work (Deeley et al., 2011).

The 20 patients had been previously treated at the Vanderbilt-Ingram Cancer Center with intensity-modulated radiation therapy (IMRT) for high-grade gliomas. Their cases were specifically chosen for the presence of large space occupying lesions often in close proximity to intracranial organs at risk, a situation that has both high clinical relevance (Amelio et al., 2010) and presents a challenge for automatic segmentation (Dawant et al., 2002). The images were x-ray computed tomography (CT) of 2 or 3 mm slice thickness and 1.5/3 T T1 magnetic resonance (MR) volumes of approximately 1 mm^3. These are typical of patients undergoing stereotactic brain biopsy. The raters were classified as senior (P_1-P_4, three attending radiation oncologists and one diagnostic radiologist) and junior (J_1-J_4, four radiation oncology residents in their final year of training).

In the first portion of our evaluation study the raters were asked to delineate brainstem, optic chiasm, eyes and optic nerves for the 20 cases utilizing fused CT/MR imaging within a clinical system. As they were given no starting point other than delineation guidelines and anatomical definitions, we refer to this as the *de novo*, or "from scratch" study. These delineations were acquired over a period of approximately one year.

Several months after concluding the *de novo* study, we initiated an editing study with the same raters. In this second round of contouring we presented the experts with fully completed contours for the brainstem, optic chiasm, eyes, and optic nerves from three sources: the automatic contours (A_1), their own contours (self), and contours delineated by their peers (peer) in the *de novo* study. In total each expert edited 60 complete sets of segmentations, three per patient. These 60 tasks were randomized and single-blind, in that the raters did not know the origin of the segmentations. In fact, to avoid presumptive guessing, we told the raters only that they would be presented segmentations for editing. We made no mention of the potential sources of segmentations, though it is likely some of them assumed the source to be the automatically generated contours. Though it was beyond the scope of this work to test, we anticipated that this time interval would be sufficient to avoid potential effects of memory on the raters' interpretations. Additionally, if effects were to exist within the current study as a result of revisiting each patient three times, these would be randomly distributed over the patients and segmentation sources.

An in-house graphical user interface was developed to present a task queue and to record editing times. The design was such that each rater was presented with each of the 20 automatic segmentation sets once, each of their own previous segmentations at least once and sometimes twice per patient, and their peers' segmentations from the *de novo* study. We will refer to these groups as "A_1", "self", and "peer", respectively. The selection of which peer to edit was also randomized and balanced such that each rater edited each peer two to three times

over the course of the tasks. The editing was done using a research version of the treatment planning system (Eclipse 8.5, Varian Medical Systems, Palo Alto, CA) identical to the clinical system. Details of this system were included in our previous work (Deeley et al., 2011). We did not specify to the experts which tools to use for editing. Several options were available such as using a paintbrush tool to take away or add to an existing contour, deleting a contour and redrawing from scratch, and moving the contours as a whole. We did not collect data on tools utilized, though generally most experts appeared to prefer the paintbrush method for making edits.

III.2.2 Automatic segmentation

The automatic segmentations presented for editing were the same generated in the *de novo* study. In summary, we segmented the organs at risk using two methods. An atlas-based, registration-driven method was used to segment the brainstem and eyes. It involves a global affine registration of the atlas to target, followed by automatic extraction of a predefined bounding box from both target and atlas. A second, now local, affine registration is performed on the bounded region, resulting in a transformation projecting the atlas to the target. Normalized mutual information (NMI) (Studholme et al., 1999) is used as the similarity measure. A local non-rigid registration is performed between the results of the local affine registration and the atlas. Lastly, the deformation fields resulting from the three registrations are used to project contours from the atlas to the target (patient) image.

The second method, used to segment the optic chiasm and nerves, is a technique we have developed for the segmentation of tubular structures, termed the atlas-navigated optimal medial axis and deformable model algorithm (NOMAD) (Noble and Dawant, 2011). NOMAD first computes the medial axis of the structure as the optimal path with respect to a cost function relative to image and shape features, and then expands using a level-set algorithm to the final structure. The statistical model employed in NOMAD was trained on image volumes outside those used in this study.

The non-rigid transformations used in our segmentation framework are provided by the adaptive bases algorithm (ABA) we have developed (Rohde et al., 2003). It utilizes NMI and models the deformation field registering the atlas and target as a linear combination of radial basis functions (Wu, 1995) with finite support.

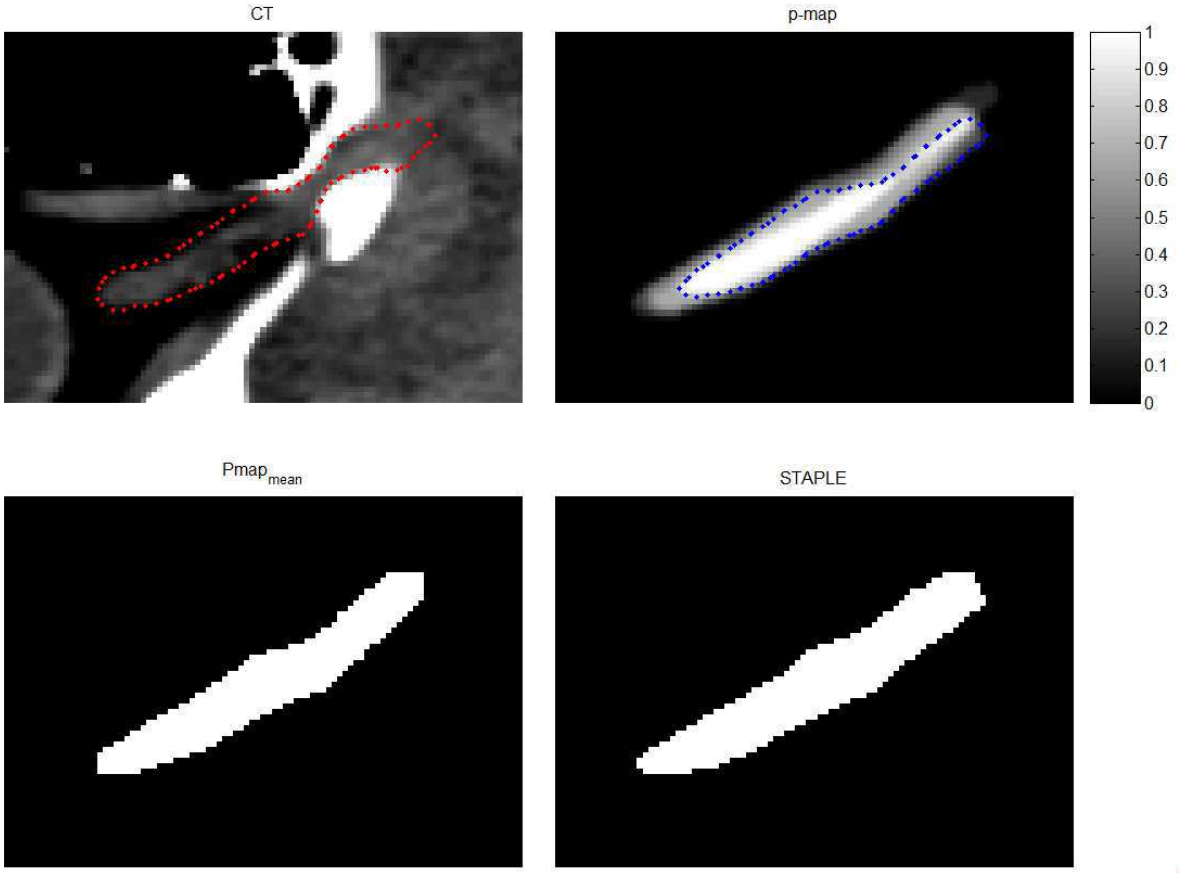

Figure III.1: Ground truth estimation. The upper left panel displays the area of an optic nerve for patient 12 on an axial CT slice. The dotted contour is the automatic segmentation after editing by rater P_3. The upper right panel plots the p-map used to estimate a ground truth for comparison against P_3 and consists of his peers' segmentations. The contour overlaying the p-map is the unedited automatic result. The ground truth estimated by thresholding the p-map at the non-zero mean is shown at lower left, and the STAPLE estimation for the same slice at lower right.

III.2.3 Ground truth estimation

To gauge accuracy we calculated ground truth estimations via two methods: the simultaneous truth and performance level estimation (STAPLE) algorithm (Warfield et al., 2004) and through dynamically thresholded probability maps (p-maps) (Meyer et al., 2006). We calculated STAPLE and the p-map derived ground truths in the same manner described previously (Deeley et al., 2011). In that study we found the two methods produced ground truths estimates with a high degree of overlap as measured by DSC, even for the small tubular structures. However, much of the impetus to use two independent estimates arose from a qualitative observation that STAPLE could be influenced disproportionately by under-segmenting individuals. We also noted that though spatial overlap between the two methods was consistently high, STAPLE also consistently under-segmented compared to the p-map method. For the purposes of this work we define under-segmented structures as those that are volumetrically larger than a reference structure.

STAPLE is designed to be robust to bias (Warfield et al., 2004), so we applied it as has been commonly done in other work (Stapleford et al., 2010) to create a single ground truth per patient from a cohort of experts. With the p-map method we removed bias explicitly and thus calculated a different ground truth mask for each rater in a leave-one-out process. That is, the p-map derived ground truth for rater P_1, for example, is generated from the p-map which excludes his own segmentations. We also eliminated all segmentations delineated by rater P_2 from the pool as evidence from the *de novo* study showed this rater was often different from the rest of the group. The p-maps are calculated as

$$\mathrm{p-map}_{i,j,k} = \left[\frac{1}{6}\left(\sum_{n=1}^{8}(E_{i,n,k}) - E_{i,2,k} - E_{i,j,k}\right)\right] \times K_\sigma \tag{III.1}$$

where i, j, and k, represent the patient case, rater, and structure, respectively; $j = 2$ indicates the rater P_2; E represents the binary mask; K is a Gaussian kernel of 3x3 pixels applied to each slice of the mask with standard deviation $\sigma = 0.65$ pixel width. A full discussion of method and rationale for thresholding the p-maps can be found in our prior work (Deeley et al., 2011).

The *de novo* and editing studies resulted in a combined 32 sets of manual segmentations for each patient (in total, 640 structure sets and 3840 individual organ structures): 8 *de novo* (P_1-P_4, J_1-J_4), and on average, 8 edited automatic (A'_1), 9 edited self (self), and 7 edited-peers (peer). From these we calculated ground truth estimations to provide a basis for accuracy assessment and calculation of distance maps (discussed in section 2.4.2). Figure III.1 illustrates a p-map for an optic nerve with a single physician contour as well as the correspond-

ing Pmap_{mean} and STAPLE ground truth estimations. An important choice had to be made as to which of the segmentations cohorts to draw from in these calculations. One could envision using all of the manual *de novo*, the manual-edited, those groups combined, or individual sets of manual-edited segmentations (e.g., edited-peers). We chose to base all analyses in this study on the class derived from the edited-peers group, the rationale for which we discuss in section 4.

III.2.4 Metrics for comparison

Volume-based metrics

We calculated two volumetric measures in this study: nominal volume and Dice similarity coefficient. Nominal volume is calculated as

$$\text{Volume} = |E_{j,j,k}| \times v_i \tag{III.2}$$

where the binary mask E for patient i, rater j, structure k, is summed over its voxels of volume v. Volume provides a summary statistic about the gross size of segmentations and can be compared readily to results from other studies utilizing different datasets as there is no dependence on ground truth estimation, group variance, or image dimensionality.

The Dice similarity coefficient (DSC) is a spatial overlap measure that can be calculated generally via

$$\text{DSC} = \frac{|E_A \cap E_B|}{\frac{1}{2}\left(|E_A| + |E_B|\right)} \tag{III.3}$$

where E_A and E_B denote any two mask volumes of the same dimensionality. Its range is [0,1], where zero signifies no overlap and 1 signifies exact overlap. The DSC can be calculated as an integrative measure over the volume segmentations or on a slice-by-slice basis. In this study we calculated DSC on volumetric basis to measure inter-rater variance, and assess accuracy, while the slice-by-slice implementation was used only in gauging amount of editing.

Distance-based metrics

Distance-based metrics complement the volume-based metrics by providing information about differences between segmentations at their edges, independent of object shape. There are a number of methods for calculating generalized distance measures; a discussion toward a generic evaluation of image segmentation using the concept of distance is provided by Cardoso (Cardoso and Corte-Real, 2005). In this work we were concerned with the end-use of segmentations in radiation treatment, an environment well-suited to Euclidean distances, sometimes

known as *ordinary* or surface normal distances.

Signed three-dimensional Euclidean distance maps were pre-calculated from the Pmap_{mean} simulated ground truths. Each voxel in the distance map contains the Euclidean distance between that voxel and the nearest edge voxel of the ground truth. Distance distributions were formed for individual rater segmentations by sampling the appropriate distance map with contour points for the segmentation in question. Contour points lying inside the boundary of the ground truth were signed negative and those lying outside signed positive. The distributions were used to calculate average min, mean, and max distances across the patients, raters, and structures.

Additionally, we calculated a quantity which we term the true positive rate. It is the fraction of total contour points falling within a shell of a specified distance from the edge of the simulated ground truth, Pmap_{mean}. We chose $\pm$ 2 mm as a relevant distance for selection of the shell. It is on the order of the slice thickness, and while one would want to minimize uncertainty from segmentation as much as possible, 2 mm is on the order of the overall geometric accuracy of most linear accelerators.

Several aspects of this implementation are noteworthy. First, as the distances are signed, we avoided summary statistics without also examining the distribution, as measures of central tendency could be washed out by positive and negative variations. Second, as the distance distributions are calculated in one direction only, from rater segmentation to ground truth, each distribution contains information regarding exclusively where a rater segmented, but not where a rater elected not to segment. In that sense, a rater could delineate one slice of a multi-slice structure and have a distance distribution of zeroes. In the absence of complementary volume-based measures, this would be a weakness. Lastly, as the distances are calculated from ground truths, they do not provide a direct pair-wise comparison of the same flavour as DSC, and hence they are less valuable in determining inter-rater variance than DSC and nominal volume.

Time-to-edit

Time is an important factor in the process of treatment planning, which can be a complex multi-step workflow with a number of checkpoints requiring input from several professionals. In this study we measured the time required by physicians to modify pre-generated segmentations, as discussed in previous sections. This was accomplished via an in-house task queue and timing program. To be as clinically realistic as possible, the software alerted the expert to the current task in need of attention and allowed for pausing and restarting. The experts were instructed not to run the timer during administrative tasks such as opening and

closing patients.

III.2.5 Analytical framework

We use the measures discussed herein in combination to test the hypotheses laid out in section 1. One tool that we use repeatedly to present the reader with a visual summary of the data is the boxplot (Tukey, 1977). Each boxplot divides the distribution into quartiles q1-q4, where the inter-quartile range q3-q1 is represented by a vertical rectangular box, and vertical lines, also known as whiskers, extend past the box to represent the statistical range of the distribution; outliers are shown as dots plotted individually that fall beyond 150 percent of the inter-quartile range. The median of the distribution is shown as a red horizontal line within the box. In cases where the distribution satisfies conditions for normality, notches, were used to provide information about significance in differences at the median. In these plots notches were represented via triangles, whose centers delimit the edges of the 95% confidence interval about the median.

Some distributions, such as that of DSC, are not normally distributed, and in such cases we have calculated measures of central tendency and 95% confidence intervals via bias corrected and accelerated bootstrapping with 1000 replicates (Davison and Hinkley, 1997). Others have achieved normality by transformation of the data, such as by using the `logit` function (Zou et al., 2004).

III.3 Results

III.3.1 Assessing editing efficiency

One aim of introducing automation into the segmentation process is to improve efficiency. We measured two variables to gauge efficiency: time to edit pre-generated contours and amount of editing required for a satisfactory end product. All measures of quality being held equal, one would choose a process that minimizes both of these factors. In our previous study we found that the experts required a mean time of approximately 14.5 minutes (total range 4.5-31 minutes, individual means 6.5-21 minutes; $N = 107$ contouring sessions as times were not collected for first six patients and on occasion raters forgot to start the timer) to segment the brainstem, chiasm, eyes, and optic nerves utilizing fused CT/MR, though individuals varied widely. Editing required considerably less time than contouring from scratch. Panel (a) of figure III.2 plots the distribution of times across all raters for the *de novo* (P_1-J_4) and editing (P'_1-J'_4) studies for the group of tasks in which the raters edited A_1, the pre-generated automatic segmentations produced by our algorithm. Editing of A_1 reduced mean time to final product

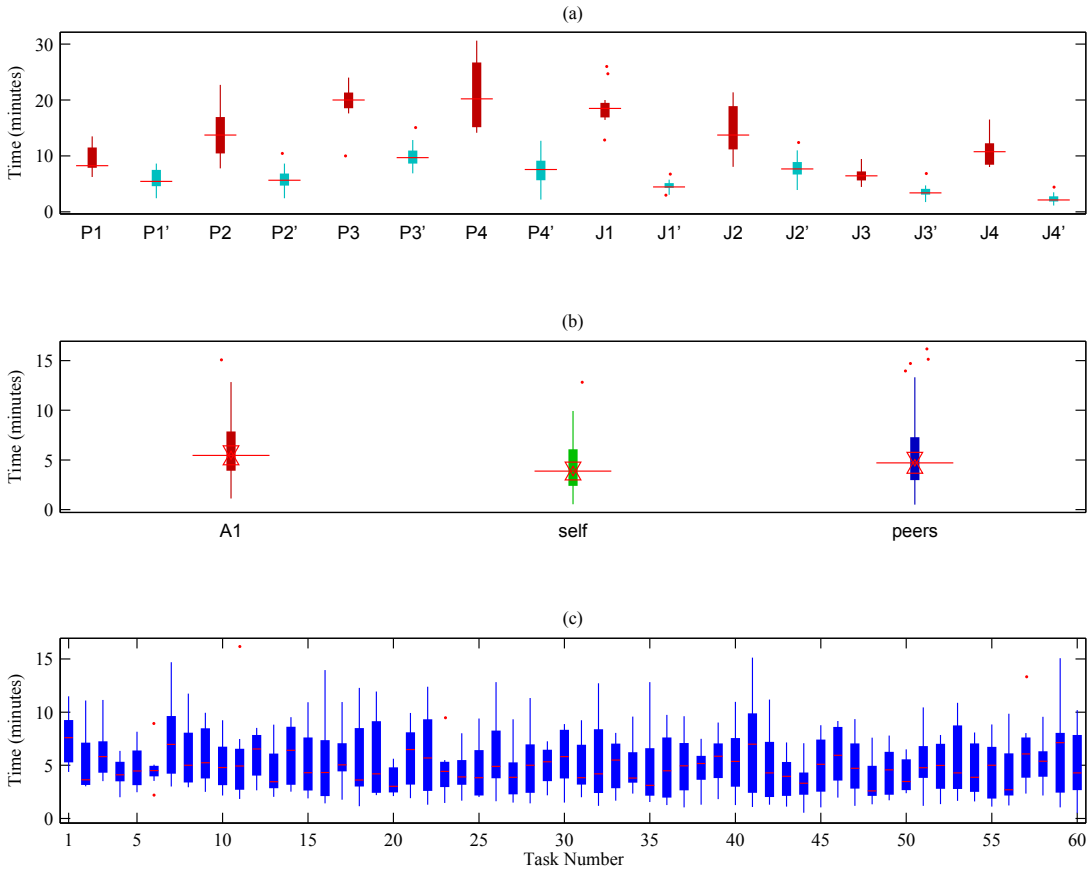

Figure III.2: Time Analysis. Panel (a) plots the distribution of times across all raters for the *de novo* (P_1-J_4) and editing (P'_1-J'_4) studies for the group of tasks in which the raters edited A_1, the pre-generated automatic segmentations. Panel (b) compares the distributions across the three sources for editing: automatic (A_1), self, peer. In all each rater completed 60 randomized tasks over the course of the editing study. Panel (c) plots the time to modify as a function of task to evaluate whether there was a learning effect.

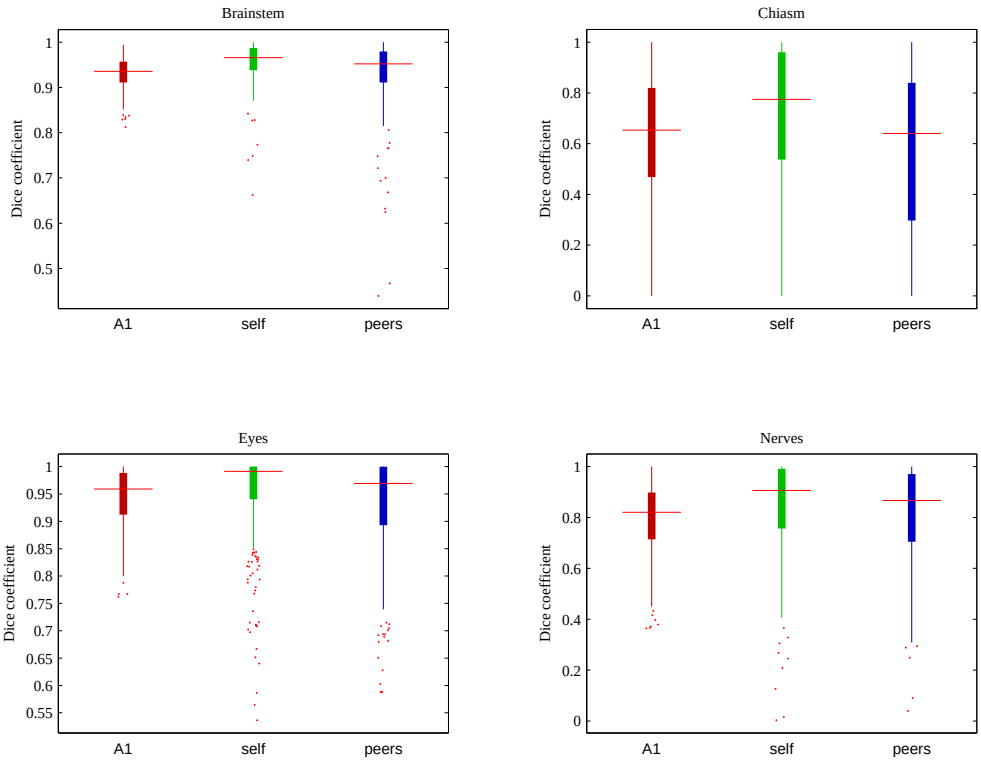

Figure III.3: Plot of volumetric Dice coefficient as a function of source for editing: A_1, self and peers. Each distribution consists of the pairwise comparisons between the *de novo* and edited segmentations for each of the eight raters. For example, P_1 *de novo* is compared via DSC to P_1 from each of the editing sources to gauge the similarity (or, equivalently, amount of editing).

to 5.9 ([5.5, 6.4] 95% confidence interval) minutes, as did editing of their own (self) contours to 4.3 [4.0, 4.7] minutes, and those of their peers to 5.5 [4.9, 6.1] minutes. Panel (b) compares the distributions across the three sources for editing: automatic (A_1), self, peer. We found there was a significant ($\alpha = 0.05$) though small reduction in time when raters were presented with their own contours segmented in the previous study as compared to those of the automatic system or their peers. As this was a task-oriented study conducted over approximately a year, we wondered if there would be an effect of learning or even potentially fatigue on time to modify. We randomized the tasks over the patient population to avoid confounding case difficulty with experience and found that taken as a group (panel (c)) there was no learning effect. We similarly found there was no influence of task number on accuracy as measured by DSC against the ground truth estimations.

