Optimising adaptive radiotherapy for head and neck cancer

A thesis submitted to the University of Manchester for the degree of
PhD
in the Faculty of Biology, Medicine and Health

2016

William John Beasley

School of Medical Sciences
Contents

1 Introduction .............................................. 15
  1.1 Overview ............................................. 15
  1.2 Head and neck cancer ............................... 15
  1.3 Radiotherapy for head and neck cancer .......... 17
  1.4 Anatomic and dosimetric changes ................. 20
    1.4.1 Target changes .................................. 21
    1.4.2 Normal structure changes ....................... 22
    1.4.3 Dosimetric consequences ....................... 26
  1.5 Adaptive radiotherapy ............................. 28
    1.5.1 ART strategies .................................. 31
    1.5.2 Predicting anatomic changes ................... 35
    1.5.3 Deformable image registration and automatic segmentation 38
  1.6 Predictive factors of radiotherapy toxicities .... 49
  1.7 Clinical benefit of ART ............................ 52
  1.8 Rationale and aims .................................. 55
  1.9 Thesis structure ..................................... 55

2 Relative plan robustness of IMRT and VMAT for weight loss 59
  2.1 Introduction .......................................... 61
  2.2 Method ............................................... 63
    2.2.1 Patient selection ................................ 63
    2.2.2 Treatment planning ............................... 63
    2.2.3 Dosimetric comparisons ........................... 65
  2.3 Results ............................................... 66
    2.3.1 Patient characteristics .......................... 66
    2.3.2 Dosimetric comparisons ........................... 66
  2.4 Discussion ............................................ 68
  2.5 Conclusion ............................................ 74

3 The suitability of common metrics for assessing parotid and
larynx automatic segmentation accuracy ............................ 75
  3.1 Introduction .......................................... 77
  3.2 Method ............................................... 79
    3.2.1 Dosimetric and geometric accuracy ............ 80
    3.2.2 Relationship between geometric and dosimetric accuracy . 82
3.3 Results .................................................. 82
   3.3.1 Dosimetric and geometric accuracy ............... 82
   3.3.2 Relationship between geometric and dosimetric accuracy 83
3.4 Discussion .............................................. 85
3.5 Conclusion .............................................. 90

4 An automated workflow for patient-specific quality control of contour propagation 91
   4.1 Introduction .......................................... 93
   4.2 Method ................................................ 94
      4.2.1 Automated workflow for contour propagation QC ...... 94
      4.2.2 Image and contour data ........................... 96
      4.2.3 Workflow validation: error scenario I ............. 96
      4.2.4 Workflow validation: error scenario II ............ 96
   4.3 Results ............................................... 98
      4.3.1 Propagation performance .......................... 98
      4.3.2 Workflow validation: error scenario I ............. 98
      4.3.3 Workflow validation: error scenario II ............ 99
   4.4 Discussion ............................................ 101
   4.5 Conclusion ............................................ 107

5 Permutation testing for image-based data mining 109
   5.1 Introduction .......................................... 112
   5.2 Method ................................................ 115
      5.2.1 Simulated data .................................... 115
         5.2.1.1 Test statistic ................................. 116
         5.2.1.2 Permutation testing .......................... 117
         5.2.1.3 Comparison of permutation testing techniques .. 120
      5.2.2 Application to radiation-induced trismus .......... 120
         5.2.2.1 Patient data .................................. 121
         5.2.2.2 Workflow ...................................... 121
   5.3 Results ............................................... 124
      5.3.1 Simulated data .................................... 124
      5.3.2 Application to radiation-induced trismus .......... 127
   5.4 Discussion ............................................ 132
   5.5 Conclusion ............................................ 139
6 Discussion

6.1 Relative robustness of IMRT and VMAT to weight loss during head and neck radiotherapy .......................... 142

6.2 The suitability of common metrics for assessing automatic segmentation accuracy ........................................ 144