To keep the study as clinically relevant as possible and the timing procedure valid,

Table III.1: The range of DSC [0,1] is divided into four categories to gauge amount of editing as a function of structure and source: major [0,0.7), moderate [0.7,0.9), minor [0.9,1), and no [1,1] editing. Each cell contains the fraction of slices via 2D DSC calculation that fell within a given range.

	Brainstem			Chiasm			Eyes			Nerves		
	A_1	self	peers	A_1	self	peers	A_1	self	peers	A_1	self	peers
none	0.28	0.43	0.32	0.15	0.29	0.17	0.37	0.59	0.43	0.26	0.44	0.37
minor	0.42	0.41	0.44	0.10	0.13	0.11	0.40	0.24	0.29	0.13	0.14	0.16
moderate	0.21	0.10	0.14	0.14	0.15	0.12	0.15	0.11	0.19	0.25	0.17	0.18
major	0.09	0.06	0.09	0.62	0.43	0.60	0.08	0.06	0.09	0.37	0.25	0.29

we did not ask raters to comment directly on the quality of the contours presented (Stapleford et al., 2010). Rather we gauged acceptability using the Dice coefficient in a pairwise calculation between the pre- and post-editing masks, both by slice and volumetrically. For example, a DSC of 1.0 indicates unequivocally that the initial segmentation matches the final segmentation. As the similarity between pre- and post-editing segmentations increases, DSC increases as a function of the overlap relative to volume or area of the segmentations, indicating smaller changes were made. The results of the volumetric calculation are shown in figure III.3, where distributions of DSC are plotted as a function of source for editing and structure. Most edits to the brainstem and the eyes resulted in a less than 10% change in spatial overlap, whereas the chiasm and nerves required more extensive editing across all sources. Mirroring the data concerning time, there was a small preference of raters for their own contours compared to the automatic and those of their peers. To evaluate the amount of editing by slice, we divided the range of DSC, [0,1], in Table III.1 into four categories: major [0,0.7), moderate [0.7,0.9), minor [0.9,1), and no [1,1] editing. Here there was a clear preference of raters for their own contours, of which they made no edits to 43%, 29%, 59% and 44% of the contours for brainstem, chiasm, eyes, and optic nerves. The chiasm and nerves underwent substantial heavy editing regardless of the original source: A_1, self, or peers.

Figure III.4 contains four panels of orthogonal MR cross-sections comparing (a) the *de novo*, (b) A_1-edited, (c) self-edited, and (d) peer-edited groups. In this example the *de novo* contours display the most variance. The unedited A_1 contours are included in red in panel (a). We also note the erroneously contoured internal carotid arteries as part of the optic chiasm shown with red arrows in the coronal (upper right) section of panel (a). Editing of A_1 reduced inter-rater variability the most and eliminated the inclusion of the internal carotids for all but one rater.

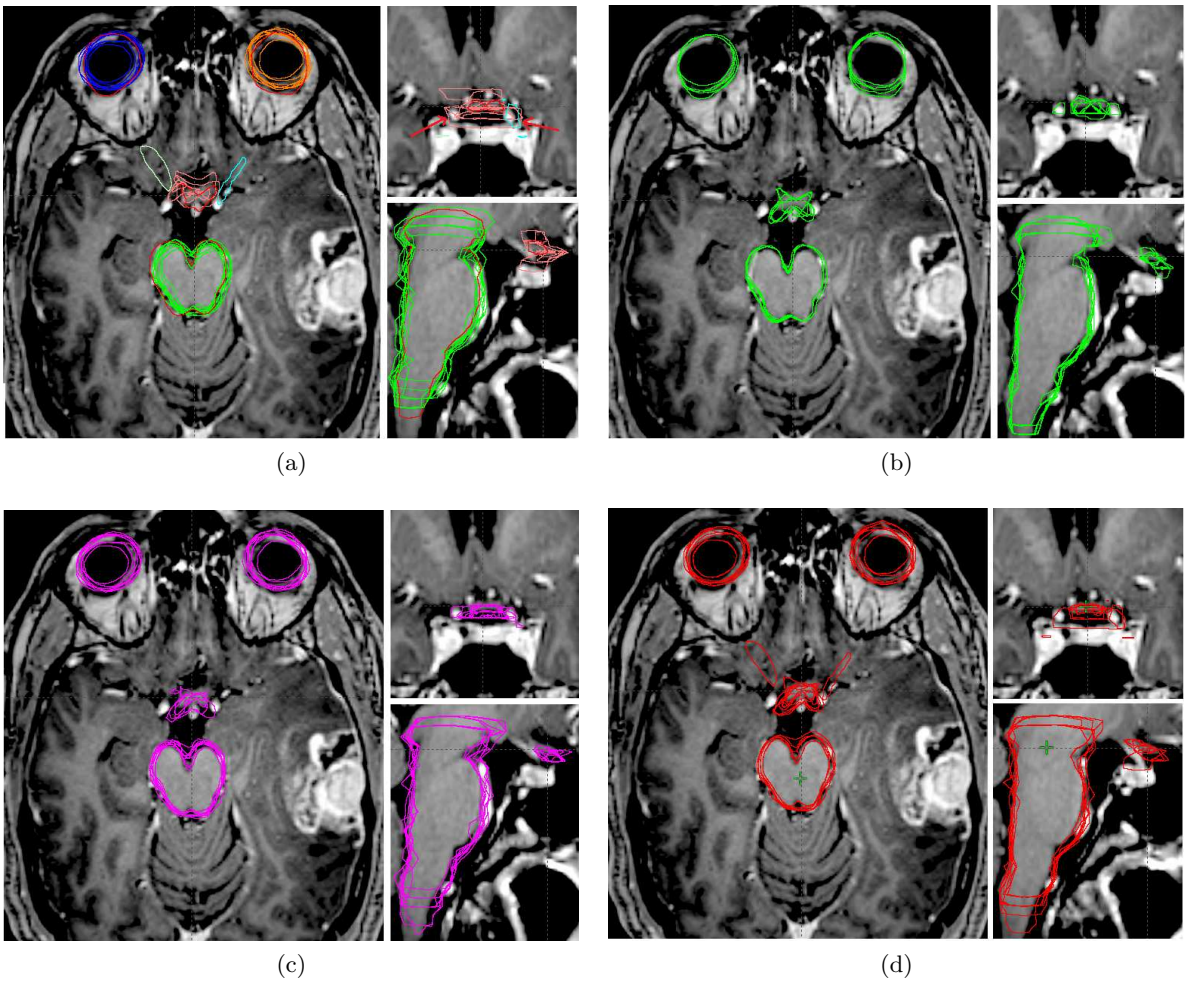

Figure III.4: Orthogonal views comparing group results from (a) *de novo*, (b) A_1-edited, (c) self-edited, (d) peer-edited. The red arrows in the upper right (coronal section) of panel (a) point to the internal carotid arteries, which were often erroneously included as part of the optic chiasm in the *de novo* study as well as self- and peer-edited groups. In panel (a) the red contours are those of the A_1 while the other colors represent manual expert segmentations.

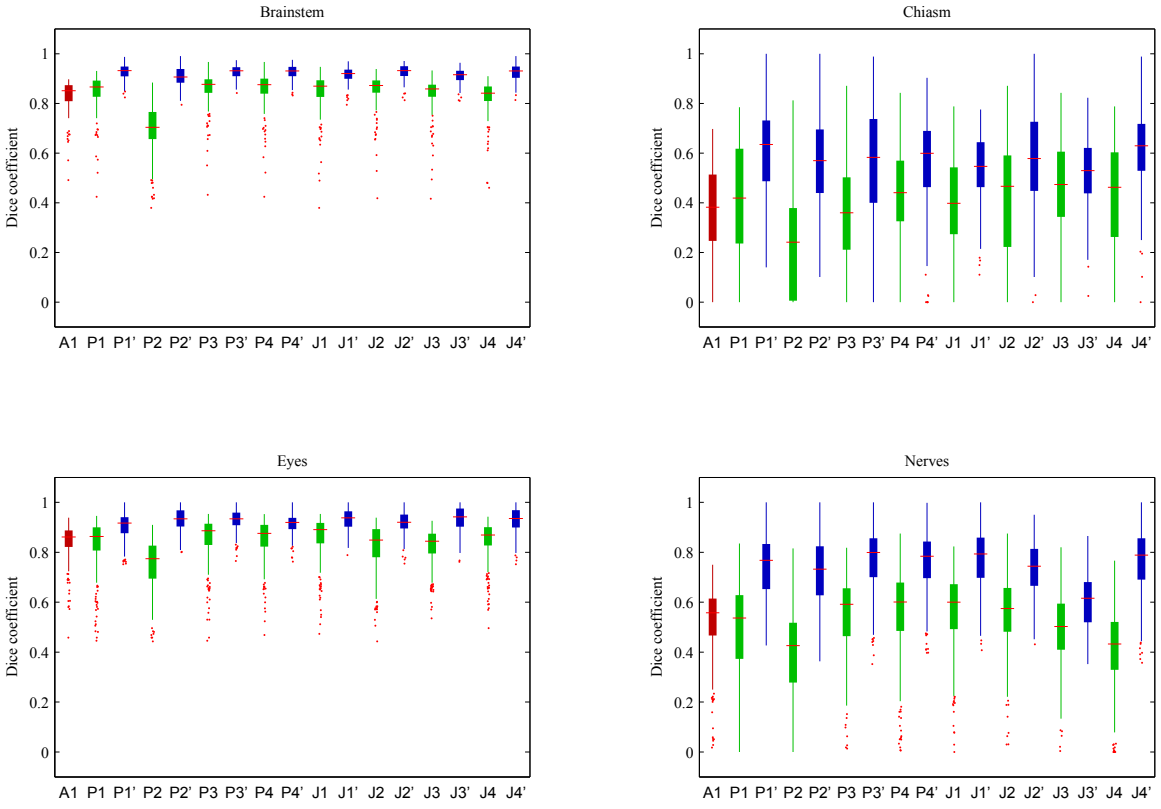

Figure III.5: Plots the distribution of volumetric Dice coefficients for the editing of the automatic (A_1) over all structures and raters providing a sense of inter-rater variance from *de novo* study versus the editing study. Columns A_1, P_1-J_4 plot the distributions of non-redundant pairwise DSC from the *de novo* study. Primed columns, P'_1-J'_4, denote the editing study results.

III.3.2 Evidence regarding hypothesis: Editing of automatic segmentations (A_1) reduces inter-rater variance

Each of the physician raters was presented complete sets of automatic segmentations (A_1) for each of the 20 patients as discussed in section 2.1. This process was blinded and randomized along with the presentation of self and peer segmentations. Figure III.5 plots the distribution of volumetric Dice coefficients over all structures and raters providing a sense of inter-rater variance from *de novo* study versus the editing study. Columns A_1, P_1-J_4 plot the distributions of non-redundant pairwise DSC from the *de novo* study. Primed columns, P'_1-J'_4, denote the editing study results. These distributions relate inter-rater variance in two key ways. The first is simply the DSC statistic itself, which can be summarized via the mean or median. The red horizontal lines in the boxplot represent the median. The means and corresponding 95% confidence intervals and standard deviations were calculated via bootstrapping and are presented in Tables III.3 and III.4. In figure III.5 it is clear that across all structures and raters the median DSC increased with editing of the automatic contours. The gains were largest for the chiasm and nerves, though even after editing agreement was still less than seen with the brainstem and eyes. The mean inter-rater DSC treating all raters as a single group increased from 0.83 (*de novo*) to 0.92 for the brainstem, 0.39 to 0.57 for the chiasm, 0.83 to 0.93 for eyes, and 0.49 to 0.73 for the optic nerves when the raters were presented with automatic contours for editing. The second way DSC relates inter-rater variance is through the spread of these distributions, which decreased as a result of editing such that there was both a reduction in outliers and standard deviation.

Similar to figure III.5 nominal volume is plotted in figure III.6, and the corresponding mean, confidence interval, and standard deviation, are present in table III.5 for all raters as a group and as a function of source: unedited A_1, *de novo*, as well as edited A_1, self and peer groups. Editing of A_1 resulted in reduction in inter-quartile range as well as the coefficient of variation over all the raters as a group and across each of the structures.

III.3.3 Evidence regarding hypothesis: Editing of automatic segmentations (A_1) maintains or improves accuracy

To assess accuracy, we compared rater segmentations to ground truth estimations via the Dice coefficient. Figure III.7 plots the distributions of pair-wise DSC of STAPLE and $Pmap_{mean}$ ground truth estimations against automatic and rater segmentation distributions. Each subplot can be divided into two: the four left columns (A/S,...,E/S) plot the segmentations against the STAPLE-derived ground truth, while the right side columns (A/P,...,E/P)

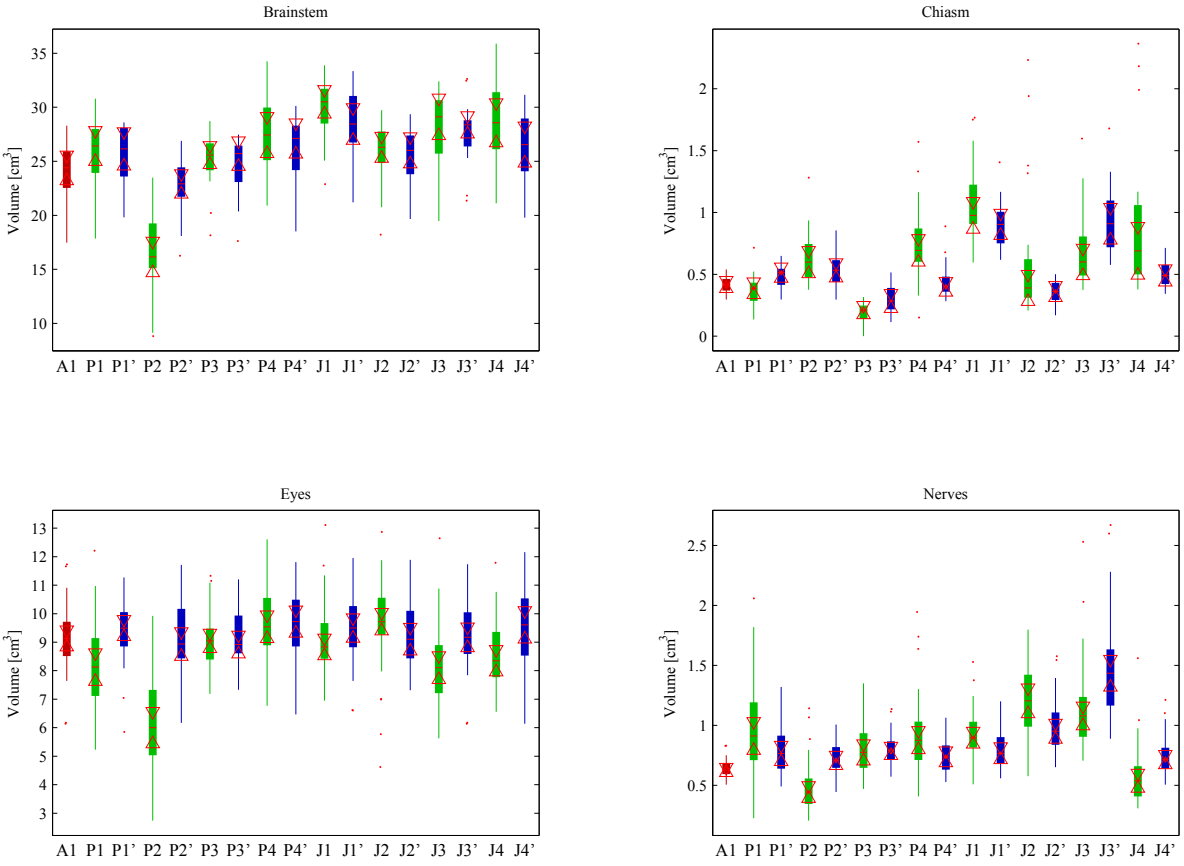

Figure III.6: Plots the distribution of nominal volume as a function of structure and rater and segmentation class: unedited automatic (A_1), *de novo* (columns P_1-J_4) and edited-A_1 (P'_1-J'_4).

Figure III.7: Plots the distributions of pair-wise DSC of STAPLE and Pmap_*mean* ground truth estimations against automatic and rater segmentation distributions. Each subplot can be divided into two: the four left columns (A/S,...,E/S) plot the segmentations against the STAPLE-derived ground truth, while the right side columns (A/P,...,E/P) plot the segmentations against the Pmap_{mean} ground truth. Columns A/,...,E/ represent the DSC distributions for the unedited automatic segmentations, unedited *de novo* segmentations, edited automatic segmentations, edited self, and edited peers, respectively.

Table III.2: Assessing accuracy of 5 classes of segmentations: unedited automatic (A_1), *de novo*, and editing groups M(A_1), M(self), and M(peers), via DSC against the ground truth estimates.

	Brainstem			Chiasm			Eyes			Nerves		
Source	Mean	Mean CI		Mean	Mean CI		Mean	Mean CI		Mean	Mean CI	
A_1	0.87	0.86	0.88	0.47	0.43	0.50	0.88	0.88	0.89	0.59	0.57	0.61
de novo	0.87	0.86	0.88	0.45	0.43	0.47	0.88	0.87	0.88	0.63	0.61	0.64
mod(A_1)	0.89	0.89	0.90	0.55	0.53	0.56	0.90	0.89	0.90	0.66	0.65	0.67
mod(self)	0.88	0.88	0.89	0.53	0.51	0.55	0.89	0.89	0.90	0.66	0.65	0.67
mod(peers)	0.90	0.90	0.91	0.59	0.57	0.61	0.90	0.90	0.91	0.70	0.69	0.71

plot the segmentations against the Pmap_{mean} ground truth. Columns A/,...,E/ represent the DSC distributions for the unedited automatic segmentations, unedited *de novo* segmentations, edited automatic segmentations, edited self, and edited peers, respectively.

Figure III.7 provides evidence that accuracy compared to unedited A_1 and *de novo* segmentations is at minimum maintained by editing, and this was consistent against both STAPLE and Pmap_{mean} ground truth estimates. However, the small tubular structures of the chiasm and nerves benefited the most from editing. The mean[95% CI] of Dice comparison against the ground truths for the chiasm increased from 0.47 [0.43,0.5] and 0.45 [0.43,0.47] for A_1 and *de novo* to 0.55 [0.53,0.56], 0.53 [0.51,0.55], and 0.59 [0.57,0.61] for edited-automatic, -self, and -peer, respectively. These mean DSC and 95% confidence intervals can be found in table III.2.

A complementary gauge of accuracy to DSC is the Euclidean, or surface normal, distance from the ground truth estimate to the test segmentation. Figure III.8 plots the signed distances, where positive indicates a point outside and negative a point inside the ground truth, Pmap_{mean}. The columns from left to right represent the distribution of distance error for the unedited A_1, *de novo* (columns $P_1,\ldots,J_4$) and edited A_1 for individual raters (columns $P'_1,\ldots,J'_4$). Blue dashed lines indicate a distance error of $\pm$ 2 mm. These distributions comprise hundreds of thousands of contour points, and there are a number of outliers. To improve clarity we have plotted them over fixed range from -5 mm to +10 mm denoted by the lower

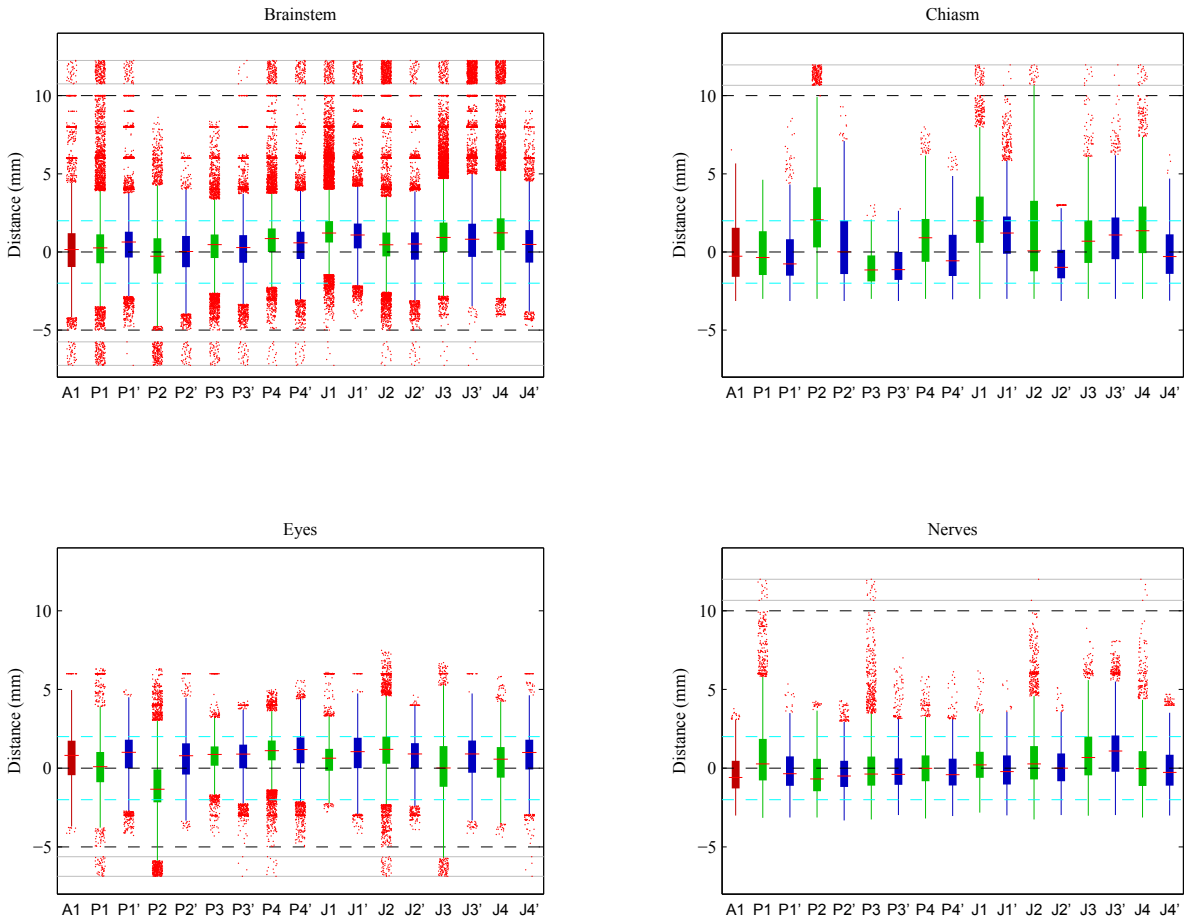

Figure III.8: Plots the signed distance errors, where positive indicates a point outside and negative a point inside the ground truth. The columns from left to right represent unedited A_1, *de novo* (columns P_1,..., J_4) and edited A_1 by raters (columns P'_1,..., J'_4). Inner dashed lines indicate a distance error of $\pm$ 2 mm. To improve clarity we plot are over fixed range from -5 mm to +10 mm denoted by the lower and upper bounded dashed lines. If outliers occur beyond these bounds, they are shown in lower and upper bands for which the density is proportional to the number of outliers.

and upper bounding dashed lines. If outliers occur beyond these bounds, they are shown in lower and upper bands for which the density is proportional to the number of outliers. The number of data points comprising each column of each subfigure (of figure III.8) varies as a function of the size of the structure drawn from one rater to the next. In so much as they are comprised of the same number of whole structure distance calculations (20 for each column), relative comparisons of outliers are valid. However, absolute comparisons of outlier density between structures are invalid, as the brainstem, for instance, has vastly more contour points than the chiasm, nerves or eyes. There were generally only small changes in median distance error between the *de novo* and edited-automatic segmentations, though the number of outliers was reduced in the edited distributions for most cases.

III.3.4 Evidence regarding hypothesis: Editing of automatic segmentations (A_1) salvages the results of low performing raters

In our previous study we noted that one rater in particular often produced segmentations different from the rest of the group. For this reason the rater in question was not included in the ground truth estimation. We hypothesized whether editing of the automatic segmentations would salvage the rater's performance. We use salvage to describe the process of preventing a negative result, such as "radiation was able to salvage the failed surgery". Looking to figures III.5 and III.6 we can see in situations where P_2 had a distribution markedly different from the group, and these deviations have been corrected by editing of A_1. The mean volumetric DSC for rater P_2 against the other experts increased from 0.69 to 0.91 (brainstem), 0.25 to 0.56 (chiasm), 0.75 to 0.93 (eyes), and 0.403 to 0.72 (nerves) through editing of the automatic segmentations.

We expect all raters on occasion to produce segmentations of low accuracy. These could be entire segmentations, such as mistaking the pituitary for the optic chiasm, or individual areas such as a single slice or series of slices omitted as part of the inferior brainstem. To this end we compared areas of low quality (slice DSC < 0.5) in the *de novo* study to the same areas post-editing of the automatic contours. First, we found the frequency of total miss, or omission of a slice, higher (16%) than the frequency of present but low quality contours (3.4%). The unedited A_1 produced fewer (12%) total misses but more low quality slices (8%) than the experts. As a result of editing of the automatic contours, the median DSC of the low quality slices increased from 0, which was skewed heavily by total misses, to a minimum of 0.5 for each of the raters. This reduced the total miss frequency by half, though the overall accuracy in these areas remained challenged. The mean DSC after editing for slices that were total misses (DSC $= 0$) in the *de novo* study improved to 0.45. Similarly for slices that contained contours

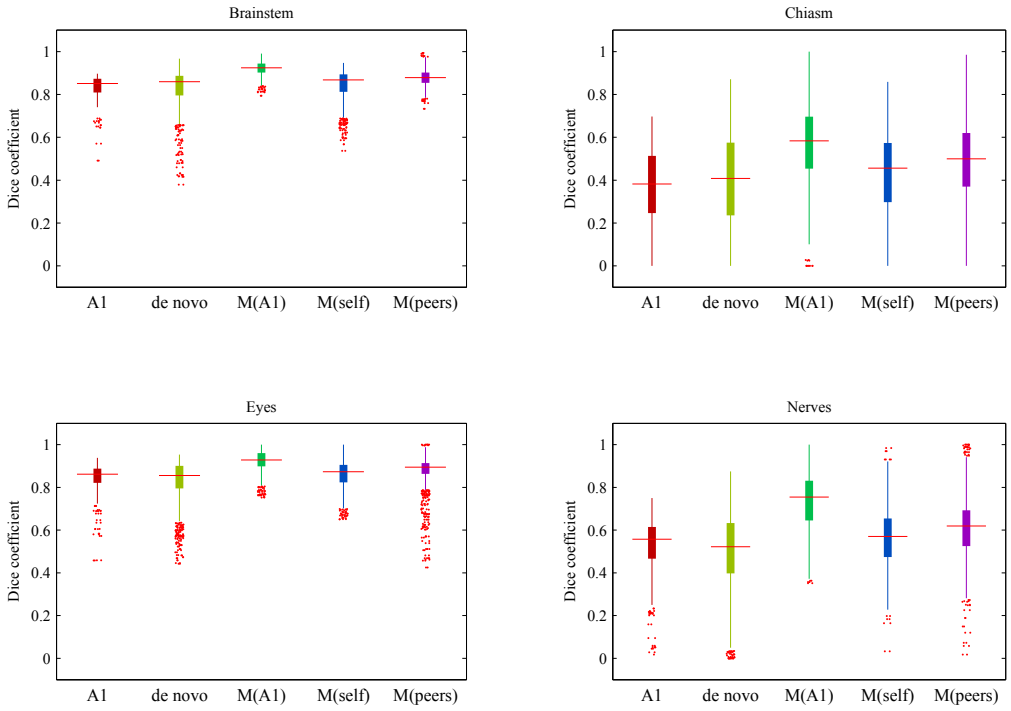

Figure III.9: Plots the distributions of volumetric DSC across each class of segmentation: unedited automatic (A_1), *de novo*, and the editing groups M(A_1), M(self), and M(peers).

but of low quality (DSC in range (0, 0.5)), the post-editing mean DSC increased from 0.34 to 0.52.

While it is clear that areas of poor performance *de novo* remained a challenge for raters during editing, the situation was improved as can be seen by the increase in both mean and median DSC and the avoidance of approximately half of total misses. Interestingly, P_2, the lowest performing rater *de novo*, saw the most dramatic improvement, from a median DSC of 0 to 0.68, the highest of the rater group, after editing of the automatic segmentations.

III.3.5 Evidence regarding hypothesis: Contour editing reduces inter-rater variation while maintaining or improving accuracy irrespective of the source segmentation

Thus far the results have focused on the performance of the automatic system in the context of editing compared to the experts' *de novo* segmentations and the unedited automatic segmentations. As outlined in section 2.1, we also asked the physicians in a blinded and randomized experiment to modify their own segmentations and those of their peers.

Their performance in these tasks is assessed in figures III.9-III.11 alongside the au-

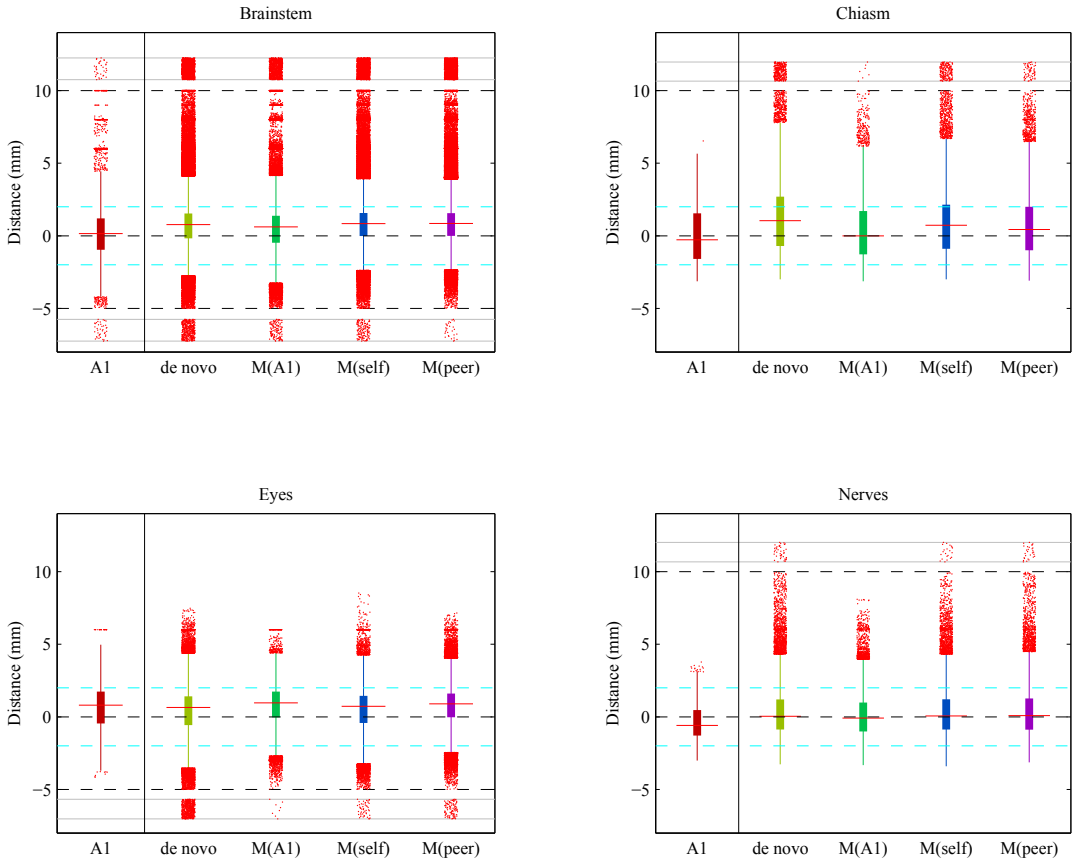

Figure III.10: Distance errors are plotted as a function of structure and segmentation class: unedited automatic (A_1), *de novo*, and edited A_1, self and peer. The inner dashed lines are drawn at $\pm$ 2 mm from the ground truth estimation. The plots are confined to a range -5 mm to + 10 mm as shown by the outer dashed lines. If a distribution has outliers beyond this range, they are plotted in the small bands at the periphery of the distributions. The density of the outliers within the bands is proportional to the number of outliers beyond the plot range.

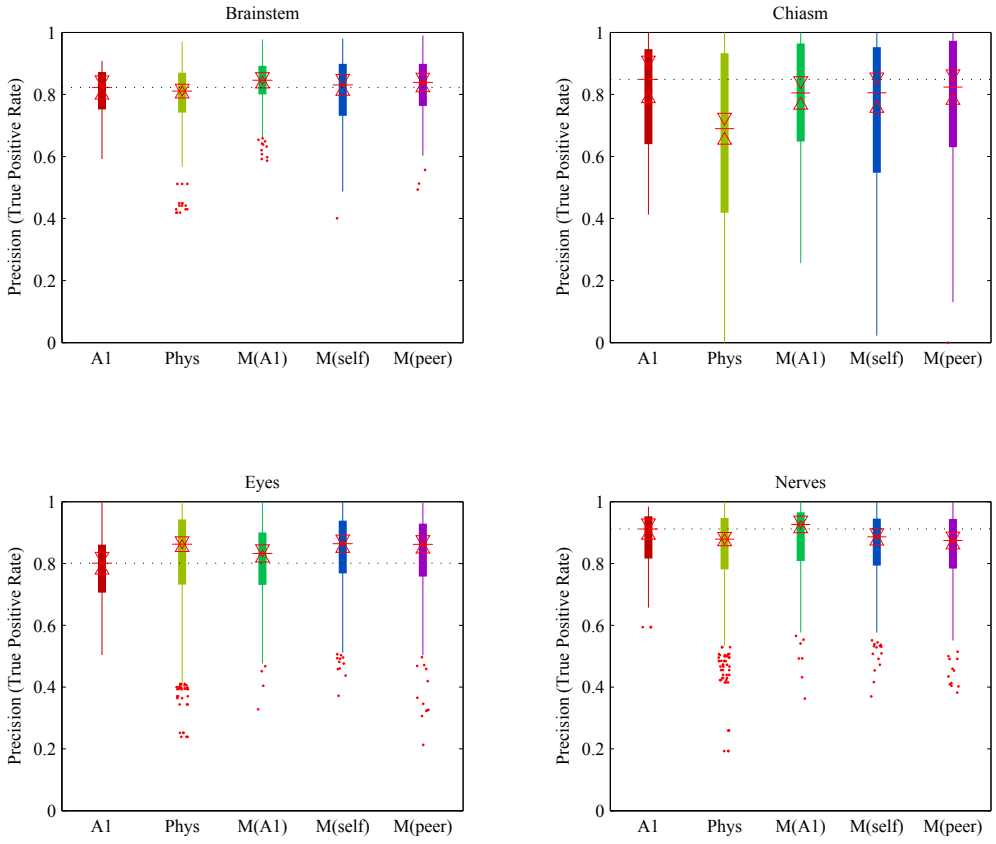

Figure III.11: True positive rate is plotted as the fraction of contour points falling within a 2 mm shell of the ground truth across the 5 segmentation classes. The dashed line is drawn at the level of the median for the unedited automatic (A_1).

tomatic and *de novo* results. The following nomenclature is used to distinguish the classes: automatic-unedited (A_1), expert-unedited (*de novo*), automatic-edited ($M(A_1)$), experts modifying their own initial segmentations (M(self)), and the expert modifying their peers' initial segmentations (M(peer)).

Figure III.9 plots the distributions of volumetric Dice coefficient across each class of segmentation, and tables III.3 and III.4 provide the mean, 95% confidence interval, and standard deviation for the same. Inter-rater variation was reduced for all editing groups as seen by both the increase in mean DSC and reduction in standard deviation. However, the best results came through editing of the automatic segmentations, which was consistent across the different structures. There was also a small but significant ($\alpha = 0.05$) advantage to modifying peers' as opposed to one's own segmentations.

Edits resulted in small differences in distance error, plotted in figure III.10, compared to the unedited automatic and *de novo* segmentations in terms of median error and with regard to the extent and number of outliers. [Note in figure III.10 when viewing the outliers shown by dots in red at the extremes of distributions, the A_1 (divided from the other groups by a vertical line) distributions are a result of only 20 segmentations each, whereas the other groups have approximately eight times the number of segmentations (one for each rater) in their distributions. Therefore, a direct comparison of outlier prevalence between A_1 and the others is not possible visually.] In fact, in the cases of the optic chiasm and brainstem the unedited automatic segmentations produced a median distance error closer to zero than either the *de novo* physicians or the any of the edited segmentations. These boxplots, however, consider the complete set of all contour points for a given rater or rater-group, which skews the results towards patients with larger structures and raters who contoured larger structures. Weighting each rater and case equally, we recalculated the mean (and 95% confidence interval), minimum and maximum distance errors provided in table III.6. Across all classes of segmentations mean distance errors were approximately equal to or less than 1 mm. Interestingly, the unedited automatic performed well in comparison to the edited classes. This was especially true in terms of maximum (signed positive) distance errors, which were smaller for all structures except the eyes.

We also used the signed distance maps to calculate the true positive rate within a 2 mm shell around the ground truth estimation. We found significant differences ($\alpha = 0.05$) from the unedited automatic and the *de novo* segmentations (figure III.11) only in the case of the optic chiasm, where both the A_1-unedited and all editing classes (A_1-, self-, and peer-edited) had higher true positive rate compared to the unedited *de novo* class. This advantage disappeared when we narrowed the shell to 1 mm (not shown), such that there were no differences ($\alpha = 0.05$)

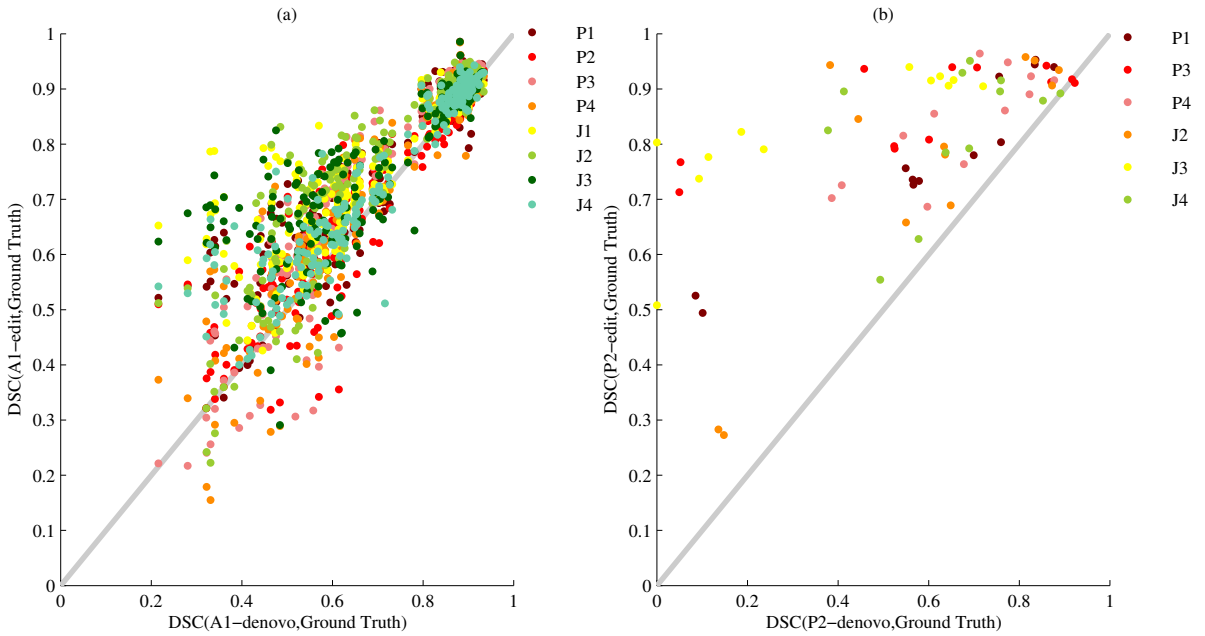

Figure III.12: Plots of DSC against the ground truth segmentations pre- and post-editing for A_1 (a) and P_2 (b). Note P_2 was not randomized to all peers in editing, thus the differences between the raters in legends of (a) and (b).

amongst the five groups of unedited and edited segmentations.

Another question that arises which we can begin to answer is that of whether editing is robust to segmentations of varying quality. We can do so by examining the correlation between pre- and post-editing accuracy. Figure III.12 plots DSC against the ground truth estimates pre- and post-editing for (a) A_1 and (b) P_2. We chose to single out P_2 to illustrate this effect as this rater generally produced the segmentations most different from the group in the *de novo* study. A line with a slope of 1 is plotted through the origin; all points above the line indicate an improvement in accuracy. We see in general that editing improves the accuracy, which is supported by figure III.7 as well, and editing appears largely robust in areas of low quality initially (low DSC *de novo*). In figure III.12 (b) we see that each time a peer edited the contours of P_2 the accuracy was improved, usually substantially, though they did not generally attain a final accuracy as high as was achieved starting with higher accuracy segmentations.

III.4 Discussion

We have undertaken a large scale behavioural study to better understand performance of automatic and manual segmentation in radiation therapy and the interaction of the two. Previously (Deeley et al., 2011) we reported on automatic segmentation for brain organs at risk in the presence of large space-occupying lesions, which challenge registration-based methods.

That study characterized inter-rater variance and found that the automatic system generally could serve as a surrogate to the physicians with potential gains in efficiency and accuracy within the treatment planning process. The basis for our experimental framework is the observation that segmentation should be evaluated in behavioural studies through 1) multi-dimensional metric analysis, e.g., volumetric and distance-based methods, 2) sufficient numbers of raters and patients chosen prospectively to ensure high power analyses, and 3) clinically realistic design that recognizes end-use of the segmentations. Here we applied this framework using the same physicians and patients in a single-blind editing study to test hypotheses concerning the impact of manual-automatic system interaction (editing of the automatic) on inter-rater variance and accuracy, the impact of manual-manual rater interaction (modifying their own and peers' segmentations), and whether the automatic system could salvage the performance of low performing raters.

III.4.1 Comparison to previous studies

Previously Chao (Chao et al., 2007) reported results of an editing study using computer-assisted delineation of head and neck structures. In this study, eight physicians manually contoured two head and neck cases and then edited contours produced from an atlas-based system. They found editing reduced inter-rater variance with significance via nominal volume, Dice coefficient, and Euclidean distance disagreement. The authors proffered that computer-assisted segmentation and contour editing may be useful to educate physicians from different training backgrounds and to improve efficiency in the treatment planning process. We found this work compelling in the design of our study, but it was limited in several ways. While our experience would indicate the number of raters generally sufficient, the overall analysis is likely of low power as the statistical analyses were performed on each of the two patients separately. It is unlikely the results reported can be used to infer to a larger population. Additionally, there is a fundamental difference from the study we have undertaken. In the Chao study, the participants contoured from scratch and then immediately were presented with automatic contours for editing. Furthermore, the raters viewed the atlas images at all times during editing. This is certainly a valid design though has a markedly different emphasis than our own. One could envision a single standard atlas for use by all radiation oncologists for every contouring task, and one could extrapolate that this method may reduce variation within the population. Our focus was in a different direction. At the outset of the study we discussed with the group of experts general guidelines for delineation. We also chose a body site, the brain, where training is more ubiquitous and expected variance lower. We suspect viewing of the atlas during editing would additionally bias the raters in a clinically unrealistic way, as this is not standard practice.

Stapleford and colleagues (Stapleford et al., 2010) also reported results of a segmentation and editing study involving the head and neck. They recruited five physicians to contour bilateral lymph node regions for five patients and to edit automatically generated contours. The data were analysed using five metrics: sensitivity, DSC, percent false positive, mean and max surface distance error, and volume. These metrics yield complementary information about the differences in segmentations. The use of percent false positive in place of specificity is particularly useful, as specificity is dependent on the image size and thus not readily comparable between studies. They found the automatic contours compared well to the manual, and editing led to improved consistency. Interestingly the experts commented that only 32 percent of contours were acceptable without editing, and the primary complaint was the automatic segmentations were too large. However, when making edits, the cumulative changes only partially recaptured the mean volume of the manual segmentations, leading one to wonder about the bias introduced by the automatic segmentations. We found the opposite in our study of brain structures. The automatic system consistently produced smaller segmentations than the experts, but expert nominal volume was generally recaptured upon editing. In fact, we found in general that editing produced a trend of increasing segmentation size, though the effect was small (table III.2 and figure III.6).