6.3 An automated workflow for quality control of contour propagation 146

6.4 Permutation testing for image-based data mining ..................... 149

7 Conclusion ............................................. 155

Word count: ~40,000
List of Figures

1.1 TCP and NTCP .................................................. 18
1.2 Head and neck VMAT treatment plan ........................... 19
1.3 Anatomic changes ............................................. 21
1.4 Parotid and submandibular gland anatomic displacements ... 24
1.5 An offline ART workflow ....................................... 29
1.6 Deformable image registration .................................. 39
1.7 Atlas-based automatic segmentation ............................ 41
1.8 Dice similarity coefficient ...................................... 44
1.9 Distance-to-agreement .......................................... 45
3.1 DSC and DTA .................................................. 78
3.2 Dosimetric inter-observer variation for the parotids .......... 83
3.3 Dosimetric inter-observer variation for the larynx ........... 84
3.4 Scatter plots showing the relationship between dosimetric and geometric accuracy ........................................... 86
4.1 Illustration of the contour QC workflow ......................... 95
4.2 Plot of the consistency metrics for error scenario I .......... 99
4.3 Variation of propagation accuracy with increasing noise .... 100
4.4 ROC curves for the logistic regression model .................. 102
5.1 Simulated data ................................................ 116
5.2 Image-based data mining for continuous variables .......... 117
5.3 Permutation testing ............................................ 120
5.4 Methodology for for image-based data mining ................. 123
5.5 Permutation testing for increasing N .......................... 125
5.6 Permutation testing for increasing outcome data noise ...... 126
5.7 Choice of threshold for STCS_{max} ............................ 127
List of Tables

1.1 Summary of anatomic changes ........................................ 23
1.2 Summary of dosimetric changes .............................. 27
1.3 Overview of automatic segmentation performance ............... 47
2.1 OAR dose constraints ............................................. 65
2.2 Patient, tumour and treatment characteristics ................. 67
2.3 Changes in clinical target volumes ............................. 68
2.4 Changes in target volume coverage ................................ 69
2.5 Differences in the changes to target volume coverage ............ 70
2.6 Changes in OAR doses ............................................ 71
2.7 Differences in the changes to OAR doses ......................... 72
3.1 OAR dose constraints ............................................ 81
3.2 Correlation coefficients for the different metrics ............... 85
4.1 Mean accuracy of the propagated structures ..................... 98
5.1 Summary of the patients included in the analysis. ............... 122
5.2 Corrected p-values and the minimum and maximum correlation coefficients ......................... 129
# Nomenclature

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3DCRT</td>
<td>Three-dimensional conformal radiotherapy</td>
</tr>
<tr>
<td>ART</td>
<td>Adaptive radiotherapy</td>
</tr>
<tr>
<td>BED</td>
<td>Biologically effective dose</td>
</tr>
<tr>
<td>CBCT</td>
<td>Cone beam computed tomography</td>
</tr>
<tr>
<td>CT</td>
<td>Computed tomography</td>
</tr>
<tr>
<td>CTCAE</td>
<td>Common terminology criteria for adverse events</td>
</tr>
<tr>
<td>CTV</td>
<td>Clinical target volume</td>
</tr>
<tr>
<td>DIR</td>
<td>Deformable image registration</td>
</tr>
<tr>
<td>DSC</td>
<td>Dice similarity coefficient</td>
</tr>
<tr>
<td>DTA</td>
<td>Distance-to-agreement</td>
</tr>
<tr>
<td>GTV</td>
<td>Gross tumour volume</td>
</tr>
<tr>
<td>HU</td>
<td>Hounsfield unit</td>
</tr>
<tr>
<td>IGRT</td>
<td>Image guided radiotherapy</td>
</tr>
<tr>
<td>IMRT</td>
<td>Intensity modulated radiotherapy</td>
</tr>
<tr>
<td>LKB</td>
<td>Lyman-Kutcher-Burman</td>
</tr>
<tr>
<td>MI</td>
<td>Mutual information</td>
</tr>
<tr>
<td>MID</td>
<td>Maximum incisor-to-incisor distance</td>
</tr>
<tr>
<td>MLC</td>
<td>Multileaf collimator</td>
</tr>
<tr>
<td>MRI</td>
<td>Magnetic resonance imaging</td>
</tr>
<tr>
<td>NCC</td>
<td>Normalised cross correlation</td>
</tr>
<tr>
<td>NTCP</td>
<td>Normal tissue complication probability</td>
</tr>
<tr>
<td>OAR</td>
<td>Organ at risk</td>
</tr>
<tr>
<td>PET</td>
<td>Positron emission tomography</td>
</tr>
<tr>
<td>Acronym</td>
<td>Description</td>
</tr>
<tr>
<td>---------</td>
<td>-------------</td>
</tr>
<tr>
<td>PTV</td>
<td>Planning target volume</td>
</tr>
<tr>
<td>QC</td>
<td>Quality control</td>
</tr>
<tr>
<td>SMG</td>
<td>Submandibular gland</td>
</tr>
<tr>
<td>SSD</td>
<td>Sum of squared differences</td>
</tr>
<tr>
<td>TCP</td>
<td>Tumour control probability</td>
</tr>
<tr>
<td>VMAT</td>
<td>Volumetric modulated arc therapy</td>
</tr>
</tbody>
</table>
Abstract

Anatomic changes occur throughout head and neck radiotherapy, and a new treatment plan is often required to mitigate the resulting changes in delivered dose to key structures. This process is known as adaptive radiotherapy (ART), and can be labour-intensive. The aim of this thesis is to optimise ART, addressing some of the technical and clinical challenges facing its routine clinical implementation.

Optimising the frequency and timing of adaptive replanning is important, and it has been shown here that intensity modulated radiotherapy (IMRT) and volumetric modulated arc therapy (VMAT) are equally robust to weight loss during head and neck radiotherapy. Plan adaptation strategies that have previously been developed for IMRT are therefore applicable to VMAT.

Contour propagation is an important component of ART, and it is essential to ensure that propagated contours are accurate. A method for assessing the suitability of a metric for measuring automatic segmentation accuracy has been developed and applied to the head and neck. For the parotids and larynx, metrics based on surface agreement were better than the commonly used Dice similarity coefficient. By establishing a consensus on which metrics should be used to assess segmentation accuracy, comparison of different algorithms is more objective and should lead to more accurate automatic segmentation.

A novel method of assessing contour propagation accuracy on a patient-specific basis has also been developed. This was demonstrated on a cohort of head and neck patients and shows potential as a tool for identifying propagated contours that are subject to a high degree of uncertainty. This is a novel tool that will increase the efficiency of automatic segmentation and, therefore, ART.

Optimum ART requires consideration of different radiotherapy-related toxicities, and image-based data mining is a powerful technique for spatially localising dose-response relationships. Correction for multiple comparisons through permutation testing is essential, but has so far only been applied to categorical data. A novel method has been developed for performing permutation testing and image-based data mining with a continuously variable clinical endpoint. Application to trismus for head and neck radiotherapy identified a region with a dose-response relationship in the ipsilateral masseter. Sparing this structure during radiotherapy should reduce the severity of radiation-induced trismus.

ART mitigates the dosimetric effects of anatomic changes, and this thesis has addressed technical and clinical challenges that have so far limited its clinical implementation. Detailed knowledge of dose-response relationships will enable selection of patients for ART based on potential clinical benefit, and accurate contour propagation will make ART more efficient, facilitating its routine implementation.

The University of Manchester
William John Beasley
Doctor of Philosophy
Optimising adaptive radiotherapy for head and neck cancer
December 2016
Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=24420), in any relevant Thesis restriction declarations deposited in the University Library, The University Library’s regulations (see http://www.manchester.ac.uk/library/aboutus/regulations) and in The University’s policy on Presentation of Theses.
Acknowledgements

I would like to thank my supervisors, Marcel van Herk, Alan McWilliam, Ranald Mackay and Nick Slevin for their support and guidance throughout my PhD. I am extremely grateful for the advice and opportunities they have given me. Thanks also to my advisor, Tim Illidge, for the helpful chats and advice, and to Carl Rowbottom and Chris Boylan for their support at the start of my PhD. I would like to thank my colleagues in the Radiotherapy Department, and CMPE for funding my PhD and providing an environment for me to progress with my research.

I have received support from several clinicians throughout my research, and I would like to extend my thanks in particular to the head and neck team. Many thanks also to Adam Aitkenhead for his help with Matlab, and Andrew Green for his speedy updates to the MIP toolkit.

I would like to extend my gratitude to the Physics Department at Memorial Sloan Kettering Cancer Centre in New York, where I performed much of the work for my final project. In particular, many thanks to Maria Thor for helping to set up the collaboration and for making me feel welcome whilst I was in New York.