There are some limitations to the methodology of the investigation by Stapleford and colleagues. The authors used STAPLE to calculate two ground truth segmentations for all pairwise comparisons: one from the cohort of manual segmentations and one from the cohort of edited segmentations. The use of simulated ground truths to assess accuracy is desirable, but their methodology rests the validity of almost all judgment on the quality of these estimations. First, we find it non-ideal to create separate ground truth estimates to compare groups of segmentations from the same imaging dataset. How can one infer with high confidence differences between segmentations when the two groups are being compared to different ground truths that were calculated from their respective groups? This requires a seemingly contradictory assumption: both ground truth estimations, while different from each other, are fully accurate ground truths, or at the very least have equal quality. If a systematic difference exists between the groups, it may be missed. We believe a more appropriate assumption, though not ideal, is to choose a single ground truth calculated from the most appropriate cohort for all comparisons. Second, basing variance analysis through intermediary comparison with simulated ground truths will produce a perception of lower overall variance and increased correlation amongst the group. We posit that the most accurate and transparent way to evaluate inter-rater variance is through standalone metrics such as nominal volume or pairwise metrics on the unadulterated segmentations.

We found that editing of pre-generated segmentations both improved efficiency (reduced contouring time by at least 60%) and reduced inter-rater variance across all sources (A_1, self, peers), structures, patients, and physicians. Though we found, interestingly, that physicians showed preference toward their own contours in terms of time and amount of editing, variance was reduced more when they edited the automatic segmentations, regardless of structure. However, as Zou (Zou et al., 2004) suggests the problem can be restated as one in which error is a function of bias and variance, or another way of stating it, as random and systematic errors. Thus, one must be careful not to overstate the implications of observed variance. In our study, each rater edited the same automatically generated structures such that the inter-rater variance before editing was zero. When modifying their own or peers' contours, the baseline variance was carried from the *de novo* study.

To determine whether pre-generation of contours impacted the raters' accuracy, we employed two ground truth estimates. The STAPLE algorithm and our own approach with p-maps as well as the rationale for using both estimates has been discussed previously (Warfield et al., 2004; Deeley et al., 2011; Meyer et al., 2006; Biancardi et al., 2010). In general, ground truth estimation is a difficult problem that is at least in part a function of size and quality of the input cohort. As discussed in reference to the work by Stapleford and colleagues, choice of ground truth cohort can be vital to the conclusions drawn from the analyses. We had several distinct classes of segmentations (A_1, *de novo*, A_1-edited, self-edited, and peer-edited), each with multiple cases, raters, and structures from which to choose a cohort for ground truth estimation and subsequent accuracy analyses. The following considerations were made. First, since all expert segmentations are valid clinically by virtue of the raters' expertise and none of the automatic segmentations would be deemed acceptable without oversight, we did not include A_1 as an input to the ground truth calculations. Second, it is also not valid to use A_1-edited for reasons already mentioned: there is no basis to know whether it will bias the raters toward higher or lower accuracy. Third, heuristically, we reasoned that including either all physician segmentations (*de novo* and edited) or just those edited would be non-ideal, as there could be significant inter-class differences (increased variance) which would presumably lead to lower quality estimates. With this in mind we chose to make all assessments against those calculated from a single class, the peer-edited class. Prospectively we anticipated that this class of edits would be the most likely to have reduced variance and similar or less bias compared to the *de novo* class (used for ground truth creation in the previous study) and the self-edited class, and this was born out in the data as can be seen in tables III.3 and III.4.

Testing against the ground truth estimates from the peer-edited class, we found that accuracy was either maintained or improved in figure III.7 via editing. Accuracy of edited

classes was similar to the unedited automatic and *de novo* classes, as seen by the 95% confidence intervals of mean DSC in tables III.3 and III.4 for the brainstem and eyes, but editing improved accuracy in the more challenging optic chiasm and nerves. The distance data paint a less clear picture. Editing regardless of source reduced the number of outliers, but in terms of mean, min, and max distance error there were only small differences from the A_1 or *de novo* classes. In fact, in the analysis of true positive rate within a 1 mm and 2 mm shell (figure III.11) of the ground truth, only for the chiasm were results notable in that the unedited automatic as well as all three editing classes had smaller distance errors than the physicians *de novo*.

The *de novo* study previously uncovered that in the group of eight experts one was often an outlier and therefore removed from the ground truth cohort. We also found the source of these differences, especially in the brainstem, was often failure to extend the organ as far cranial-caudal as the group. It would be very useful clinically for automatic systems to correct these errors, which we term total miss errors. Looking at every slice from the *de novo* study against the ground truth estimates we isolated low quality contours, anything with a DSC < 0.5. The prevalence of total miss over present-but-low-quality slices suggests that edges of structures in the cranial-caudal plane are a challenge for manual raters. This is likely both a result of lack of natural boundary (e.g., brainstem and spinal cord) and partial volume effects (e.g., chiasm, nerves and eyes). Editing of A_1 was generally successful at salvaging the total misses. The improvement for present-but-low-quality slices was less remarkable and is likely a result of the automatic system and the manual raters being generally challenged in areas of low contrast.

This study provides strong evidence that editing of pre-generated segmentations, independent of source, reduces inter-rater variance while maintaining or improving accuracy and increasing efficiency. This suggests given a starting point, even if the starting points are different, experts tend to converge. We postulate that raters focus on the task of segmentation differently when modifying than when starting from a blank slate. The data showed, for instance, though differences at the edges of contours (distance error) were not dramatically different from the *de novo* study, raters focused more on capturing the entire extent and correct location of a structure, suggesting a good starting place to develop delineation standards may be to propose contours for editing to experts in the field. These results also lend evidence to the suggestions made in prior work (Chao et al., 2007; Beyer et al., 2006) that automatic methods can help improve consistency in radiation therapy treatment planning, especially situations wherein users are less experienced. Finally, the two studies we have undertaken provide evidence that our unedited automatic segmentations perform quite well, and after editing provide an even more robust alternative to manual segmentation.

III.4.2 Limitations and future work

There were several limitations in this work. First, we have attempted to extend an experimental framework for segmentation analysis beyond what has been done previously using a behavioural approach and statistically robust design. However, our study though large by comparison to others is limited to a single institution that may have systematic bias. Additionally, as this is an ongoing project extending the prior *de novo* study, we have not evaluated the results of other algorithms or now commercially available systems. We have focused on only one body site. Most of these choices were a function of resources, since behavioural studies require prolonged time for longitudinal tasks (over 2 years to collect data in our case) and are costly. Many more questions could be investigated with less global uncertainty if a framework such as that we have proposed could be implemented on a multiple institution, body site, and algorithm basis. This would also help to gain useful interaction about the users and the system for contouring, such as which tools were utilized for contours or editing and whether those choices impacted results.

Second, the choice of metrics is important. We believe multiple complementary and cross-study compatible metrics such as the Dice coefficient, distance-based measures, and nominal volume increase the value of the analyses. However, the metrics as used herein can only characterize the data and describe differences in and relationships between groups or classes of segmentations. A valuable analysis would involve an understanding of what are the sources of these differences, such as has been done in prior work by Meyer (Meyer et al., 2006) and Zou (Zou et al., 2004) using analysis of variance and multiple regression. We did not include that analysis herein as the scope was already extensive. However, given sufficient categorical understanding of the data, this could be done retrospectively. This type of analysis in a targeted study with multiple different sources of varying quality would also help to further answer questions about the interaction of source segmentation quality and the editing process.

Third, the end point of the segmentations in our context is radiation therapy treatment plans, which was not considered herein. The ultimate impact of differences will manifest in dose coverage of target volumes and normal tissues. Others have looked at dosimetric end points (Weiss et al., 2008; Tsuji et al., 2010) but generally not in the context of a large scale study with multiple raters. Nelms and colleagues (Nelms et al., 2012) conducted a "Plan Challenge" evaluating the dosimetric impact of differences in normal tissue contouring in the head and neck. However, only a single patient was analysed over 32 raters. Extending studies such as these with more raters, patients, and anatomical sites would provide valuable information about the impact of segmentation variance as well as help guide clinical users.

Lastly, the lack of a known ground truth is an ongoing challenge in segmentation

evaluation. We have discussed the choice and importance as well as the pitfalls of ground truth estimation. The wealth of data generated in the editing study presented a problem of choosing a cohort of segmentations as inputs to the ground truth calculations. We reasoned that the peer-editing group was the most desirable class to use for truth estimation, and it was applied in all analyses for all groups. In post-hoc analysis we also looked at the impact had different assumptions been made, namely using the other classes to compose the truth estimation. We found that these assumptions did produce small differences, most notably when A_1-edited was used. The choice of A_1-edited in ground truth composition resulted in higher accuracy for the A_1-edited class as compared to other classes, though the magnitude of accuracy in the other groups did not change remarkably. This is likely a result of the reduced variance of the A_1-edited class compared to the other edited classes. It is also possible that the results we have presented favor accuracy toward the peer-edited class at the expense of the other classes, including the automatic and automatic-edited. However, it was determined this was a better choice than to potentially bias accuracy toward the automatic system.

Acknowledgements

This work was funded in part through a grant from the National Institute of Biomedical Imaging and Bioengineering (award R01EB006193). All images used in this study were acquired with institutional review board approval through the Vanderbilt-Ingram Department of Radiation Oncology. The manual segmentations in this study were acquired via a research Eclipse treatment planning system through a grant from Varian Medical Systems (Palo Alto, CA). The authors would also like to thank George Ding for his thoughtful advice in implementation of this project, as well Jenny Lu for assistance in data entry and extraction from the treatment planning system.

Appendix

Table III.3: DSC for brainstem and chiasm, each rater modifying A_1, self, and peers.

Rater	Source	Brainstem				Chiasm			
		Mean	Mean CI		std	Mean	Mean CI		std
P_1	A_1	0.927	0.922	0.931	0.029	0.612	0.586	0.644	0.172
	self	0.863	0.851	0.872	0.066	0.468	0.437	0.498	0.190
	peers	0.873	0.863	0.881	0.047	0.488	0.457	0.519	0.169
P_2	A_1	0.907	0.900	0.913	0.038	0.562	0.530	0.592	0.191
	self	0.726	0.716	0.737	0.064	0.251	0.220	0.285	0.194
	peers	0.867	0.859	0.874	0.038	0.402	0.364	0.439	0.198
P_3	A_1	0.926	0.922	0.931	0.027	0.562	0.523	0.597	0.218
	self	0.864	0.855	0.874	0.059	0.396	0.364	0.427	0.186
	peers	0.884	0.877	0.891	0.038	0.491	0.452	0.528	0.191
P_4	A_1	0.926	0.921	0.931	0.029	0.550	0.514	0.585	0.207
	self	0.867	0.856	0.878	0.066	0.443	0.415	0.474	0.170
	peers	0.886	0.879	0.892	0.036	0.506	0.470	0.537	0.180
J_1	A_1	0.916	0.910	0.920	0.032	0.539	0.513	0.560	0.144
	self	0.855	0.842	0.866	0.075	0.444	0.417	0.472	0.172
	peers	0.877	0.871	0.884	0.037	0.508	0.479	0.537	0.154
J_2	A_1	0.927	0.922	0.931	0.030	0.572	0.537	0.606	0.209
	self	0.861	0.851	0.870	0.060	0.486	0.453	0.518	0.197
	peers	0.882	0.875	0.889	0.036	0.484	0.450	0.521	0.182
J_3	A_1	0.910	0.904	0.915	0.030	0.524	0.497	0.548	0.153
	self	0.842	0.830	0.852	0.066	0.471	0.443	0.502	0.176
	peers	0.870	0.862	0.875	0.033	0.486	0.452	0.517	0.166
J_4	A_1	0.924	0.919	0.929	0.032	0.609	0.582	0.636	0.166
	self	0.841	0.831	0.850	0.061	0.490	0.459	0.518	0.177
	peers	0.872	0.863	0.880	0.044	0.518	0.483	0.550	0.168
Senior	A_1	0.922	0.919	0.924	0.032	0.572	0.554	0.590	0.200
	self	0.830	0.822	0.837	0.088	0.389	0.371	0.406	0.204
	peers	0.877	0.874	0.881	0.041	0.472	0.454	0.490	0.190
Junior	A_1	0.919	0.916	0.921	0.032	0.561	0.545	0.575	0.174
	self	0.850	0.844	0.856	0.066	0.473	0.459	0.488	0.182
	peers	0.875	0.872	0.879	0.038	0.499	0.483	0.515	0.169
All Phys	A_1	0.920	0.918	0.922	0.032	0.566	0.555	0.577	0.187
	self	0.840	0.835	0.845	0.079	0.431	0.420	0.443	0.198
	peers	0.876	0.874	0.879	0.039	0.486	0.473	0.497	0.180

Table III.4: DSC for eyes and optic Nerves.

Rater	Source	Eyes				Nerves			
		Mean	Mean CI		std	Mean	Mean CI		std
P_1	A_1	0.904	0.898	0.911	0.052	0.746	0.731	0.759	0.122
	self	0.860	0.853	0.867	0.058	0.539	0.524	0.555	0.134
	peers	0.881	0.873	0.887	0.051	0.603	0.585	0.621	0.128
P_2	A_1	0.932	0.926	0.937	0.045	0.718	0.703	0.736	0.142
	self	0.810	0.805	0.816	0.046	0.469	0.456	0.483	0.123
	peers	0.880	0.869	0.887	0.065	0.599	0.582	0.617	0.130
P_3	A_1	0.930	0.926	0.935	0.039	0.768	0.752	0.781	0.121
	self	0.877	0.869	0.883	0.058	0.595	0.580	0.608	0.121
	peers	0.887	0.876	0.894	0.065	0.609	0.583	0.632	0.167
P_4	A_1	0.915	0.910	0.919	0.039	0.760	0.746	0.774	0.121
	self	0.873	0.866	0.879	0.056	0.610	0.595	0.624	0.121
	peers	0.878	0.867	0.887	0.073	0.618	0.601	0.635	0.124
J_1	A_1	0.932	0.927	0.937	0.042	0.775	0.762	0.788	0.114
	self	0.880	0.874	0.887	0.059	0.607	0.593	0.621	0.121
	peers	0.878	0.869	0.885	0.058	0.611	0.590	0.632	0.158
J_2	A_1	0.918	0.912	0.923	0.042	0.733	0.721	0.746	0.107
	self	0.874	0.866	0.880	0.057	0.607	0.592	0.618	0.113
	peers	0.888	0.877	0.896	0.064	0.648	0.631	0.663	0.121
J_3	A_1	0.936	0.931	0.942	0.048	0.600	0.589	0.613	0.105
	self	0.832	0.824	0.839	0.063	0.525	0.511	0.538	0.117
	peers	0.823	0.805	0.839	0.121	0.574	0.559	0.591	0.117
J_4	A_1	0.928	0.923	0.934	0.050	0.761	0.746	0.776	0.131
	self	0.868	0.862	0.874	0.052	0.515	0.499	0.529	0.122
	peers	0.879	0.869	0.887	0.069	0.612	0.597	0.630	0.132
Senior	A_1	0.920	0.918	0.923	0.046	0.748	0.740	0.755	0.128
	self	0.855	0.851	0.858	0.061	0.553	0.545	0.561	0.137
	peers	0.881	0.877	0.885	0.064	0.608	0.597	0.616	0.138
Junior	A_1	0.929	0.926	0.931	0.046	0.717	0.710	0.726	0.134
	self	0.864	0.860	0.867	0.061	0.563	0.556	0.571	0.126
	peers	0.867	0.862	0.873	0.086	0.611	0.603	0.621	0.136
All Phys	A_1	0.925	0.923	0.926	0.046	0.733	0.727	0.738	0.132
	self	0.859	0.857	0.862	0.061	0.558	0.553	0.563	0.132
	peers	0.874	0.870	0.877	0.076	0.610	0.603	0.615	0.137

Table III.5: Volume [cm^3]. Mean, 95% confidence interval on the mean, and the coefficient of variation of nominal volume for the unedited automatic (A_1), *de novo*, and editing groups M(A_1), M(self), and M(peers).

	Brainstem				Chiasm			
Source	Mean	Mean CI		cov	Mean	Mean CI		cov
A_1	23.99	22.82	24.87	11.01	0.41	0.39	0.45	16.07
de novo	25.88	25.01	26.62	19.59	0.66	0.60	0.74	67.41
mod(A_1)	25.84	25.33	26.31	12.55	0.56	0.52	0.61	48.30
mod(self)	26.76	25.98	27.42	18.51	0.67	0.62	0.73	59.04
mod(peers)	27.16	26.55	27.83	13.37	0.67	0.60	0.73	57.31

	Eyes				Optic Nerves			
Source	Mean	Mean CI		cov	Mean	Mean CI		cov
A_1	9.13	8.59	9.57	17.56	0.64	0.61	0.67	54.98
de novo	8.59	8.40	8.77	20.15	0.87	0.83	0.91	41.75
mod(A_1)	9.39	9.25	9.52	12.78	0.89	0.85	0.92	35.16
mod(self)	8.88	8.71	9.03	17.13	0.95	0.91	0.98	38.22
mod(peers)	9.27	9.10	9.44	16.06	1.01	0.97	1.04	33.90

Table III.6: Distance error [mm]. Presents the mean, confidence interval, minimum and maximum signed distance errors for the 5 classes of segmentations: unedited automatic (A_1), *de novo*, and edited A_1, self and peer. These distances were determined weighting each rater and patient equally, e.g., the maximum can be thought of as the maximum distance error averaged over the 20 patients.

	Brainstem					Chiasm				
	Mean	Mean CI		Min	Max	Mean	Mean CI		Min	Max
A_1	0.18	0.02	0.35	-3.42	4.83	-0.08	-0.25	0.17	-2.02	2.00
de novo	0.72	0.60	0.82	-3.90	7.23	1.08	0.78	1.62	-1.90	5.07
mod(A_1)	0.57	0.48	0.65	-3.61	6.38	0.04	-0.12	0.21	-2.28	3.58
mod(self)	0.85	0.73	0.96	-3.21	7.40	0.51	0.28	0.79	-2.16	4.54
mod(peers)	0.88	0.80	0.97	-3.01	7.37	0.31	0.09	0.54	-2.14	4.00

	Eyes					Optic Nerves				
	Mean	Mean CI		Min	Max	Mean	Mean CI		Min	Max
A_1	0.63	0.51	0.75	-2.59	3.75	-0.39	-0.50	-0.27	-2.59	2.38
de novo	0.32	0.17	0.46	-2.78	3.15	0.31	0.16	0.45	-2.89	3.21
mod(A_1)	0.74	0.66	0.81	-2.31	3.54	0.79	0.73	0.85	-2.33	3.49
mod(self)	0.44	0.30	0.54	-2.36	3.19	0.42	0.28	0.54	-2.31	3.16
mod(peers)	0.68	0.56	0.77	-2.19	3.31	0.71	0.59	0.80	-2.21	3.43

CHAPTER IV

DOSIMETRIC IMPACT OF AUTOMATIC SEGMENTATION

IV.1 Introduction

Image segmentation is a vital component of modern radiotherapy treatment planning and will become only more so with the increased use of inverse and adaptive planning methods (Hansen et al., 2006; Ding et al., 2006; Schwartz and Dong, 2011; Gregoire et al., 2012; Jensen et al., 2012; Peroni et al., 2012; Schwartz et al., 2013). Traditionally, segmentation has been accomplished through manual human intervention, but in recent years algorithms have been developed to segment the structures needed for treatment planning in several body sites such as the pelvis, head and neck, and brain.

These algorithms have been incorporated clinically and in commercial products with relatively few published reports focused on evaluation in a clinically realistic context (Chao et al., 2007; Stapleford et al., 2010; Deeley et al., 2011, 2013). We have undertaken a multi-rater behavioral study to gauge the impact of automatic segmentation as well as differences amongst experts for intracranial organs at risk in the presence of large space-occupying lesions. Our work is motivated by the observation that segmentation is an inherently noisy process (Meyer et al., 2006) for which an individual rater may not serve as a robust reference standard. Previously we reported results regarding the geometric quality of automatic segmentations in the context of accuracy and variability of the experts for the brainstem, optic chiasm, eyes, and optic nerves. We used three comparison metrics, nominal volume, Dice similarity coefficient (DSC), and Euclidean distance, to test the hypothesis that there was no geometric difference between the automatic and expert segmentations. We found that differences in raters could be large, that at least one rater was often markedly different from the group, and that the automatic system performed well in this context, though both the automatic and manual raters were challenged considerably in the area of the small tubular structures: the optic chiasm and nerves. We used two simulated ground truth methods to assess accuracy: the simulataneous truth and performance level estimation (STAPLE) algorithm and a novel implementation of probability map thresholding (Meyer et al., 2006; Deeley et al., 2011), similar to voting rule (Kittler et al., 1998).

In a second study we tested hypotheses concerning the impact of segmentation editing by experts, as this is likely how systems for segmentation will be used. We presented seg-

mentations for editing to the same group of eight raters in a single-blind, randomized design. The sources of the segmentations were 1) the automatic segmentations (A_1), 2) their own segmentations (self), and 3) their peers' segmentations (peer) from the first (*de novo*) study. We found that editing improved efficiency while reducing variation and maintaining or improving accuracy, regardless of original segmentation source. That is, in a geometric sense, editing was efficacious regardless of whether the experts edited A_1, self, or peers. We also found that editing generally improved accuracy in the areas where raters had performed poorly compared to the group in the *de novo* study. Even when experts were presented the lowest quality results from that study, editing generally salvaged the final segmentation result.

The ultimate test of segmentation acceptability is that of impact on the end-use. In radiation therapy that corresponds to dosimetry. While large multi-rater studies for which we advocate have been rare, recent work has begun to incorporate some of these aspects into studies of impact on geometry, dosimetry, and the interaction of the two. This work has primarily been focused in the area of head-and-neck cancer (Nelms et al., 2012) and in the context of adaptive therapy (Tsuji et al., 2010; Voet et al., 2011), an area that would arguably benefit the most from accurate and robust automatic segmentation. The design of clinically evaluative studies of segmentation, whether dosimetric or not, should reflect the clinical variance that exists. These can be separated into the variance resulting from differences in normal and pathologic patient anatomy, imaging protocols, treament planning systems and planners, and physicians. The resources needed for such exhaustive studies are enormous and scarce, which has limited both the scope and likely the statistical power of prior studies. Nelms and colleagues (Nelms et al., 2012) examined the dosimetric impact of manual contouring differences in a multi-institutional study of 32 raters, but the scope was limited to a single patient and all comparisons were made against an assumed ground truth from one institution. In another study by Tsuji and colleagues (Tsuji et al., 2010), again only one rater was used as ground truth. Teguh (Teguh et al., 2011), again in the head-and-neck, utilized a number of different raters over 12 patients, some raters segmenting *de novo* and others editing automatic contours, such that impact of inter-rater variance was unclear. In what appears to be a follow-up study to that of Teguh and colleagues, Voet (Voet et al., 2011) examined the dosimetric impact over 9 patients with 2 editing raters. Yet another study (La Macchia et al., 2012) evaluated segmentation results from three commercial systems over three body sites (head-and-neck, pleura, and pelvis) using 5 patients and 3 raters, though rater variance was not considered. We posit that while each of these studies has added important information, study design has been such that strong inferences cannot be made.

In this work, we retained the behavioral framework of our previous two studies to

gauge the impact on inverse-planned intensity modulated radiation therapy (IMRT). We have limited the study to the *de novo* segmentations as a worst-case scenario in terms of dosimetric variability, as we have shown previously that editing reduces variation and improves or maintains accuracy.

There were three main considerations in this study concerning the dosimetric impact of segmentation differences. First, we tested whether target coverage was impacted by segmentation differences. In our prior studies we find considerable inter-rater variability and overall reduced accuracy for the small tubular structures of the optic chiasm and nerves. As this study was comprised of large tumors often in close proximity to the normal tissues, it is plausible that coverage could be compromised. Second, we tested the impact of using different segmentations on dose to the ground truth simulated organs at risk. This is a departure from other studies that have compared planned dose using a single manual rater as a reference standard (Nelms et al., 2012; Voet et al., 2011). In so doing, we tested whether the automatic system or other rater-derived plans deviated from the group in such a way to negatively impact normal tissue toxicity. Third, we evaluated the multi-rater plans in terms of dose reported versus dose to ground truth.

The rationale for measuring dose to the targets is simple. The goal is to treat the targets while sparing the normal tissues. Since the target segmentations and all other variables were held constant between raters, differences in target coverage can be attributed to differences in normal tissue segmentations. The rationale for our approach to evaluating impact via the normal tissues is more complex. First, measuring and comparing doses to ground truth estimates from the various rater-derived plans is a way of testing the impact of those raters' segmentations on our best estimate of reality; that is, what is the true dose to a normal tissue by a plan. Second, in a clinical situation the ground truth will not be known, and thus the dose that is reported will be that as measured by the segmentation used in the optimization, the raters' own segmentation. We call the difference in the two the dose reporting discrepancy. The first measurement is important in understanding potential toxicity. The second is important for clinical decision making (e.g., a physician deciding whether to undertreat a tumor as a result of the dose reported to a normal tissue) and in the broader perspective of evidence-based studies of toxicity.

Lastly, while confined to critical organs in the brain, it is our hope that this work yields useful information for other investigators and a framework for the design of future evaluative studies, especially regarding head-and-neck lymph node region and organ at risk segmentation.

IV.2 Materials and Methods

This study follows our previous *de novo* study (Deeley et al., 2011, 2013), in which automatic segmentations (A_1) were evaluated in the context of eight expert raters. Many additional details can be found in that work. In short, eight raters (four senior, P1-P_4, and four junior, J_1-J_4) contoured the brainstem, optic chiasm, eyes, and optic nerves from scratch over 20 patients who had been treated previously with IMRT for large space-occupying lesions in the brain, mostly glioblastoma multiforme. This site was chosen as a benchmark site for the ubiquity of physician training in cross-sectional anatomy as well as for the challenge presented by the large lesions to registration-based segmentation algorithms. The number of study patients was originally chosen to provide a power of 0.9 ($\beta = 0.1$) to detect a difference of 0.1 in Dice coefficient while setting long term type I errors to 5% ($\alpha = 0.05$). Dosimetric estimates were not available to guide study design.

IV.2.1 Segmentation

Manual and automatic contouring was accomplished utilizing fused sets of x-ray computed tomography (CT) (2 or 3 mm slice thickness) and magnetic resonance (T1-weighted, 1.5 or 3 T, approximately 1 mm^3 voxels) images. Our atlas-based registration-driven segmentation methods for the brainstem and eyes have been discussed at length in previous work (Rohde et al., 2003; D'Haese et al., 2003; Deeley et al., 2011) as well as the atlas-navigated optimal medial axis and deformable model algorithm (NOMAD) (Noble and Dawant, 2011) we use for segmenting the optic chiasm and nerves.

IV.2.2 Treatment planning

The unedited automatic and manual contours from the *de novo* study were used to generate inversely optimized IMRT plans (Philips Pinnacle v.9.0) in the current study. Planning was dictated by the radiation therapy oncology group (RTOG) 0837 clinical protocol, which we believe is well centered within the standard of care. Gross, clinical and planning target volumes (GTV, CTV, and PTV) were contoured as per the protocol using T1 and T2 MR images, with approximately 2 cm of expansion from GTV to CTV and an additional 3 mm of expasion from CTV to PTV. GTV1 was defined from the post-operative T2 or FLAIR as enhancement plus surgical cavity, while GTV2 was defined as the enhancement plus surgical cavity on post-op contrasted T1 MR. PTV1 (mean volume 435, $\sigma = 142$ cm^3) and PTV2 (mean volume 260, $\sigma = 102$ cm^3) were prescribed doses of 51 Gy (an increase of 5 Gy from the RTOG protocol) and 60 Gy, respectively, in a single IMRT plan. In four of the cases as a result of edema or

other factors making the PTVs very similar, only a single PTV was used with a prescription of 60 Gy.

Inverse optimization requires several inputs, of which target and normal tissue segmentations is one type, and is subject to many variables in the process of producing a treatment plan. As this study was focused on dosimetric differences as a result of normal tissue segmentation, attempts were made to control other variables as tightly as possible while keeping the overall process clinically realistic. For each patient, nine IMRT plans were inversely optimized, one each for the expert rater (P_1-J_4) and automatic (A_1) segmentation sets. A different rater was randomized to each patient for initial optimization (excluding A_1 and P_2, who was often different from the group geometrically). Using this rater's structures, a planner chose five to seven static gantry angles with energies of 6 or 10 MV (at most two beams of 10 MV) for step-and-shoot IMRT and determined an optimization strategy of constraints and priorities. After perhaps several rounds to determine a good set of parameters, the optimization routine was reset and run for 75 iterations, with a maximum of 100 segments, 4 cm^2 minimum segment size, minimum 2 MU per segment, and a dose grid of 2.5 mm in each dimension. Then, for each of the remaining 7 raters and A_1 the trial was copied and rerun with the only change being that of the normal tissue inputs. There was no human intervention within the optimization after determination of adequate parameters using the first rater. Normal tissue tolerances for planning purposes were those of RTOG 0837: maximum doses of 60 Gy to brainstem, 56 Gy to optic chiasm, 55 Gy to optic nerves, and 50 Gy to eyes.

IV.2.3 Data analysis

Dose matrices were exported via DICOMRT, and dose to specific organ structures was captured through convolution of each plan dose matrix with binary masks of the *de novo*, ground truth, and target segmentations co-registered in CT-space. Dose volume histograms were calculated as well as dosimetric figures of merit. Dose to the ground truth segmentations was used to evaluate impact of segmentation differences to our best estimate of the true organs at risk (the ground truths estimations). We also calculated the difference in figures of merit, such as maximum dose, *reported* by a particular plan's native segmentations and the same as measured by the ground truth estimations to gauge what we term discrepancies in dose reporting. Whereas the first method provides a best predictor of reality had a patient been treated with the plan, the second method assesses how different the reality is from what was planned.

Dosimetric figures of merit

To evaluate relative quality of treatment plans, we calculated several dosimetric figures of merit with general guidance from the quantitative evaluation of normal tissue effects in the clinic (QUANTEC) recommendations (Mayo, Martel, Marks, Flickinger, Nam and Kirkpatrick, 2010; Mayo, Yorke and Merchant, 2010; Lawrence et al., 2010; Jackson et al., 2010). For each of the 180 plans (9 per patient) mean, maximum, V_{45}, V_{54}, V_{59}, V_{64}, and D_{1mL} were determined. The volume doses are defined as the percent volume of a structure receiving the specified dose (Gy), while the dose to 1 mL (D_{1mL}) is defined as the minimum dose of the highest dosed 1 mL of a structure. On a DVH plotted for absolute volume, D_{1mL} is the dose corresponding to a volume of 1 mL.

We calculate figures of merit and their differences with the following. First, we declare $F(i, dose_j, struct_{k,l})$ as a figure of merit, F, for patient $i \in \{1, 20\}$ on the dose distribution j, which is the distribution produced by optimizing a plan on segmentations belonging to rater j. Rater $j \in \{1, 9\}$ represents A_1 and experts P_1-J_4, respectively. To calculate the figure of merit, we must also specify the measuring segmentation, that is, the segmentation used to assess the distribution. We have segmentation types $k \in \{1, 6\}$ for brainstem, chiasm, left and right eyes, and nerves, and segmentation sources $l \in \{1, 11\}$ representing A_1, the 8 experts raters, and the two ground truth estimates, respectively. We could in fact measure any dose distribution via any rater, but in this study we have restricted the calculations to either $l \in \{10, 11\}$, the ground truths, or $j = l$, in which case a distribution is measured via the segmentations used to produce it. We calculate the figures of merit as

$$F_{GT}(i, j, k) = \texttt{mean}\ F(i, dose_j, struct_{k,10:11}) \tag{IV.1}$$

as measured by the ground truths and as

$$F_{self}(i, j, k) = F(i, dose_j, struct_{k,j}) \tag{IV.2}$$

when one considers what would actually be reported by the structures used to generate the plan. The difference in a figure of merit measured by the ground truths for a given rater from that of his peers is calculated as

$$\Delta F_{GT}(i, j, k) = F_{GT}(i, j, k) - \texttt{mean}\ F_{GT}(i, peers, k) \tag{IV.3}$$

where *peers* indicates F is calculated for each of the rater j's peers, excluding A_1. We calculate

the reporting discrepancies as

$$\Delta F_{report}(i,j,k) = F_{self}(i,j,k) - F_{GT}(i,j,k) \tag{IV.4}$$

or, the difference measured by self from the that measured by the ground truths.

Average dose-volume histograms

Before calculating dosimetric figures of merit, which are by definition data reduction measures, there is some use in examining the dose distributions in total. Clinically, this is done via qualitative examination of the isodose distributions and quantitatively through inspection of the entire DVH. In our situation, we have 9 plans with 6 critical organs over 20 patients; individual comparisons are too cumbersome over so many curves. To evaluate plans over all rater-structure-patient combinations, we have generated average DVHs that weight each patient equally. These DVHs, relative in volume, may not be representative of any particular patient DVH but should highlight any gross, systematic differences over all rater-structure combinations.

Statistical inference

As most distributions were non-normal, non-parametric tests were used to test for significance. We used Friedmans test (Friedman, 1937) on ranks first to look for family-wise significance of differences between rater-derived plans (for each figure of merit), followed by pair-wise comparisons via Wilcoxon signed-rank test where reasonable strength of evidence of differences was found. If we were to make all pair-wise comparisons between raters there would be 36 non-redundant comparisons for each family (each figure of merit). Rather than compare individuals, we computed the mean of each individuals peers (this is equivalent to the last term in equation IV.3), and made those comparisons. Since no rater was deemed *a priori* to produce a superior dose distribution than the others, interpretation of any single comparison would be ambiguous. We used a right-tailed signed-rank test as we were most interested in plans that overdose the normal tissues compared to their peers. We did not correct for effects of multiplicity on the family-wise type I error rate, as in this situation we did not want to sacrifice power. We approached significance in this work from the Fisherian perspective, which eschews long term error rates (e.g., $\alpha = 0.05$, $P < 0.05$) in favor of interpreting P-values as indices of evidence; an excellent discussion of such is provided by Lew (Lew, 2012).

As both an indication of the agreement of figures of merit amongst the raters and to gauge whether use of their mean would be appropriate in tests of significance, we calculated

Table IV.1: Doses to targets from the A_1-derived plans in comparison to the range of the physician-derived plans.

Plan	Mean Dose [Gy]		V95 [%]		Min Dose [Gy]		Max Dose[Gy]	
	CTV1	CTV2	CTV1	CTV2	CTV1	CTV2	CTV1	CTV2
A_1	58.72	60.05	100.00	97.1	50.56	54.80	62.66	62.90
Lowest rater	58.70	60.03	99.99	97.1	49.74	54.39	62.54	62.68
Highest rater	58.74	60.06	100.00	97.2	50.57	54.94	62.82	63.02

intra-class correlation coefficients (ICC). The ICC can be thought of as the proportion of observed variance that is true; that is, it separates rater-contributed (within-subject) variance from patient (between-subject) variance. In rater reliability studies it is often used as evidence of interchangeability of raters or as the validity in using the mean of a group of raters as an outcome measure. As we expect the ICCs to be potentially overly optimistic, we chose a conservative one-way model (ICC(1,8), Shrout and Fleiss nomenclature (Shrout and Fleiss, 1979; McGraw and Wong, 1996), which models random effects of subject (patient) and treats rater effects as random error. Calculated ICCs ranged from 0.92 to 1.0, indicating high agreement between the experts in relation to the variability between-patient. This adds validity to our decision to compare raters against the mean of their peers rather than making many more pairwise comparisons. However, it is noteworthy that ICC measures agreement rather than variability (Haber et al., 2005) and we do not employ it to assess the latter.

IV.3 Results

IV.3.1 Impact on target coverage

We measured the impact of segmentation differences on target coverage via the mean dose, V95 (volume to 95% of prescription dose), minimum dose, and maximum dose. We found no clinically important differences in dose coverage as a result of utilizing different rater-derived critical structures in the optimization. The mean doses to CTV1 and CTV2 over all patients and expert-rater derived plans were 58.73 and 60.05 Gy, respectively, and the variation between the lowest and highest of the 9 plans (using OARs from A_1,P_1-J_4) was less than 10 cGy. Table IV.1 compares the target doses as a result of A_1-derived plans to the range of physician-derived plans.

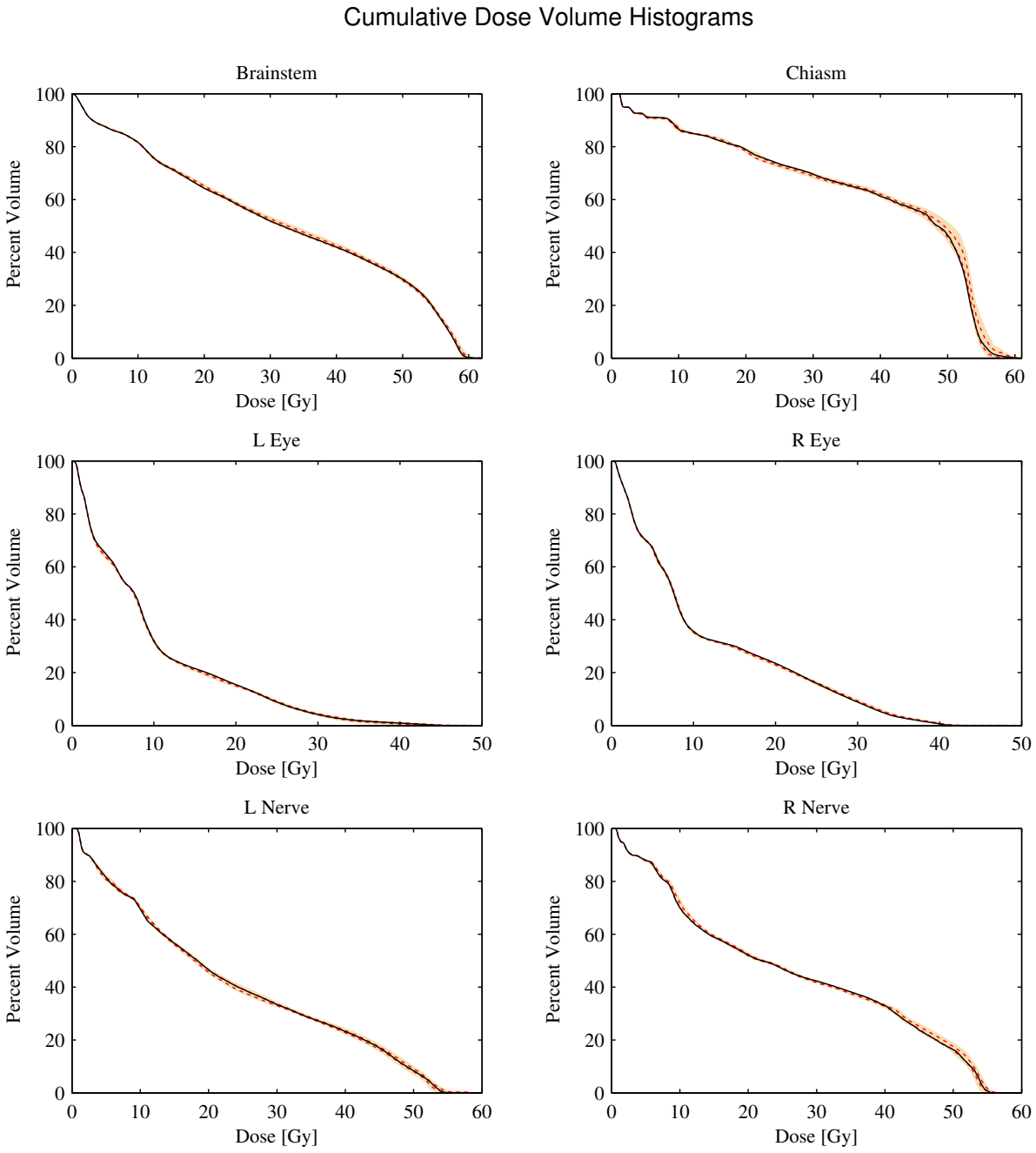

Figure IV.1: Average dose volume histograms over the 20 patients and 9 rater-derived plans. The solid area denotes the mean minimum and maximum extent of DVHs from plans P_1-J_4 with the 95% CI about their mean in red. A_1 is displayed as a solid black line.

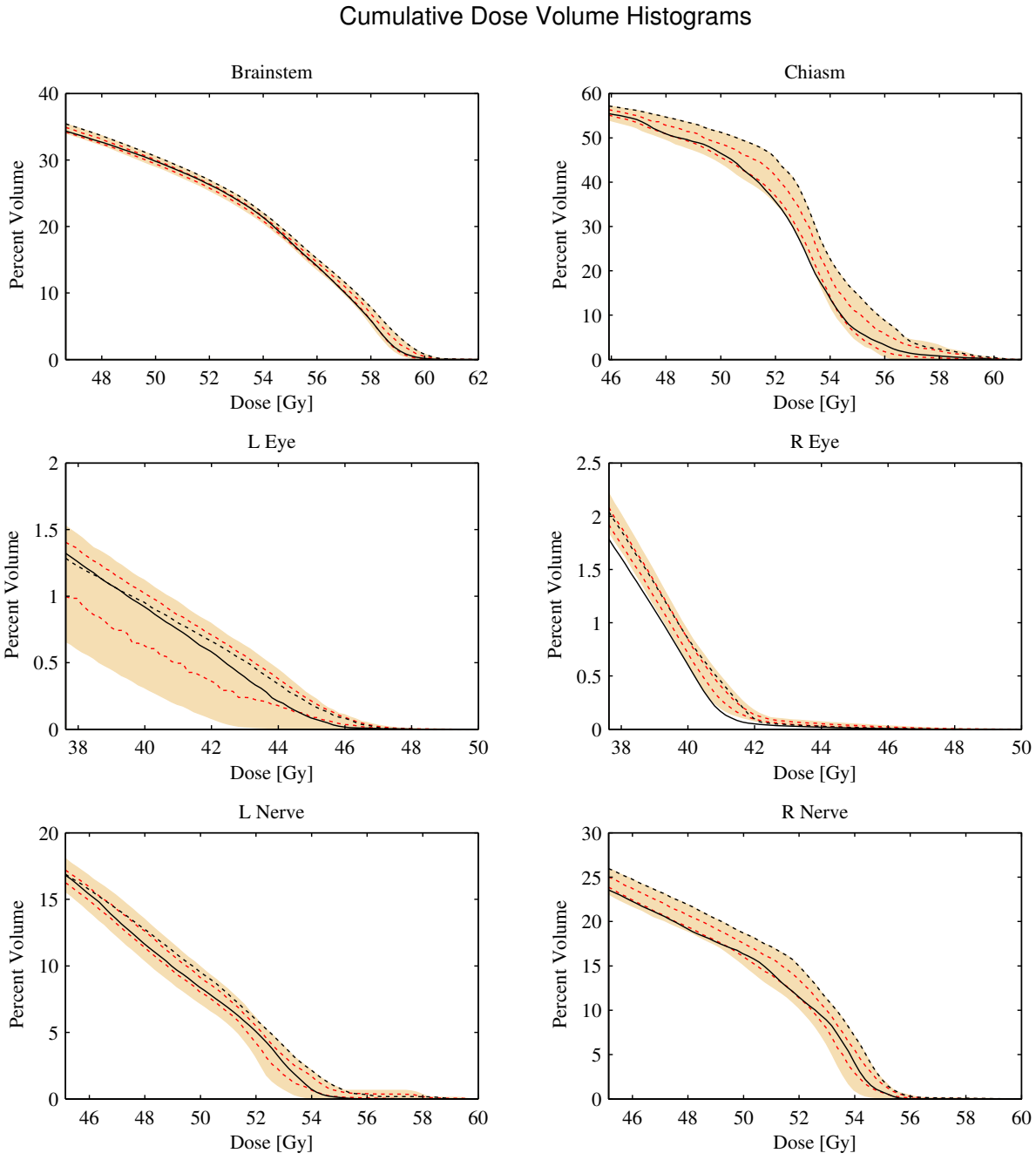

Figure IV.2: Zoomed to the upper 25% of the dose ranges in figure IV.1

Table IV.2: P-values from Friedman's test for significance in differences between raters.

	Mean	Max	V_{45}	V_{54}	V_{59}	D_{1mL}
Brainstem	0.005	0.004	0.049	0.157	< 0.001	<0.001
Chiasm	0.001	< 0.001	0.011	0.078		
Left Eye	0.107	0.381				
Right Eye	0.267	0.003				
Left Nerve	0.004	< 0.001	0.014			
Right Nerve	0.123	0.023	0.147			

Table IV.3: P-values from Wilcoxon sign-rank test for significance in differences between raters and the mean of their peers.