Finally, I would like to thank my parents for their support and guidance over the last 30 years, and Amy for her continuing encouragement and patience.

The author

- HCPC-registered Clinical Scientist since October 2013
- MSc Physics and Computing in Medicine and Biology, University of Manchester, 2010
- MPhys Physics, University of Warwick, 2009
Alternative format

Permission has been obtained for submission of this thesis in alternative format, in which each chapter is written in the style of a scientific publication. Justification for submission in alternative format is provided in Section 1.9. Several of the chapters included in this thesis have been published, and a list of publications is provided below.


1 Introduction

1.1 Overview

Radiotherapy for head and neck cancer is typically delivered in daily fractions over the course of up to seven weeks. Treatment is planned such that the tumour receives the prescription dose, with the dose to nearby sensitive structures reduced as much as reasonably possible. This requires steep dose gradients, making the treatment plan extremely sensitive to any changes that may occur during treatment. This treatment planning is performed on computed tomography (CT) images obtained before treatment, and anatomic changes that often occur during the course of treatment mean that the original treatment plan is not always delivered optimally. A new treatment plan is often required to mitigate the dosimetric changes that result from these anatomic variations, and this process is referred to as adaptive radiotherapy (ART). ART can be extremely time-consuming, which has prevented it from being routinely adopted, and several technical challenges remain before ART can be routinely implemented.

The overall aim of this thesis is to optimise ART, addressing some of the technical and clinical challenges that have so far prevented routine clinical implementation of ART. In this introduction, the background to radiotherapy for head and neck cancer is first discussed, and a review of the relevant literature on ART for head and neck cancer is provided. Finally, the specific aims of the thesis are introduced.

1.2 Head and neck cancer

Head and neck cancer is a group of cancers that comprise several different sites within the head and neck, including the oropharynx, hypopharynx, nasopharynx, larynx, oral cavity, salivary glands, sinuses and lip. In 2011, 356,860 people were
diagnosed with cancer, of which 11,449 diagnoses were for head and neck cancer (Cancer Research UK, 2014). The mean 5-year survival rate is dependent on the specific subsite; for example, it is 46 % for cancers of the tongue, 47 % for oral cavity, 39 % for the oropharynx, 20 % for the hypopharynx, and 65 % for the larynx, calculated from patients diagnosed with head and neck cancer between 1995 and 1999 (Zigon et al., 2011). Treatment depends on the site and stage of disease, but typically consists of a combination of surgery, radiotherapy, and systemic therapy.

Side effects of radiotherapy for head and neck cancer consist of acute toxicities, in which symptoms can be present for up to three months post-radiotherapy, and late toxicities, which tend to persist several months or years after the completion of radiotherapy. These side effects can result in significant deterioration in quality of life (Kelly et al., 2007), and so their management and prevention is important for radiotherapy.

One of the main late toxicities after head and neck radiotherapy is xerostomia, in which salivary flow is reduced. This can affect speech and swallowing and increases the risk of dental caries, which results in significant degradation of quality of life (Wijers et al., 2002). The incidence of radiotherapy-related xerostomia varies depending on the specific radiotherapy technique used and the dose delivered to the parotid glands. For example, in a trial investigating the benefit of minimising the dose to the parotid glands during radiotherapy, the incidence of xerostomia was reduced from 74 % when the parotids were not spared to 40 % when the dose to the parotids was minimised (Nutting et al., 2011).

Other side effects relating to head and neck radiotherapy include dysphagia, trismus, osteoradionecrosis of the jaws, hearing loss and skin fibrosis. Dysphagia is an impairment to the ability to swallow, and is a significant late complication that can severely affect a patient’s quality of life (Denaro et al., 2013). The incidence of severe dysphagia that results in feeding tube dependency varies, with
rates between 12 % and 50 % (Bhide et al., 2012) reported. Trismus is another late side effect of head and neck radiotherapy that affects between 5 % and 38 % of patients and is characterised by reduced mouth opening (Dijkstra et al., 2004). It progresses rapidly from 1 to 9 months after radiotherapy, with a slower rate of progression in later years (Wang et al., 2005). Although less common, other late side effects such as osteoradionecrosis and hearing loss can also have a significant effect on quality of life (Bhide et al., 2012).