	A_1	P_1	P_2	P_3	P_4	J_1	J_2	J_3	J_4
					Brainstem				
Mean	0.433	0.955	0.001	0.652	0.552	0.996	0.652	0.981	0.106
Max	0.222	0.463	0.005	0.233	0.979	0.999	0.281	0.939	0.070
V_{45}	0.552	0.975	0.012	0.880	0.942	0.966	0.729	0.448	0.058
V_{59}	0.883	0.849	0.006	0.396	0.994	1.000	0.867	0.994	0.235
D_{1mL}	0.180	0.848	0.001	0.552	0.820	0.999	0.086	0.998	0.171
					Optic Chiasm				
Mean	0.829	0.995	0.001	0.680	0.507	0.925	0.433	0.639	0.994
Max	0.027	0.778	<0.001	0.004	0.375	1.000	0.010	0.939	0.998
V_{45}	0.926	0.999	0.012	0.416	0.160	0.681	0.618	0.120	0.897
V_{54}	0.924	0.953	0.042	0.555	0.896	0.640	0.756	0.820	0.849
					Right Eye				
Max	0.996	0.200	0.652	0.887	0.999	0.778	0.680	0.004	0.581
					Left Nerve				
Mean	0.433	0.959	0.032	0.537	0.035	0.004	0.968	0.522	0.987
Max	0.200	0.999	0.004	0.981	0.120	0.005	0.865	0.743	0.778
V_{45}	0.515	0.961	0.311	0.425	0.396	0.001	0.810	0.485	0.485
					Right Nerve				
Max	0.988	0.995	0.002	0.086	0.639	0.245	0.810	0.755	0.820

IV.3.2 Impact of segmentation on plan quality as measured by dose to ground truth

We measured plan quality via the doses delivered to the ground truth segmentations. These calculations provide our best estimate of true dose to the organs at risk from the A_1 and expert-derived plans. The tabulated mean, 95% confidence interval on the mean, and the coefficients of variation are presented in full as supplemental material (section IV.6). First, we found the maximum dose delivered over all plan-patient-structure combinations was 61.6 Gy, which indicates immediately that the highest risk factor for brainstem injury, V_{64}, was not impacted by segmentation differences. We evaluated relative plan quality and variability by calculating for each rater the differences from their peers. Figure IV.3 plots the distributions of differences in maximum dose (IV.3) for each rater with the mean and 95% CI over all structures. Figure IV.4 plots the same for V_{45}, V_{54}, V_{59}, and D_{1mL} for the brainstem. It can be clearly seen that plans derived from A_1 vary generally no more from the physician plans than the physicians vary from their peers. The results of Friedman's test for significance over the entire group and the paired Wilcoxon sign-rank tests are provided in tables IV.2 and IV.3.

IV.3.3 Discrepancy in dose reporting

In the previous section, we characterized differences in plans through their impact on the ground truth segmentation dosimetry. This analysis did not require direct use of the rater segmentations; that is, we ignored the doses as delivered to structures that were actually used to derive the plans. Here we evaluate the plans as a function of the difference in dose between the rater segmentation (about which the plan was optimized) and the ground truth segmentations (IV.4).

We found dose reporting discrepancies were commonplace and large within this study. Across all patients, rater-derived plans, and structures, 60% of reported maximum doses differed from the ground truth doses by greater than 0.5 Gy, 54% greater than 1 Gy, and 33% by greater than 2 Gy. Table IV.4 provides the maximum under- and over-reported doses for A_1, P_2, and the other physicians as a group. To evaluate both bias and variability as well as a relationship to segmentation accuracy, in figure IV.5 we plotted the reporting differences against DSC over A_1, P_2, and the other physicians as a group. Similar to what has been proposed by Bland and Altman (Altman and Bland, 1983; Bland and Altman, 1986), we provided the limits of agreement as the 95% confidence interval on the mean difference between A_1 and the ground truths. As suggested by Kelley (Kelley, 2005), we calculated both the parametric and the bootstrap intervals, considering the parametric limits as a worst case scenario. One can see that A_1-derived plans compare favorably to the physicians. For simplicity we plotted physicians

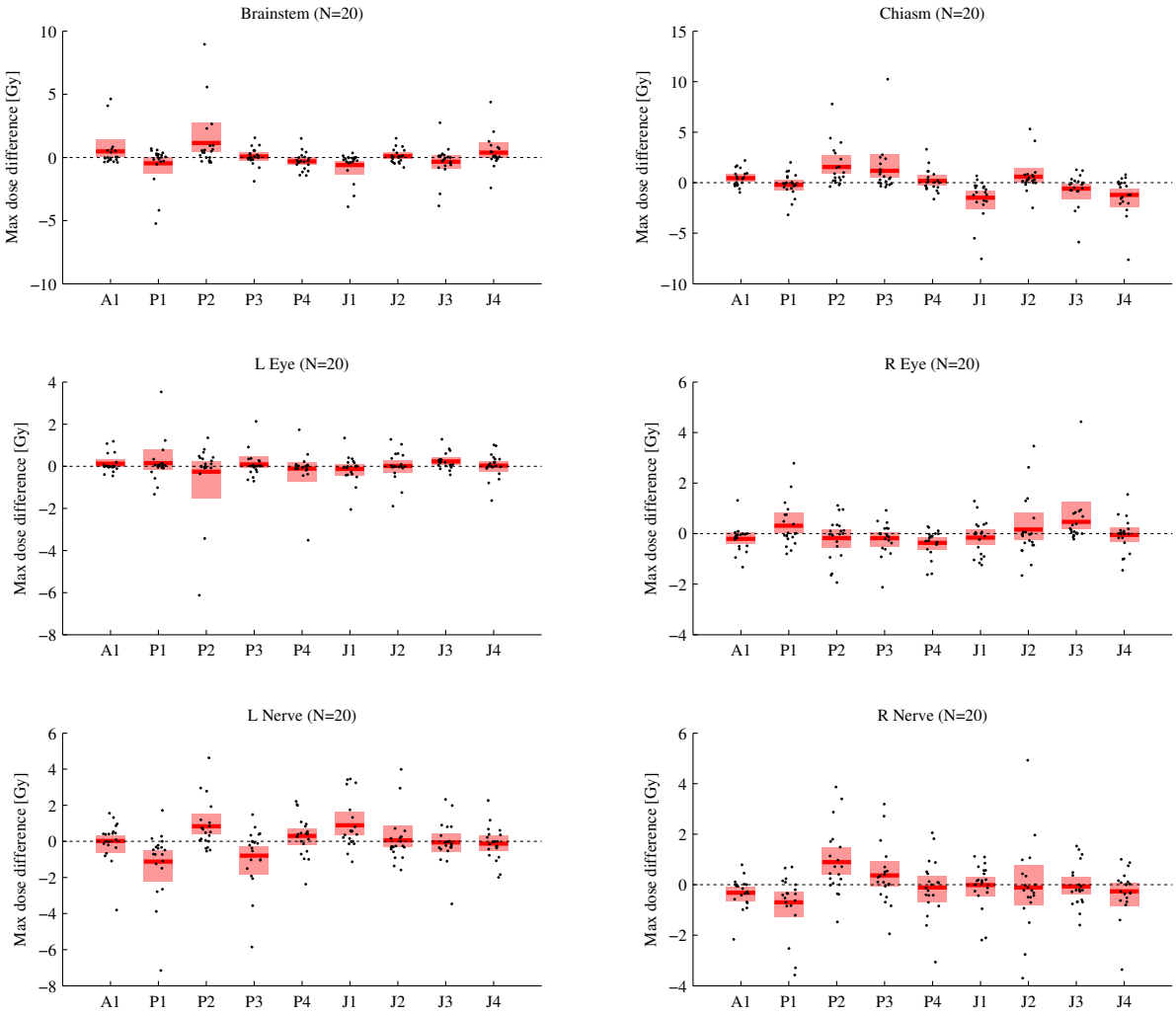

Figure IV.3: Difference in maximum dose from peers. Each black dot represents a difference in maximum dose from the mean of the rater's peers. The mean difference and 95% confidence interval are displayed via the red horizontal line and encompassing box.

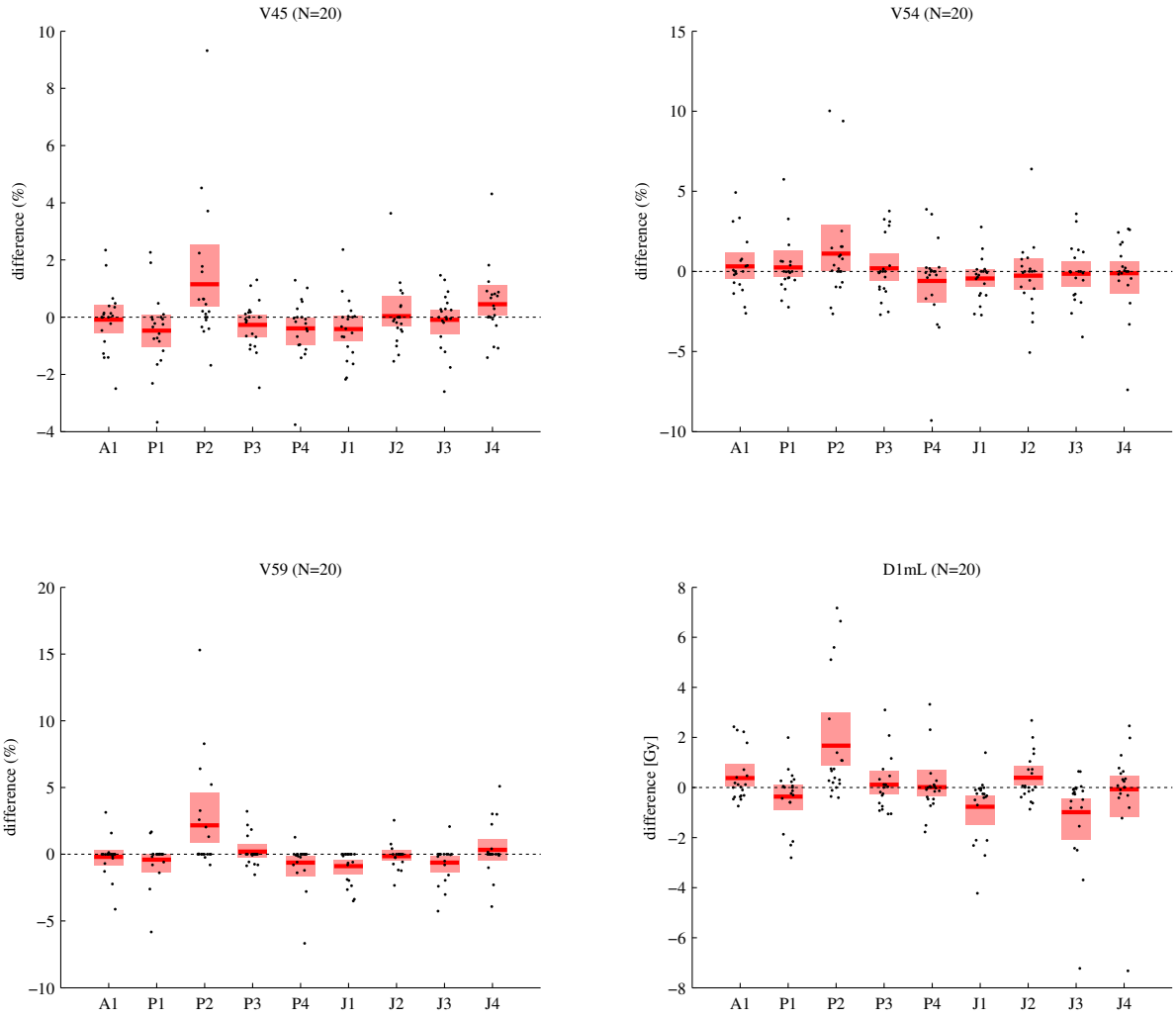

Figure IV.4: Difference in volume dose metrics for the brainstem. The distributions, mean, and 95% CI on the mean of differences in volume dose metrics are plotted. The unit [%] of the y-axis in the V_{xx} plots is absolute difference in percent volume, not a percent error.

Table IV.4: Discrepancies in dose reporting from the ground truth. Maximum under- and over- reported doses in Gy are presented for A_1, P_2, and the other raters as a group. The last row reports the rater associated the value from row three, "others".

	Brainstem		Chiasm		Eyes		Nerves	
Rater	under	over	under	over	under	over	under	over
A_1	-8.24	+2.27	-12.30	+0.02	-2.33	+5.77	-5.21	+6.67
P_2	-33.52	+0.61	-21.70	0.00	-13.36	+4.10	-36.08	+1.26
Others	-6.17	+7.58	-20.08	+17.39	-16.78	+5.88	-27.01	+8.31
Identity	P_4	J_1	P_3	J_1	J_3	J_2	J_1	P_1

P_1, P_3-J_4 as a single group, and though we found some individual physicians performed as well or better than A_1, as a group they exhibit more variability. From both table IV.4 and figure IV.5 we found P_2 clearly results in plans with the most bias and variability in dose reporting, typically under-reporting of dose to the normal tissues.

IV.4 Discussion

Understanding the impact of segmentation variability is important because segmentations are a principal input to treatment plan optimization. Predicting the impact of segmentation differences is challenging as they act in a complex process that also involves tumor geometry, beam characteristics, and clinical dosimetric requirements. Likewise, interpreting impact is also difficult as there is no omnibus measure of quality.

In terms of target coverage, we found the optimization algorithm was invariant to differences in normal tissue segmentation. The differences in mean dose to the targets were less than 10 cGy, and similarly for V95 differences were within tenths of a percent. The minimum and maximum doses to the targets varied slightly more, but their range of variation over all plans, including that of A_1, was less than 2% of the prescription dose. We expected to find larger differences in target coverage for several reasons. First, the tumors were large and sometimes quite close to the normal tissues. Second, the dose grid and optimization parameters were chosen so as not to be a limiting factor. Third, higher priority was given to sparing the normal tissues than covering the targets. We postulate that the optimization simply was driven much more strongly by target volume and proximity than by what were relatively small (in comparison to target) volume differences in rater segmentations. Our study cannot address whether the same would be found with other treatment planning systems. However, in so much as we believe our planning process to be reflective of what is common, these results should be indicative of at least the sizable population of users employing the same treatment planning system.

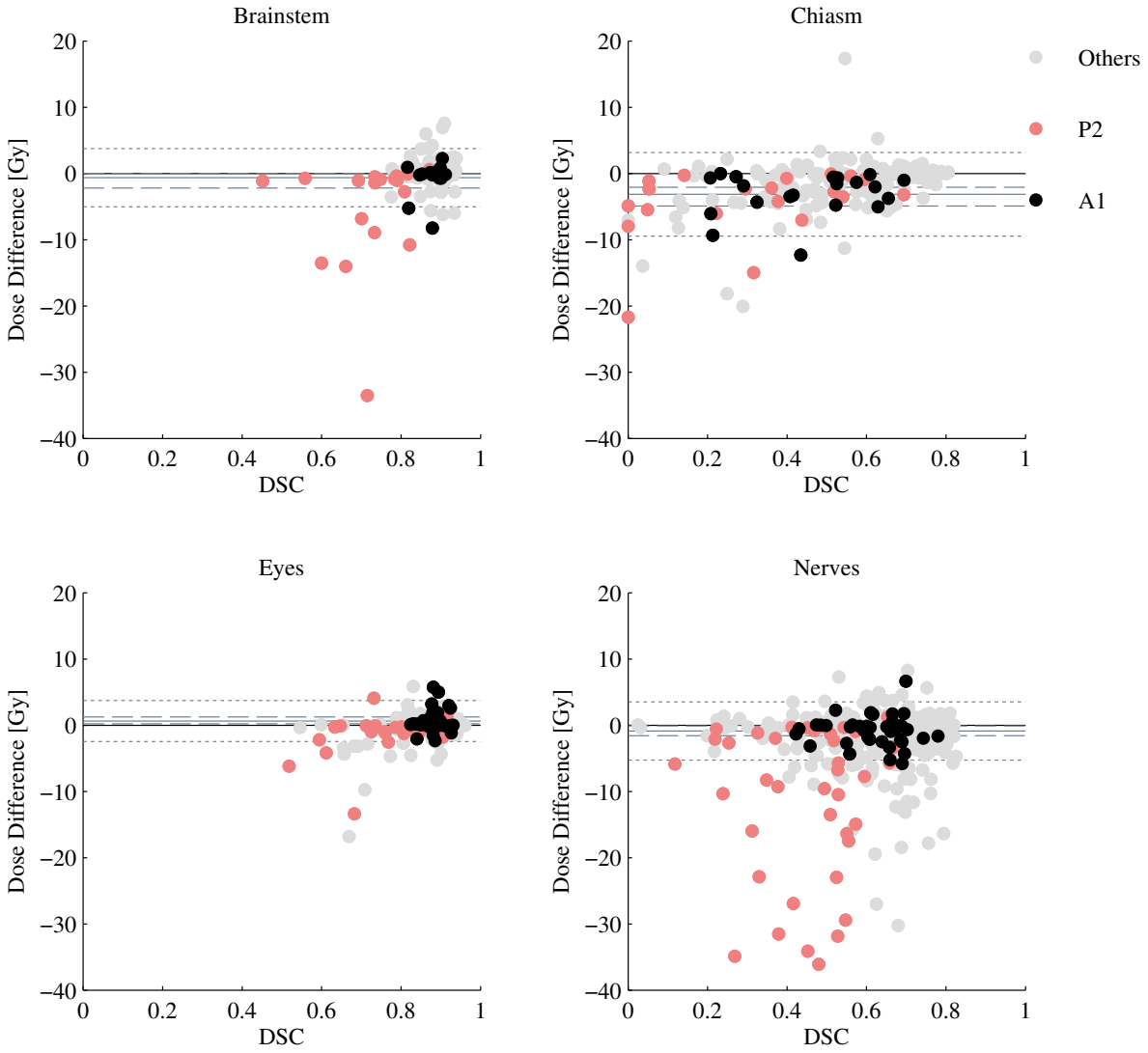

Figure IV.5: Dose reporting differences against DSC. The differences between rater-reported and ground truth reported maximum dose are plotted as a function of Dice coefficient. Three groups are plotted A_1 in black, P_2 in pink, and the other experts in gray background. The mean (solid gray line) and limits of agreement for A_1 reporting differences with both a parametric (gray dotted lines) and bootstrap (gray dashed lines) estimates.

A distinction between our work and that of others is that we approach the problem regarding normal tissues from two different yet important perspectives: dose to ground truth and dose reporting discrepancies. Nelms and colleagues (Nelms et al., 2012) undertook a similar study (single patient, 32 raters) with a narrower aim: to characterize dosimetric differences that occur as a result solely of segmentation differences. This is equivalent to what we term dose reporting discrepancy. That study, however, based all assessments of variability, both geometric and dosimetric, on a singular reference standard (a single manual contour set). We argue this is not an optimal approach. One cannot know whether the reference standard has high or low quality. Even when the goal of the work is confined to characterizing variance, the quality of the reference standard is quite important. If it is skewed toward one end or the other of the spectrum, variability will look very different. Better alternatives are to examine the distributions of pairwise comparisons between raters and/or to calculate a more optimal ground truth from the cohort of segmentations available. In this work we used ground truth estimates, which we have shown previously (Deeley et al., 2011) to produce higher group consensus than using an individual as a standard. In so doing, we argue that rather than assessing variability we can make the stronger statement that this study has assessed relative quality.

The 95% confidence intervals (tables in IV.6) were large, indicating that likely a much larger patient sample size would be needed to yield good absolute estimates of the figures of merit. However, the goal of the study was not to make inferences about the population average maximum dose, for example. Rather, the goal was to compare plan quality. Using a non-parametric test for group differences of matched samples, we found reasonable strength of evidence (Friedman's test, table IV.2) that differences did exist among the rater-derived plans for several figures of merit. In subsequent positive pairwise comparisons between individuals and the mean of their peers, we found the strongest evidence that P_2 segmentations results in plans with higher dose figures of merit. The only figure of merit-structure combination in which the A_1-derived plan differed was for the maximum dose to the chiasm. Three of the experts also produced credible differences in this venue. Interestingly, in our *de novo* study we found A_1, P_1, P_3, and J_3 scored best in a test of distance error on the chiasm; on average at least 80% of the points they contoured lay within 2 mm of the ground truth. However, of those, A_1 and P_3 showed evidence of difference in maximum dose. The reason may be explained by the following. For P_1 and J_3, distances errors tended to be positive (these raters erred on the side of more conservative, outwardly larger though generally well-scaled contours), while those of A_1 and P_3 were close to zero or negative (erred toward smaller, more central contours). P_2 produced chiasm contours that were inconsistent, sometimes missing the ground truth chiasm entirely. The evidence of difference in A_1 and P_3 plans is likely a result of their tighter boundaries

being less protective within the optimization. This underscores some of the complexity of the potential relationships between segmentation quality and dosimetric output.

In contrast to the significance tests, the average DVHs and distribution of dose difference plots offer a sense of effect size. The DVHs show that on average there were very small, likely clinically unimportant, differences in volume dose to the ground truths between the plans. This with the target dose evidence is an indication that differences in the normal tissues segmentations did not produce clinically meaningful differences, on average, in the treatment plans. There were individual instances, however, that may have been clinically important, with differences in maximum dose as high as 10 Gy more to the brainstem and chiasm than delivered by the rest of the group. The automatic system resulted in plans that performed quite well within the variation of the physician-derived plans.

There was much more variability in dose reporting than in the analysis of dose to ground truth. We focused on discrepancies in maximum dose reporting as this is the most commonly reported figure of merit for the structures in our study. These discrepancies are important for several reasons. In the process of optimizing a plan, if the goals cannot be met, a clinical decision will have to be made whether to spare the normal tissue or compromise tumor coverage. When dose has been over-reported, unnecessary action may be taken that results in suboptimal tumor dose. Likewise, an under-reporting plan may result in a decision to treat the tumor more generously, thereby inadvertently overdosing the normal tissues. Dose reporting could also have implications for evidence-based medicine as well. Clinical trials collect the reported doses from enrolled patients and correlate with toxicity. Mayo and colleagues (Mayo, Martel, Marks, Flickinger, Nam and Kirkpatrick, 2010; Mayo, Yorke and Merchant, 2010) reviewed toxicity studies of the brainstem and optic pathway and found there little consensus. A contributing factor may be high variability and often inaccurate reporting, especially concerning maximum dose. Even for the small tubular structures, reporting of mean and volume doses seems to be indicated; we found these figures of merit were more accurate and less variable than maximum dose.

IV.4.1 Limitations and future work

This study was designed to test the impact of segmentation differences, principally with respect to the feasibility of our automatic system, on plan quality. To the best of our knowledge this is a step further than has been done previously. In doing so we employed 8 experts and 20 patients with challenging tumor geometry. However, our inferences are tempered by the realization that there are additional potentially important variables not captured. We cannot know whether our sample is representative of the population of patients with large brain

tumors, physicians, or treatment planning procedures. As mentioned previously, alternate dose optimization processes could be more less sensitive to segmentation differences than that which we employed. The same can be said for dose prescription. We chose a treatment regimen similar to a common clinical protocol, but the spectrum of protocols and institutional preferences is broad. For instance, in a dose escalation study such that of Tsien and colleagues (Tsien et al., 2009), figures of merit and potentially dose coverage may be more sensitive to differences in segmentations than captured in our study.

IV.5 Conclusions

Our system for automatic segmentation resulted in IMRT treatment plans within the range of those produced by expert physician segmentations as measured by dose to ground truth through a number of figures of merit. Target dose coverage was robust to segmentation differences, and average normal tissue DVHs were similar as measured by the ground truth estimates. Measurement via the ground truth estimates provided our best guess as to the true impact of differences. The variation in this analysis was muted compared to that of the dose reporting discrepancies. We found reporting, especially of maximum dose, varied widely to as much as 10-30 Gy and favored under-reporting. This could have implications in studies of normal tissue toxicity that assume accurate reporting. Maximum dose, while the most commonly employed figure of merit for the normal tissues of the brain, is more susceptible to these variations than volume doses. Our results indicate one should report volume as well as maximum doses for all critical structures.

IV.6 Supplemental Material

Table IV.5: Brainstem figures of merit for the 9 rater-derived plans as well as the junior, senior, and all physicians as a single group.

	Mean Dose [Gy]				Max Dose [Gy]				V45 [%]				V54 [%]			
Rater	Mean	Mean CI		cv	Mean	Mean CI		cv	Mean	Mean CI		cv	Mean	Mean CI		cv
A_1	31.85	26.48	35.99	0.48	54.20	48.58	57.00	0.24	36.37	28.33	46.42	0.83	21.36	15.07	29.17	1.08
P_1	31.76	27.06	36.19	0.48	53.32	47.08	56.28	0.26	36.04	27.41	45.81	0.85	21.27	14.14	28.50	1.09
P_2	32.43	27.59	37.01	0.46	54.73	49.21	57.81	0.24	37.46	28.08	47.71	0.81	22.03	15.68	29.01	1.01
P_3	31.87	27.00	36.73	0.48	53.79	48.24	56.68	0.25	36.22	27.51	45.83	0.85	21.22	14.34	28.67	1.09
P_4	31.84	27.32	36.38	0.47	53.45	47.24	56.26	0.25	36.11	27.51	45.56	0.84	20.53	13.89	28.12	1.09
J_1	31.67	26.96	36.34	0.48	53.20	46.45	56.26	0.25	36.09	27.15	45.56	0.85	20.67	14.10	27.67	1.08
J_2	31.87	27.25	36.27	0.47	53.82	48.58	57.07	0.25	36.49	27.63	46.44	0.83	20.82	14.13	27.74	1.06
J_3	31.75	27.05	36.24	0.48	53.42	47.28	56.65	0.26	36.37	27.65	46.00	0.84	20.93	15.00	28.73	1.11
J_4	31.99	26.94	36.26	0.48	54.06	48.01	57.08	0.25	36.85	27.61	45.88	0.83	20.96	14.38	27.59	1.07
All senior	31.97	29.36	34.04	0.47	53.82	51.53	55.53	0.25	36.46	31.75	41.12	0.83	21.26	18.07	24.86	1.06
All junior	31.82	29.21	33.97	0.47	53.62	51.06	55.57	0.25	36.45	31.63	41.10	0.83	20.84	17.52	24.50	1.07
All experts	31.90	30.16	33.55	0.47	53.72	52.04	55.06	0.25	36.45	33.24	39.91	0.83	21.05	18.83	23.61	1.06

	V59 [%]				D_{1mL} [Gy]			
Rater	Mean	Mean CI		cv	Mean	Mean CI		cv
A_1	1.48	0.89	2.58	1.77	49.80	40.27	53.75	0.28
P_1	1.33	0.74	2.06	1.64	49.10	40.60	53.92	0.30
P_2	3.58	2.19	5.79	1.60	50.88	41.85	54.97	0.28
P_3	1.86	1.13	2.93	1.53	49.52	40.81	54.06	0.29
P_4	1.13	0.63	2.09	1.87	49.44	41.25	53.72	0.29
J_1	0.90	0.48	1.42	1.70	48.75	41.60	53.52	0.30
J_2	1.56	0.97	2.26	1.43	49.76	40.83	54.33	0.29
J_3	1.13	0.63	1.76	1.56	48.56	40.03	53.31	0.31
J_4	1.97	1.20	3.12	1.63	49.36	40.85	54.19	0.31
All senior	1.97	1.51	2.66	1.84	49.73	46.21	52.26	0.28
All junior	1.39	1.06	1.76	1.65	49.11	45.55	52.01	0.30
All experts	1.68	1.39	2.05	1.81	49.42	46.84	51.25	0.29

Table IV.6: Chiasm dosimetric figures of merit.

	Mean Dose [Gy]				Max Dose [Gy]				V45 [%]				V54 [%]			
Rater	Mean	Mean CI		cv	Mean	Mean CI		cv	Mean	Mean CI		cv	Mean	Mean CI		cv
A_1	38.60	32.37	43.31	0.46	46.79	39.89	51.26	0.38	56.39	42.90	68.30	0.77	13.81	8.44	21.70	1.54
P_1	38.26	32.22	42.97	0.46	46.17	40.18	50.69	0.38	55.12	40.75	67.64	0.78	13.01	8.24	19.60	1.44
P_2	39.54	33.02	44.57	0.46	47.71	40.97	52.08	0.38	57.89	44.27	71.34	0.74	22.73	14.40	32.37	1.26
P_3	38.70	32.68	43.55	0.45	47.38	40.05	51.86	0.38	56.88	43.82	70.06	0.75	15.02	10.03	20.52	1.15
P_4	38.63	32.96	43.88	0.46	46.51	40.26	51.73	0.38	57.07	43.50	70.47	0.75	12.12	7.74	18.55	1.38
J_1	38.52	33.03	43.62	0.47	45.05	38.28	49.71	0.40	57.01	43.80	70.21	0.77	16.27	9.40	25.98	1.66
J_2	38.82	32.73	43.76	0.46	46.86	40.51	51.55	0.38	56.74	42.49	69.31	0.76	17.76	10.64	27.39	1.52
J_3	38.63	32.86	43.40	0.46	45.83	39.29	50.69	0.39	57.36	43.18	70.49	0.76	14.22	8.50	22.72	1.61
J_4	38.33	32.22	43.42	0.47	45.29	38.44	50.28	0.40	56.24	43.11	69.52	0.78	15.39	9.07	24.05	1.60
All senior	38.78	35.96	41.33	0.45	46.94	44.09	49.35	0.38	56.74	49.36	62.61	0.75	15.72	12.67	19.34	1.35
All junior	38.57	35.72	41.20	0.46	45.76	42.72	48.50	0.39	56.84	50.47	63.80	0.76	15.91	12.13	19.99	1.58
All experts	38.68	36.71	40.63	0.46	46.35	44.36	48.27	0.38	56.79	51.64	60.80	0.75	15.81	13.53	18.84	1.47

Table IV.7: Left and right eye dosimetric figures of merit.

	Left Eye								Right Eye							
	Mean Dose [Gy]				Max Dose [Gy]				Mean Dose [Gy]				Max Dose [Gy]			
Rater	Mean	Mean CI		cv	Mean	Mean CI		cv	Mean	Mean CI		cv	Mean	Mean CI		cv
A_1	9.64	7.54	12.64	0.88	16.35	12.09	21.45	0.89	11.56	8.83	14.79	0.83	20.91	16.22	25.80	0.77
P_1	9.54	7.27	12.48	0.87	16.35	12.61	21.62	0.90	11.66	9.20	14.81	0.83	21.39	16.52	26.72	0.76
P_2	9.62	7.24	12.39	0.88	16.00	12.54	20.91	0.87	11.51	8.78	14.46	0.83	20.96	16.03	26.25	0.76
P_3	9.66	7.15	12.88	0.89	16.30	12.48	20.97	0.89	11.56	8.94	14.85	0.84	20.96	16.42	25.95	0.78
P_4	9.51	7.27	12.25	0.86	16.13	12.26	20.77	0.88	11.49	8.95	14.81	0.84	20.80	15.98	25.82	0.79
J_1	9.65	7.18	12.34	0.89	16.11	12.11	20.82	0.90	11.54	8.78	14.66	0.83	20.98	16.35	25.85	0.77
J_2	9.64	7.00	12.65	0.90	16.24	12.38	21.04	0.90	11.58	8.66	14.82	0.84	21.26	16.23	26.17	0.78
J_3	9.55	7.18	12.31	0.89	16.43	11.89	21.13	0.89	11.59	9.03	14.89	0.82	21.53	16.85	27.06	0.77
J_4	9.50	7.18	12.21	0.89	16.25	12.29	21.26	0.89	11.51	8.94	14.71	0.83	21.07	16.44	26.31	0.77
All senior	9.58	8.35	10.90	0.87	16.20	14.02	18.48	0.88	11.56	10.06	13.00	0.83	21.03	18.52	23.58	0.76
All junior	9.59	8.40	10.98	0.88	16.26	13.95	18.59	0.89	11.55	10.19	13.14	0.82	21.21	18.79	23.87	0.77
All experts	9.58	8.74	10.50	0.87	16.23	14.83	17.90	0.88	11.55	10.55	12.56	0.83	21.12	19.47	22.94	0.76

Table IV.8: Left and right optic nerve dosimetric figures of merit.

Rater	Left Nerve								Right Nerve							
	Mean Dose [Gy]				Max Dose [Gy]				Mean Dose [Gy]				Max Dose [Gy]			
	Mean	Mean CI		cv	Mean	Mean CI		cv	Mean	Mean CI		cv	Mean	Mean CI		cv
A_1	22.64	18.35	27.46	0.67	38.91	32.16	44.06	0.49	26.06	21.25	31.46	0.64	39.07	32.51	44.54	0.49
P_1	22.30	18.05	27.14	0.67	37.91	31.90	42.97	0.49	26.07	21.06	31.36	0.65	38.77	32.27	44.35	0.50
P_2	22.77	18.18	27.66	0.66	39.62	33.01	44.90	0.49	26.54	21.50	31.65	0.64	40.16	33.32	45.88	0.49
P_3	22.53	17.93	27.30	0.67	38.20	31.67	43.21	0.49	26.23	21.24	31.78	0.65	39.71	32.70	44.73	0.49
P_4	22.69	18.68	27.41	0.65	39.16	32.92	44.33	0.48	26.14	21.25	31.96	0.65	39.29	32.36	44.37	0.49
J_1	23.07	18.60	27.90	0.66	39.67	32.62	45.36	0.49	26.24	21.27	31.51	0.64	39.37	32.49	44.56	0.49
J_2	22.39	18.02	27.27	0.68	38.95	32.12	44.16	0.50	26.06	21.09	31.11	0.64	39.28	32.69	44.71	0.49
J_3	22.54	17.87	27.54	0.67	38.85	32.29	44.08	0.49	25.98	21.39	31.30	0.63	39.32	32.97	44.57	0.49
J_4	22.37	18.06	27.24	0.67	38.79	32.15	43.84	0.49	25.97	20.93	31.03	0.64	39.15	32.05	44.21	0.49
All senior	22.57	20.32	24.87	0.66	38.72	35.89	41.24	0.48	26.24	23.60	28.82	0.64	39.48	36.50	42.46	0.49
All junior	22.59	20.06	24.88	0.67	39.06	35.98	41.95	0.49	26.06	23.82	28.59	0.63	39.28	36.09	42.03	0.49
All experts	22.58	20.88	24.24	0.66	38.89	37.08	41.05	0.48	26.15	24.27	27.77	0.63	39.38	37.10	41.37	0.49

CHAPTER V

DISCUSSION AND FUTURE DIRECTIONS OF RESEARCH

We have undertaken an investigation to characterize the impact and potential benefit of automatic segmentation in the brain. The three studies presented in this dissertation have shared a common framework of a multi-rater behavioral design. The following observations motivated this design:

1. Segmentations are most basically geometric, but they are used in a complex process resulting in a treatment plan and dose distribution to a patient. Evaluation should reflect both the native form of the segmentation as well as end-use and any interactions in the process, such as when a human rater reviews and edits the automatic segmentations.

2. Medical image segmentation is a problem lacking a well-defined ground truth. Studies utilizing a single expert segmentation as a reference standard can be subject to considerable bias.

3. Segmentations and the product of their end-use, dosimetry, are not objects that can be well-captured by single tests or metrics. Each should be examined by a number of complementary measures to reveal information that may be lost using single measures.

We employed these principles in the design of the studies comprising the main chapters of this dissertation.

In chapter II we recruited 8 experts to segment *de novo* the brainstem, optic chiasm, eyes, and optic nerves of 20 patients who had been previously treated for large brain tumors. We tested our automatic segmentations within the context of the expert variability and accuracy. To test accuracy we calculated ground truth estimates, one using a common approach and another via a simple yet novel approach. We found that the automatic segmentations could serve as a surrogate to the experts. We uncovered several areas in which experts are challenged, particularly the visual pathway of the optic chiasm and nerves. Previous works not employing our multi-rater design have concluded with dissatisfaction that perhaps automatic methods are not well-suited for the segmentation of the visual pathway. Indeed, they are challenged, but no more so than the experts. In this context, the benefit of automatic segmentation is principally one of efficiency. Whereas the automatic system requires no user input, we found the average expert required 15 minutes. The efficiency impact may be greater in the

future, however. Other body sites require much more expert time (greater than an hour) for segmentation, particularly in the head and neck. Additionally, if adaptive replanning becomes the new paradigm, segmentation workload could increase substantially.

In chapter III we presented results of a study to test the interaction of the automatic segmentations and the human experts. Even considering the results of the *de novo* study, it is likely automatic segmentations will be reviewed by the end-user. In fact, even when humans segment manually it is advisable that they review their own segmentations for correctness. With this in mind we tested the interaction of the humans and automatic system through editing. We were also interested in the effect of editing beyond the context of the automatic system. Our hypothesis was that editing, regardless of source, may be beneficial to improve the results of what is otherwise a noisy process when starting from a blank slate. To that end we designed a single-blinded randomized set of tasks in which each rater was called to edit the automatic contours, their own contours from the *de novo* study, and those of their peers. We found that editing reduced inter-rater variability and at minimum maintained accuracy across all sources. In areas where raters had performed poorly *de novo*, such as missing slices at the superior and inferior borders of structures, editing A_1 improved performance. This process was even robust to using the lowest quality segmentations as a starting point. We found that efficiency was still improved, as editing of automatic contours for a single patient required on average 6 minutes. Thus, we conclude from this study that the automatic segmentations not just improve efficiency, but they have the potential to reduce geometric inter-rater variability without introducing unwanted bias. They could also prove useful as a learning tool. Rather than be confined to traditional anatomical atlases, users may invoke the automatic system to incorporate the knowledge-base in the atlases to the target patient as a starting point.

The last test is that of the end-use of segmentations. Chapter IV presented a study to test the impact of segmentations differences on radiation therapy treatment plans and to determine whether the automatic system resulted in plans within the variability of the experts-derived plans. Once again, we found the automatic system performed well within the context of the experts. First, we found target dose coverage was robust to all segmentation differences across all patients and raters. Second, we measured the true impact of differences via the ground truth estimates. Statistically significant differences were found, but the magnitude of these difference was not clinically important on average. The only consistent differences across patients and structures were a result of a single expert rater. Third, we found dose reporting discrepancies were common and could be large and tended toward underdosing. This could have important implications for clinical trials and toxicity studies and may explain some of the variability noted between current studies. To our knowledge, this work is the first to separate

true dose (to ground truth) from the concept of dose reporting. Previous work has measured only the latter without distinguishing between the two perspectives.

Studies such as those undertaken in this work are long, costly, and require scarce resources, but the framework and data gathered can be used for a number of future studies.

Algorithm validation studies. In our *de novo* study we developed a framework and gathered rich data from several raters, but we did not test the system on other algorithms. Many algorithms have been developed in recent years to segment the same normal tissues, and perhaps as many as a dozen commercial vendors now offer automatic segmentation as part of a treatment planning system or stand-alone platform. A web-based study to test and compare these algorithms could be undertaken with relative ease now that the rater data and framework exist.

Extension to other body sites. As noted, many algorithms and commercial systems are already being employed clinically. The anatomical site of most interest is the head and neck, as many tissues must be segmented for IMRT planning, both normal and diseased, in a process that often exceeds 1-2 hours. Our group has also begun developing methods to segment the lymph nodes, thyroid, and parotid glands (Chen et al., 2010, 2012). We propose similar studies to the ones presented in the present work but with a different method for data acquisition. In the previous studies we collected data via the clinical treatment planning systems in a tedious process requiring much manual intervention to input and output the necessary data structures. Future studies could employ web-based segmentation tools. This approach trades clinical realism for feasibility and removes barriers to recruiting diverse raters from a number of institutions. In parallel many algorithms could segment structures of interest. Both *de novo* and editing studies could be implemented in this design. This would be a major contribution, especially if an editing study could be accomplished, as prior evaluation work in the head and neck has been exploratory, utilizing only a few raters and patients (Chao et al., 2007; Stapleford et al., 2010).

Studies without matched sets. Fundamental to our previous studies was the collection of data from all experts on the same subjects (patients). This design was motivated by the observation that experts will disagree, and accordingly a single expert should not be used as a reference standard. However, we have now characterized this variability in the brain. If estimates can be had for the head and neck, it is conceivable that a study could be designed for that site utilizing a number of independent samples (each with a single rater from the population of available raters) as the reference standard. Some of these samples will be of low quality, but with an estimate of how often that might happen, one could calculate the sample size needed to provide sound statistics. As in the previous example, this design aims to achieve similar results

to the previous studies with increased efficiency. Along these lines one could undertake a *de novo* study using volumes that have been previously segmented for patient treatment, requiring no new manual segmentation.

BIBLIOGRAPHY

Aird, E. and Conway, J. (2002), 'Ct simulation for radiotherapy treatment planning', *Br J Radiol* **75**(900), 937–949.

Altman, D. and Bland, J. (1983), 'Measurement in medicine - the analysis of method comparison studies', *Statistician* **32**(3), 307–317.

Amelio, D., Lorentini, S., Schwarz, M. and Amichetti, M. (2010), 'Intensity-modulated radiation therapy in newly diagnosed glioblastoma: A systematic review on clinical and technical issues', *Radiother Oncol* **97**(3), 361 – 369.
URL: *http://www.sciencedirect.com/science/article/pii/S0167814010005232*

Asman, A. J. and Landman, B. A. (2012), Non-local staple: An intensity-driven multi-atlas rater model, *in* 'MICCAI (3)', pp. 426–434.

Babalola, K., Patenaude, B., Aljabar, P., Schnabel, J., Kennedy, D., Crum, W., Smith, S., Cootes, T., Jenkinson, M. and Rueckert, D. (2009), 'An evaluation of four automatic methods of segmenting the subcortical structures in the brain', *Neuroimage* **4**, 1435–47.

Bach Caudra, M., De Craene, M., Duay, V., Macq, B., Pollo, C. and Thiran, J. (2006), 'Dense deformation field estimation for atlas-based segmentation of pathological MR brain images', *Compt Methods Programs Biomed* **84**, 66–75.

Bach Cuadra, M., Pollo, C., Bardera, A., Cuisenaire, O., Villemure, J. and Thiran, J. (2004), 'Atlas-based segmentation of pathological MR brain images using a model of lesion growth', *IEEE Trans Med Imaging* **23**, 1301–14.

Beyer, G. P., Velthuizen, R. P., Murtagh, F. R. and Pearlman, J. L. (2006), 'Technical aspects and evaluation methodology for the application of two automated brain MRI tumor segmentation methods in radiation therapy planning', *Magn Reson Imaging* **24**(9), 1167–1178.

Biancardi, A., Jirapatnakul, A. and Reeves, A. (2010), 'A comparison of ground truth estimations', *IJCARS* **5**, 295–305.

Bland, J. and Altman, D. (1986), 'Statistical methods for assessing agreement between two methods of clinical measurement', *Lancet* **1**(8476), 307–310.

Bland, J. M. and Altman, D. G. (1999), 'Measuring agreement in method comparison studies', *Stat Methods Med Res* **8**(2), 135–160.

Bondiau, P., Malandain, G., Chanalet, S., Marcy, P., Habrand, J., Fauchon, F., Paquis, P., Courdi, A., Commowick, O., Rutten, I. and Ayache, N. (2005), 'Atlas-based automatic segmentation of MR images: validation study on the brainstem in radiotherapy context', *Int J Radiat Oncol Biol Phys* **61**(1), 289–98.

Bouchet, L. G., Meeks, S. L., Goodchild, G., Bova, F. J., Buatti, J. M. and Friedman, W. A. (2001), 'Calibration of three-dimensional ultrasound images for image-guided radiation therapy', *Phys Med Biol* **46**(2), 559–577.

Božica, V. and Bojana, B. (2010), 'Radiotherapy: Past and present', *Archive of Oncology* **18**, 140–142.

Burr, D. (1981), 'A dynamic model for image registration', *Comput Graph Image Process* **15**, 102–12.

Cardoso, J. S. and Corte-Real, L. (2005), 'Toward a generic evaluation of image segmentation', *IEEE Trans Med Imaging* **14**(11), 1773–1782.

Chao, K. S., Bhide, S., Chen, H., Asper, J., Bush, S., Franklin, G., Kavadi, V., Liengswangwong, V., Gordon, W., Raben, A., Strasser, J., Koprowski, C., Frank, S., Chronowski, G., Ahamad, A., Malyapa, R., Zhang, L. and Dong, L. (2007), 'Reduce in variation and improve efficiency of target volume delineation by a computer-assisted system using a deformable image registration approach', *Int J Radiat Oncol Biol Phys* **68**, 1512–1521.

Chao, M., Xie, Y. and Xing, L. (2008), 'Auto-propagation of contours for adaptive prostate radiation therapy', *Phys Med Biol* **53**(17), 4533–42.