1.3 Radiotherapy for head and neck cancer

The aim of radiotherapy is to deliver an homogeneous radiation dose to the target whilst sparing sensitive normal tissue. The tumour control probability (TCP) and normal tissue complication probability (NTCP) are often used to describe the probability of achieving local control and normal tissue toxicity for a given dose, respectively, as illustrated in Figure 1.1. In the figure, TCP is shown by the solid line, and NTCP is shown by the dotted line. Also illustrated by the green shaded region is the therapeutic window, which highlights a range of doses for which there is an acceptable balance between the probability of cure and the risk of excessive toxicity. Increasing the dose beyond this range will result in improved local control, but at the cost of increased risk of normal tissue complications, and it is this risk of toxicity that limits the dose that can be delivered to a tumour. Conformal treatment plans that deliver a high dose to the target whilst sparing healthy tissue are therefore required to obtain the highest probability of cure.

The number of healthy structures, or organs at risk (OARs) in the head and neck, and their proximity to typical disease sites, means that complex treatment
Figure 1.1: Tumour control probability (TCP) and normal tissue complication probability (NTCP) as a function of dose. TCP is shown by the solid line, NTCP is shown by the dotted line, and the therapeutic window is illustrated by the green shaded region.

Plans are required, with steep dose gradients to provide sparing for different sensitive structures (see Figure 1.2 for an example of a typical head and neck treatment plan). This is achieved through intensity modulated radiotherapy (IMRT), in which non-uniform beam intensities from different gantry angles are optimised inversely to build the desired dose distribution. Compared to three-dimensional conformal radiotherapy (3DCRT), IMRT has led to improved target coverage and OAR sparing (Eisbruch et al., 1998; van Dieren et al., 2000; Xia et al., 2000). An extension of IMRT, volumetric modulated arc therapy (VMAT) is also capable of producing highly conformal treatment plans (Scorsetti et al., 2010). Similarly to IMRT, beam intensities are inversely optimised to create the desired treatment plan, but for VMAT the radiation source rotates continuously around the patient during treatment, with simultaneous variation of the multileaf collimator (MLC) positions, dose rate and gantry speed. VMAT is capable of producing treatment plans of comparable or better quality to IMRT, and is more efficient in terms of delivery time (Guckenberger et al., 2009; Bertelsen et al., 2010; Dai et al., 2015).
Both IMRT and VMAT can produce extremely conformal treatment plans that spare OARs whilst maintaining target coverage, and this is particularly useful for sparing the parotid glands, which are often in close proximity to the target and are responsible for salivary function. Indeed, the PARSPORT trial, which compared conformal radiotherapy and parotid-sparing IMRT, showed that IMRT resulted in a reduced incidence of xerostomia (Nutting et al., 2011). However, the steep dose gradients resulting from head and neck IMRT are particularly sensitive to positional errors and anatomic changes, meaning that small changes in position can result in large changes in dose to the tumour and OARs.

A radiotherapy treatment plan is traditionally based on a single CT scan obtained prior to treatment, and the dose distribution is tailored to conform to the anatomic structures delineated on this scan. Geometric uncertainties,
consisting of random and systematic components, arise as the patient is set up for
treatment at each treatment fraction. Random uncertainties effectively blur the
dose distribution and systematic uncertainties shift the dose distribution relative
to the target (van Herk et al., 2000). To ensure that the clinical target volume
(CTV), which is the tumour plus a margin for microscopic tumour spread, is
treated adequately, a geometric margin is added to the CTV, making the planning
target volume (PTV). This margin ensures that the CTV is adequately treated
during treatment, but also limits the degree to which nearby OARs can be spared.

The size of this margin can be reduced by performing suitable image guidance,
in which images of the patient in the treatment position are used to ensure that
the tumour is in the correct position relative to the radiation beam (Jaffray,
2005). Adjustments of the treatment couch can then be performed to ensure that
the patient is correctly aligned with the radiation isocentre. This is known as
image guided radiotherapy (IGRT).

1.4 Anatomic and dosimetric changes

Regular imaging in the treatment position enables the accuracy and precision of
radiotherapy to be increased. Rigid couch shifts based on the position of these on-
treatment images relative to the original CT scan enable a reduction of random
and systematic positioning errors that occur during treatment. However, changes
in size, shape or relative position of the target or OARs since the initial planning
CT cannot be corrected by rigid couch shifts, limiting the benefits afforded by
IGRT. Head and neck radiotherapy typically lasts for up to seven weeks, and
anatomic changes can occur, affecting the delivered dose to targets and OARs.
In this section, a review of the literature on anatomic and dosimetric changes
during head and neck radiotherapy is provided, and is summarised in Table 1.1.
Figure 1.3: Example of anatomic changes. Compared to the planning CT (a), a rescan mid-way through treatment (b) highlights weight loss and changes to the parotids. The external contour is shown by the red line and the parotids by the blue line.

1.4.1 Target changes

Several studies have shown that the gross tumour volume (GTV) and CTV reduce in volume throughout radiotherapy. Barker et al. (2004) studied changes in GTV over the course of radiotherapy by obtaining three CT scans per week for 14 patients. They measured a median volume loss of 0.2 cc (1.8\%) per treatment day, resulting in a median total volume loss of 70\% of the initial tumour volume. They noted that tumour regression was asymmetric and that the total volume loss was correlated with the initial tumour volume. Conversely, Yan et al. (2013) did not find a correlation between initial tumour volume and volume loss in their study of 20 patients. They did report, however, that CTV volume loss was correlated with pre-treatment body mass index. Bhide et al. (2010) also investigated anatomic changes during therapy in their study of 20 patients receiving weekly CT. They reported that the most significant changes occurred between the start of treatment
and week 2, with CTVs for macroscopic and microscopic disease losing 3.2 % and 10 % of their initial volumes, respectively.

Although the GTV may shrink throughout the course of therapy, it should be noted that it is not necessarily desirable to redefine the GTV during treatment (Hansen et al., 2006; Hamming-Vrieze et al., 2017), especially when considering that its delineation is often influenced by other imaging modalities, such as magnetic resonance imaging (MRI) or positron emission tomography (PET).

1.4.2 Normal structure changes

Weight loss during head and neck radiotherapy is common, with patients often losing between 5 % and 15 % of their initial weight (Barker et al., 2004; Bhide et al., 2010; Broggi et al., 2010; Yan et al., 2013). Barker et al. (2004) reported that this weight loss correlates with the skin contour and volume at the level of the C2 vertebral body, indicating that weight loss is associated with tissue loss in the neck. They also measured volumetric changes in the parotid using 3-weekly CT imaging and reported a daily volume loss of 0.6 %, resulting in an overall volume loss of 28.1 % by the end of treatment. Similarly, Bhide et al. (2010) measured weekly parotid volume changes in 20 patients and reported an overall volume loss of 35 %, with the largest change occurring between the start of treatment and week 2, by which time the parotids had reduced in volume by 15 %. A similar finding was reported by Lee et al. (2008b), who assessed 330 parotid structures obtained from daily megavoltage CT imaging of 10 patients. They reported that the rate of volume loss was greater during the early stages of treatment, and that a hyperbolic function could describe the rate of volume loss better than a linear function.

Whilst these studies grouped the parotids together, or considered the left and
Table 1.1: Summary of anatomic changes throughout treatment for head and neck radiotherapy.