Chen, A., Deeley, M. A., Niermann, K. J., Moretti, L. and Dawant, B. M. (2010), 'Combining registration and active shape models for the automatic segmentation of the lymph node regions in head and neck CT images', *Med Phys* **37**, 6338–6346.

Chen, A., Niermann, K. J., Deeley, M. A. and Dawant, B. M. (2012), 'Evaluation of multiple-atlas-based strategies for segmentation of the thyroid gland in head and neck CT images for IMRT', *Phys Med Biol* **57**(1), 93–111.

Cheng, C. W., Wong, J., Grimm, L., Chow, M., Uematsu, M. and Fung, A. (2003), 'Commissioning and clinical implementation of a sliding gantry CT scanner installed in an existing treatment room and early clinical experience for precise tumor localization', *Am J Clin Oncol* **26**(3), 28–36.

Christensen, G. and Johnson, H. (2001), 'Consistent image registration', *IEEE Trans Med Imaging* **20**, 568–82.

Cohen, J. (1968), 'Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit', *Psychol Bull* **70**(4), 213–220.

Crum, W., Camara, O. and Hill, D. (2006), 'Generalized overlap measures for evaluation and validation in medical image analysis', *IEEE Trans Med Imaging* **25**(11), 1451–61.

Crum, W., Hartkens, T. and Hill, D. (2004), 'Non-rigid image registration: theory and practice', *Br J Radiol* **77**(Spec 2), S140–53.

Das, I. J., Moskvin, V. and Johnstone, P. A. (2009), 'Analysis of treatment planning time among systems and planners for intensity-modulated radiation therapy', *JACR* **6**(7), 514–517.

Davison, A. and Hinkley, D. (1997), *Bootstrap methods and their applications.*, Cambridge University Press.

Dawant, B., Hartmann, S., Pan, S. and Gadamsetty, S. (2002), 'Brain atlas deformation in the presence of small and large space-occupying tumors', *Comput Aided Surg* **7**, 1–10.

Deeley, M. A., Chen, A. Datteri, R., Noble, J., Cmelak, A., Donnelly, E. F., Malcolm, A., Moretti, L., Jaboin, J., Niermann, K., Yang, E. S., Yu, D. S. and Dawant, B. M. (2013), 'Segmentation editing improves efficiency while reducing inter-expert variation and maintaining accuracy for normal brain tissues in the presence of space-occupying lesions', *Phys Med Biol* **58**(12), 4071–4097.

Deeley, M., Chen, A., Datteri, R., Noble, J., Cmelak, A., Donnelly, E., Malcolm, A., Moretti, L., Jaboin, J., Niermann, K., Yang, E., Yu, D., Yei, F., Koyama, T., Ding, G. and Dawant, B. (2011), 'Comparison of manual and automatic segmentation methods for brain structures in the presence of space-occupying lesions: a multi-expert study', *Phys Med Biol* **56**, 4557–77.

D'Haese, P.-F. D., Duay, V., Li, R., du Bois d'Aische, A., Merchant, T. E., Cmelak, A. J., Donnelly, E. F., Niermann, K. J., Macq, B. M. M. and Dawant, B. M. (2003), 'Automatic segmentation of brain structures for radiation therapy planning', *Proceedings of SPIE: Medical Imaging* **5032**, 517–526.
URL: *http://dx.doi.org/10.1117/12.480392*

Dice, L. (1945), 'Measures of the amount of ecologic association between species', *Ecology* **26**(3), 297–302.

Ding, G., Duggan, D., Coffey, C., Hallahan, D. and Deeley, M. (2006), 'Adaptive IMRT Using Cone-Beam CT: A Case Study On Patients with Bulky Head and Neck Tumors', *Med Phys* **33**(6), 2024.

Driver, D., Drzymala, M., Dobbs, H., Faulkner, S. and Harris, S. (2004), 'Virtual simulation in palliative lung radiotherapy', *Clinical Oncology* **16**(7), 461–466.

Fallone, B. G., Murray, B., Rathee, S., Stanescu, T., Steciw, S., Vidakovic, S., Blosser, E. and Tymofichuk, D. (2009), 'First MR images obtained during megavoltage photon irradiation from a prototype integrated linac-MR system', *Med Phys* **36**(6), 2084–2088.

Feng, J., Ip, H. and Cheng, S. (2004), A 3d geometric deformable model for tubular structure segmentation, *in* Y.-P. P. Chen, ed., '10th International Multimedia Modeling Conference', IEEE Computer Society.

Ford, E. C., Chang, J., Mueller, K., Sidhu, K., Todor, D., Mageras, G., Yorke, E., Ling, C. C. and Amols, H. (2002), 'Cone-beam CT with megavoltage beams and an amorphous silicon electronic portal imaging device: potential for verification of radiotherapy of lung cancer', *Med Phys* **29**(12), 2913–2924.

Friedman, M. (1937), 'The use of ranks to avoid the assumption of normality implicit in the analysis of variance', *Journal of the American Statistical Association* **32**(200), 675–701.

Gorthi, S., Duay, V., Houhou, N., Bach Cuadra, M., Schick, U., Becker, M., Allal, A. and Thiran, J. (2009), 'Segmentation of head and neck lymph node regions for radiotherapy planning using active contour-based atlas registration', *IEEE J Select topics* **3**, 135–47.

Goshtasby, A. A. (2005), *2-D and 3-D image registration: for medical, remote sensing, and industrial applications*, Wiley-Interscience.

Gregoire, V., Jeraj, R., Lee, J. A. and OSullivan, B. (2012), 'Radiotherapy for head and neck

tumours in 2012 and beyond: conformal, tailored, and adaptive?', *The Lancet Oncology* **13**(7), e292 – e300.
URL: *http://www.sciencedirect.com/science/article/pii/S1470204512702371*

Haber, M., Barnhart, H. X., Song, J. and Gruden, J. (2005), 'Observer variability: a new approach in evaluating interobserver agreement', *Journal of Data Science* **3**, 69–83.

Hall, E. J. (1994), *Radiobiology for the Radiologist*, Vol. 117, JB Lippincott Philadelphia.

Hanal, J., Hill, D. and Hawkes, D. (2001), *Medical Image Registration*, first edn, CRC Press.

Hansen, E. K., Bucci, M. K., Quivey, J. M., Weinberg, V. and Xia, P. (2006), 'Repeat CT imaging and replanning during the course of IMRT for head-and-neck cancer', *Int J Radiat Oncol Biol Phys* **64**(2), 355–362.

Harnsberger, H., Osborn, A. and Ross, J. (2006), *Diagnostic and surgical imaging anatomy: brain, head and neck, and spine.*, first edn, Lippincott Williams and Wilkins.

Hollinshead, H. (1974), *Textbook of anatomy*, third edn, Harper and Row.

Isambert, A., Dhermain, F., Bidault, F., Commowick, O., Bondiau, P., Malandain, G. and Lefkopoulos, D. (2008), 'Evaluation of an atlas-based automatic segmentation software for the delineation of brain organs at risk in a radiation therapy clinical context', *Radiother Oncol* **87**(1), 93–9.

Jaccard, P. (1908), 'Nouvelles recherches sur la distribution florale', *Bulletin de la Societe Vaudoise des Sciences Naturelles* **44**, 223–70.

Jackson, A., Marks, L. B., Bentzen, S. M., Eisbruch, A., Yorke, E. D., Ten Haken, R. K., Constine, L. S. and Deasy, J. O. (2010), 'The lessons of QUANTEC: recommendations for reporting and gathering data on dose-volume dependencies of treatment outcome', *Int J Radiat Oncol Biol Phys* **76**(3 Suppl), S155–160.

Jacobs, R. A. (1995), 'Methods for combining experts' probability assessments', *Neural Comput* **7**(5), 867–888.

Jaffray, D. A., Drake, D. G., Moreau, M., Martinez, A. A. and Wong, J. W. (1999), 'A radiographic and tomographic imaging system integrated into a medical linear accelerator for localization of bone and soft-tissue targets', *Int J Radiat Oncol Biol Phys* **45**(3), 773–789.

Jaffray, D. A., Siewerdsen, J. H., Wong, J. W. and Martinez, A. A. (2002), 'Flat-panel cone-beam computed tomography for image-guided radiation therapy', *Int J Radiat Oncol Biol Phys* **53**(5), 1337–1349.

Jemal, A., Simard, E. P., Dorell, C., Noone, A. M., Markowitz, L. E., Kohler, B., Eheman, C., Saraiya, M., Bandi, P., Saslow, D., Cronin, K. A., Watson, M., Schiffman, M., Henley, S. J., Schymura, M. J., Anderson, R. N., Yankey, D. and Edwards, B. K. (2013), 'Annual Report to the Nation on the Status of Cancer, 1975-2009, featuring the burden and trends in human papillomavirus(HPV)-associated cancers and HPV vaccination coverage levels', *J. Natl. Cancer Inst.* **105**(3), 175–201.

Jensen, A. D., Nill, S., Huber, P. E., Bendl, R., Debus, J. and Munter, M. W. (2012), 'A

clinical concept for interfractional adaptive radiation therapy in the treatment of head and neck cancer', *Int J Radiat Oncol Biol Phys* **82**(2), 590–596.

Kelley, K. (2005), 'The effects of nonnormal distributions on confidence intervals around the standardized mean difference: Bootstrap and parametric confidence intervals', *Educat Psychol Meas* **65**(1), 51–69.

Kittler, J., Hatef, M., Duin, R. and Matas, J. (1998), 'On combining classifiers', *IEEE Trans Pattern Anal Mach Intell* **20**(3), 226 –239.

La Macchia, M., Fellin, F., Amichetti, M., Cianchetti, M., Gianolini, S., Paola, V., Lomax, A. J. and Widesott, L. (2012), 'Systematic evaluation of three different commercial software solutions for automatic segmentation for adaptive therapy in head-and-neck, prostate and pleural cancer', *Radiat Oncol* **7**, 160.

Lagendijk, J., Raaymakers, B., Raaijmakers, A., Overweg, J., Brown, K., Kerkhof, E., van der Put, R., Hrdemark, B., van Vulpen, M. and van der Heide, U. (2008), 'MRI/linac integration', *Radiother Oncol* **86**(1), 25–9.

Lawrence, Y. R., Li, X. A., el Naqa, I., Hahn, C. A., Marks, L. B., Merchant, T. E. and Dicker, A. P. (2010), 'Radiation dose-volume effects in the brain', *Int J Radiat Oncol Biol Phys* **76**(3 Suppl), S20–27.

Lew, M. J. (2012), 'Bad statistical practice in pharmacology (and other basic biomedical disciplines): you probably don't know P', *Br J Pharmacol* **166**(5), 1559–1567.

Lu, W., Chen, M., Olivera, G., Ruchala, K. and Mackie, T. (2004), 'Fast free-form deformable registration via calculus of variations', *Phys Med Biol* **49**(14), 3067–87.

Lu, W., Olivera, G., Chen, Q., Ruchala, K., Haimerl, J., Meeks, S., Langen, K. and Kupelian, P. (2006), 'Deformable registration of the planning image (kVCT) and the daily images (MVCT) for adaptive radiation therapy', *Phys Med Biol* **51**(17), 4357–74.

Ludbrook, J. (2002), 'Statistical techniques for comparing measurers and methods of measurement: a critical review', *Clin Exp Pharmacol Physiol* **29**(7), 527–536.

Malsch, U., Thieke, C., Huber, P. and Bendl, R. (2006), 'An enhanced block matching algorithm for fast elastic registration in adaptive therapy', *Phys Med Biol* **51**, 4789–4806.

Martin, D., Fowlkes, C., Tal, D. and Malik, J. (2001), A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics, *in* 'Proc. 8th Int'l Conf. Computer Vision', Vol. 2, pp. 416–423.

Maurer, C., Rensheng, Q. and Raghavan, V. (2003), 'A linear time algorithm for computing exact Euclidean distance transforms of binary images in arbitrary dimensions', *IEEE Trans Pattern Anal Mach Intell* **25**(2), 265–70.

Mayo, C., Martel, M. K., Marks, L. B., Flickinger, J., Nam, J. and Kirkpatrick, J. (2010), 'Radiation dose-volume effects of optic nerves and chiasm', *Int J Radiat Oncol Biol Phys* **76**(3 Suppl), 28–35.

Mayo, C., Yorke, E. and Merchant, T. E. (2010), 'Radiation associated brainstem injury', *Int*

J Radiat Oncol Biol Phys **76**(3 Suppl), 36–41.

McGraw, K. and Wong, S. (1996), 'Forming inferences about some intraclass correlation coefficients', *Psychological Methods* **1**(1), 30–46.

Mell, L., Mehrotra, A. and Mundt, A. (2005), 'Intensity-modulated radiation therapy use in the U.S. 2004.', *Cancer* **104**, 1296–1303.

Mell, L., Roeske, J. and Mundt, A. (2003), 'A survery of intensity-modulated radiation therapy use in the United States.', *Cancer* **98**, 204–11.

Meyer, C., Johnson, T., McLennan, D., Aberle, D., Kazerooni, E., MacMahon, H., Mullan, B., Yankelevitz, D., van Beek, J., Armato III, S., McNitt-Gray, M., Reeves, A., Gur, D. and et al. (2006), 'Evaluation of lung MDCT nodule annotation across radiologists and methods', *Acad Radiol* **13**(10), 1254–65.

Molloy, J. A., Chan, G., Markovic, A., McNeeley, S., Pfeiffer, D., Salter, B. and Tome, W. A. (2011), 'Quality assurance of U.S.-guided external beam radiotherapy for prostate cancer: report of AAPM Task Group 154', *Med Phys* **38**(2), 857–871.

Mosleh-Shirazi, M. A., Evans, P. M., Swindell, W., Webb, S. and Partridge, M. (1998), 'A cone-beam megavoltage CT scanner for treatment verification in conformal radiotherapy', *Radiother Oncol* **48**(3), 319–328.

Nelms, B. E., Tome, W. A., Robinson, G. and Wheeler, J. (2012), 'Variations in the contouring of organs at risk: test case from a patient with oropharyngeal cancer', *Int J Radiat Oncol Biol Phys* **82**(1), 368–378.

Noble, J. and Dawant, B. (2009), Automatic segmentation of the optic nerves and chiasm in ct and mr using the atlas-navigated optimal medial axis and deformable-model algorithm, *in* 'Proceedings of the SPIE: Medical Imaging'.

Noble, J. H. and Dawant, B. M. (2011), 'An atlas-navigated optimal medial axis and deformable model algorithm (NOMAD) for the segmentation of the optic nerves and chiasm in MR and CT images', *Med Image Anal* **epub**, epub.

Noble, J., Warren, F., Labadie, R. and Dawant, B. (2008), 'Automatic segmentation of the facial nerve and chorda tympani in ct images using spatially dependent feature values.', *Med Phys* **35**, 5375–84.

Pasquier, D., Lacornerie, T., Vermandel, M., Rosseeau, J., Lartigau, E. and Betrouni, N. (2007), 'Automatic segmentation of pelvic structures from magnetic resonance images for prostate cancer radiotherapy', *Int J Radiat Oncol Biol Phys* **68**, 592–600.

Peroni, M., Ciardo, D., Spadea, M. F., Riboldi, M., Comi, S., Alterio, D., Baroni, G. and Orecchia, R. (2012), 'Automatic segmentation and online virtualCT in head-and-neck adaptive radiation therapy', *Int J Radiat Oncol Biol Phys* **84**(3), e427–433.

Pluim, J. P. and Fitzpatrick, J. M. (2003), 'Image registration', *IEEE Trans Med Imaging* **22**(11), 1341–1343.

Popovic, A., Fuente, M., Engelhardt, M. and Radermacher, K. (2007), 'Statistical validation

metric for accuracy assessment in medical image segmentation', *IJCARS* **2**, 169–181.
URL: *http://dx.doi.org/10.1007/s11548-007-0125-1*

Reed, V., Woodward, W., Zhang, L., Strom, E., Perkins, G., Tereffe, W., Oh, J., Yu, T., Bedrosian, I., Whitman, G., Bucholz, T. and Dong, L. (2008), 'Automatic segmentation of whole breat using atlas approach and deformable image registration', *Int J Radiat Oncol Biol Phys* **73**, 1493–500.

Rohde, G., Aldroubi, A. and Dawant, B. (2003), 'The adaptive bases algorithm for intensity-based nonrigid image registration', *IEEE Trans Med Imaging* **22**(11), 1470–9.

Schwartz, D. L. and Dong, L. (2011), 'Adaptive radiation therapy for head and neck cancer-can an old goal evolve into a new standard?', *J Oncol* **2011**.

Schwartz, D. L., Garden, A. S., Shah, S. J., Chronowski, G., Sejpal, S., Rosenthal, D. I., Chen, Y., Zhang, Y., Zhang, L., Wong, P. F., Garcia, J. A., Kian Ang, K. and Dong, L. (2013), 'Adaptive radiotherapy for head and neck cancer–dosimetric results from a prospective clinical trial', *Radiother Oncol* **106**(1), 80–84.

Shiu, A. S., Chang, E. L., Ye, J. S., Lii, M., Rhines, L. D., Mendel, E., Weinberg, J., Singh, S., Maor, M. H., Mohan, R. and Cox, J. D. (2003), 'Near simultaneous computed tomography image-guided stereotactic spinal radiotherapy: an emerging paradigm for achieving true stereotaxy', *Int J Radiat Oncol Biol Phys* **57**(3), 605–613.

Shrout, P. and Fleiss, J. (1979), 'Intraclass correlations - uses in assessing rater reliability', *Physcological Bulletin* **86**(2), 420–428.

Stapleford, L., Lawson, J., Perkins, C., Edelman, S., Davis, L., McDonald, M., Waller, A., Schreibmann, E. and Fox, T. (2010), 'Evaluation of automatic atlas-based lymph node segmentation for head-and-neck cancer.', *Int J Radiat Oncol Biol Phys* **77**, 959–66.

Studholme, C., Hill, D. and Hawkes, D. (1999), 'An overlap invariant entropy measure of 3D medical image alignment', *Pattern Recognit* **32**, 71–86.

Teguh, D. N., Levendag, P. C., Voet, P. W., Al-Mamgani, A., Han, X., Wolf, T. K., Hibbard, L. S., Nowak, P., Akhiat, H., Dirkx, M. L., Heijmen, B. J. and Hoogeman, M. S. (2011), 'Clinical validation of atlas-based auto-segmentation of multiple target volumes and normal tissue (swallowing/mastication) structures in the head and neck', *Int J Radiat Oncol Biol Phys* **81**(4), 950–957.

Tome, W. A., Meeks, S. L., Orton, N. P., Bouchet, L. G. and Bova, F. J. (2002), 'Commissioning and quality assurance of an optically guided three-dimensional ultrasound target localization system for radiotherapy', *Med Phys* **29**(8), 1781–1788.

Tsien, C., Moughan, J., Michalski, J. M., Gilbert, M. R., Purdy, J., Simpson, J., Kresel, J. J., Curran, W. J., Diaz, A. and Mehta, M. P. (2009), 'Phase I three-dimensional conformal radiation dose escalation study in newly diagnosed glioblastoma: Radiation Therapy Oncology Group Trial 98-03', *Int J Radiat Oncol Biol Phys* **73**(3), 699–708.

Tsuji, S. Y., Hwang, A., Weinberg, V., Yom, S. S., Quivey, J. M. and Xia, P. (2010), 'Dosimetric evaluation of automatic segmentation for adaptive IMRT for head-and-neck cancer', *Int J*

Radiat Oncol Biol Phys **77**(3), 707–714.

Tukey, J. W. (1977), *Exploratory Data Analysis*, Addison-Wesley.

van Vulpen, M., van der Heide, U. and Moorselaar, J. (2008), 'How quality influences the clinical outcome of external beam radiotherapy for localized prostate cancer', *BJU Int.* **101**(8), 944–7.

Voet, P. W., Dirkx, M. L., Teguh, D. N., Hoogeman, M. S., Levendag, P. C. and Heijmen, B. J. (2011), 'Does atlas-based autosegmentation of neck levels require subsequent manual contour editing to avoid risk of severe target underdosage? A dosimetric analysis', *Radiother Oncol* **98**(3), 373–377.

Warfield, S., Zou, K. and Wells, W. (2004), 'Simultaneous Truth and Performance Level Estimation (STAPLE): an algorithm for the validation of image segmentation', *IEEE Trans Med Imaging* **23**(7), 903–21.

Weber, A. L. and Sabates, N. R. (1996), 'Survey of CT and MR imaging of the orbit', *Eur J Radiol* **22**(1), 42–52.

Weiss, E., Wijesooriya, K., Ramakrishnan, V. and Keall, P. J. (2008), 'Comparison of intensity-modulated radiotherapy planning based on manual and automatically generated contours using deformable image registration in four-dimensional computed tomography of lung cancer patients', *Int J Radiat Oncol Biol Phys* **70**(2), 572–581.

Windridge, D. and Kittler, J. (2003), 'A morphologically optimal strategy for classifier combination: multiple expert fusion as a tomographic process', *IEEE Trans Pattern Anal Mach Intell* **25**(3), 343 – 353.

Wu, Z. (1995), 'Compactly supported positive definite radial functions', *Adv Comput Math* **4**, 283–92.

Xie, Y., Chao, M. and Xing, L. (2008), 'Feature-based rectal contour propagation from planning CT to cone beam CT', *Med Phys* **35**(10), 4450–9.

Yim, P., Cebral, J., Mullick, R., Marcos, H. and Choyke, P. (2001), 'Vessel surface reconstruction with a tubular deformable model', *IEEE Trans Med Imaging* **20**, 1411–21.

Zhang, T., Chi, Y., Meldolesi, E. and Yan, D. (2007), 'Automatic delineation of on-line head-and-neck computed tomography images: toward on-line adaptive radiotherapy', *Int J Radiat Oncol Biol Phys* **68**, 522–30.

Zhu, Y., Huang, X., Wang, W., Lopresti, D., Long, R., Antani, S., Xue, Z. and Thoma, G. (2008), 'Balancing the Role of Priors in Multi-Observer Segmentation Evaluation', *J Signal Process Syst* **55**(1-3), 185–207.

Zijdenbos, A., Dawant, B., Margolin, A. and Palmer, A. (1994), 'Morphometric analysis of white matter lesions in MR images: Method and validation', *IEEE Trans Med Imaging* **13**(4), 716–724.

Zou, K., Warfield, S., Bharatha, A., Tempany, C., Kaus, M., Haker, S., Wells III, W., Jolesz, F. and Kikinis, R. (2004), 'Statistical validation of image segmentation quality based on a spatial overlap index', *Acad Radiol* **11**(2), 178–89.