<table>
<thead>
<tr>
<th>Reference</th>
<th>N</th>
<th>Imaging</th>
<th>Anatomic changes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ho et al. (2012)</td>
<td>10</td>
<td>Weekly CBCT</td>
<td>Parotid: V↓25 %</td>
</tr>
<tr>
<td>Lee et al. (2008b)</td>
<td>10</td>
<td>Daily MVCT</td>
<td>Parotid: V↓21 % (0.7 %/day); 2.6 mm medial shift</td>
</tr>
<tr>
<td>Barker et al. (2004)</td>
<td>14</td>
<td>3 CT per week</td>
<td>Parotid: V↓28 % (0.6 %/day); 3.1 mm medial shift \ GTV: V↓70 % (1.8 %/day)</td>
</tr>
<tr>
<td>Robar et al. (2007)</td>
<td>15</td>
<td>Weekly CT</td>
<td>Parotid: V↓4.9 %/week; 0.85 mm/week medial shift</td>
</tr>
<tr>
<td>Hunter et al. (2013)</td>
<td>18</td>
<td>Daily CBCT</td>
<td>Parotid: V↓13 %</td>
</tr>
<tr>
<td>Jin et al. (2013)</td>
<td>10</td>
<td>Weekly CBCT</td>
<td>Parotid: 4.5 % - 4.7 %/week</td>
</tr>
<tr>
<td>Bhide et al. (2010)</td>
<td>20</td>
<td>Weekly CT</td>
<td>Parotid: V↓ 35 %; 2.3 mm medial shift \ CTV1: V↓3.2 % (week 2) \ CTV2: V↓10 % (week 2)</td>
</tr>
<tr>
<td>Wu et al. (2009)</td>
<td>11</td>
<td>Weekly CT</td>
<td>Parotid: V↓15 %; CTV: V↓10 %</td>
</tr>
<tr>
<td>Nishi et al., 2013</td>
<td>20</td>
<td>Repeat CT</td>
<td>GTV: V↓63 % (primary); V↓52 % (nodal) \ Parotid: V↓18 %; 4.2 mm medial shift</td>
</tr>
<tr>
<td>Vásquez Osorio et al. (2008)</td>
<td>10</td>
<td>Repeat CT</td>
<td>Parotid: V↓17 % (ipsi); V↓5 % (contra) \ SMG: V↓20 % (ipsi); V↓11 % (contra)</td>
</tr>
</tbody>
</table>

**Abbreviations:** V = volume; contra = contralateral; ipsi = ipsilateral.
right glands separately, some studies have instead grouped parotid glands by their proximity to the primary treatment volume. Vásquez Osorio et al. (2008) separated the parotids into ‘irradiated’ glands, belonging to the treated neck, and ‘spared’ glands, belonging to the untreated neck. They reported that the volume loss of the irradiated glands, at $17\% \pm 7\%$ was greater than that of the spared glands ($5\% \pm 4\%$). Other authors have reported similar findings for volume loss of the parotids during radiotherapy, some of which are summarised in Table 1.1.

The position of the parotid has also been shown to change during radiotherapy. Barker et al. (2004), Lee et al. (2008b) and Bhide et al. (2010) measured the positional change of the parotid during treatment, and reported that the centre of mass moved between 2 mm and 3 mm medially. These studies noted a shift in the centre of mass of the parotids, but Robar et al. (2007) reported that it is the lateral portion of the parotid that moves medially. They performed weekly CT imaging on 15 consecutive head and neck patients and placed a landmark at a repeatable medial and lateral position on each image. They observed that the medial landmark remained stationary and that the lateral landmark moved medially by 2.6 mm and 1.9 mm for the left and right parotid, respectively. In addition, Castelli et al. (2015) reported that the mean distance between the primary CTV and parotid glands, measured using the minimum distance between the two
delineated structures, reduced by 4.3 mm for 74 % of the parotid glands included in their analysis.

A more complex method was used by Vásquez Osorio et al. (2008), who outlined the parotid glands on a repeat CT image taken two weeks after completion of radiotherapy. They used deformable image registration to generate a deformation map of the parotid contours and found that the medial and lateral regions of the parotids moved medially by 3 mm and 1 mm, respectively. This is shown in Figure 1.4a, which is adapted from Vásquez Osorio et al. (2008).

Whilst the majority of the literature has focussed on geometric changes in the parotid, some other structures have also been studied. Vásquez Osorio et al. (2008) also measured anatomic changes in the submandibular glands (SMG) during radiotherapy. They reported that the volume of the SMGs reduced during treatment, with the irradiated and spared glands reducing by 20 % ± 10 % and 11 % ± 7 %, respectively. Similar to their analysis of the parotids, they reported a trend for the SMGs to move superiorly and posteriorly during treatment, as illustrated in Figure 1.4b. Wang et al. (2009) also studied anatomic changes in the SMG, and reported a volume loss of 16.8 % by the end of treatment. Furthermore, they found that volume loss for the SMGs, as well as the parotids, was largest in the first three weeks of treatment.

Studies on geometric changes considering other ROIs in the head and neck are relatively fewer than those on the parotids. Ricchetti et al. (2011) studied the volumetric changes of different structures of up to 16 patients during the course of radiotherapy. They reported reductions in volume by the end of treatment for the larynx (16 % ± 10 %), constrictor muscles (17 % ± 19 %), and masticatory muscles (7 % ± 7 %). Structures such as the spinal cord and brainstem do not tend to change volume during treatment (Jin et al., 2013).
1.4.3 Dosimetric consequences

Anatomic changes over the course of radiotherapy can induce dosimetric changes, and Table 1.2 summarises some of the observations from the literature. Dosimetric coverage of the target volumes tends to be robust during radiotherapy. For example, Wu et al. (2009) investigated the anatomic and dosimetric changes to several structures during head and neck radiotherapy for 11 patients who underwent weekly CT scanning. They reported no change in the delivered dose to the primary CTV, with small a small increase in the minimum dose delivered to the nodal CTV, likely caused by the larger volume and anatomic changes experienced by the nodal CTV. Similarly, Nishi et al. (2013) also reported a slight increase in dose, albeit to the primary GTV, as a result of anatomic changes in their study of 20 patients who underwent a repeat CT scan part-way through treatment. They reported no changes in the minimum delivered dose to the nodal GTV. Castadot et al. (2011) also investigated the impact of anatomic changes on target coverage in their study of ten patients who underwent repeat CT scanning during treatment. They reported that the dose to the primary and nodal CTVs remained unchanged as a result of anatomic changes throughout radiotherapy.

Although the dosimetric changes to the target volumes tend to be small, the volume reduction and medial migration of the parotids tends to result in an increased cumulative dose to these structures. For example, Wu et al. (2009) investigated the anatomic and dosimetric changes to several structures during head and neck radiotherapy for 11 patients who underwent weekly CT scanning. They reported a volume reduction of approximately 15 %, which is in line with the observations of other authors (Table 1.1). This resulted in an increase in cumulative dose of approximately 10 % to the parotids. Castelli et al. (2015) investigated the parotid gland overdose resulting from anatomic changes in 15
<table>
<thead>
<tr>
<th>Reference</th>
<th>N</th>
<th>Imaging</th>
<th>Dosimetric changes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ho et al. (2012)</td>
<td>10</td>
<td>Weekly CBCT</td>
<td><strong>Parotid, SC, BS</strong>: no change</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>Larynx, OC</strong>: no change</td>
</tr>
<tr>
<td>Robar et al. (2007)</td>
<td>15</td>
<td>Weekly CT</td>
<td><strong>Parotid</strong>: $D_{\text{mean}} \uparrow 2.6% ; (L)$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>SC</strong>: $\uparrow 0.2%$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>BS</strong>: $\uparrow 1.0%$</td>
</tr>
<tr>
<td>Hunter et al. (2013)</td>
<td>18</td>
<td>Daily CBCT</td>
<td><strong>Parotid</strong>: $D_{\text{mean}} \uparrow 0.9; \text{Gy}$</td>
</tr>
<tr>
<td>Jin et al. (2013)</td>
<td>10</td>
<td>Weekly CBCT</td>
<td><strong>Parotid</strong>: $V_{26\text{%}} \uparrow 7.5% ; (R)$; $V_{26\text{%}} \uparrow 8.8% ; (L)$</td>
</tr>
<tr>
<td>Bhide et al. (2010)</td>
<td>20</td>
<td>Weekly CT</td>
<td><strong>Parotid</strong>: $D_{\text{mean}} \uparrow 7% ; \text{(ipsi)}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>SC</strong>: $D_{\text{max}} \uparrow 2%$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>BS</strong>: $D_{\text{max}} \uparrow 4%$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>PTV1</strong>: $D_{\text{min}} \downarrow 3%$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>PTV2</strong>: $D_{\text{min}} \downarrow 5%$</td>
</tr>
<tr>
<td>Wu et al. (2009)</td>
<td>11</td>
<td>Weekly CT</td>
<td><strong>Parotid</strong>: $D_{\text{mean}} \uparrow 10%$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>CTV, BS, SC</strong>: No change</td>
</tr>
<tr>
<td>Castadot et al. (2011)</td>
<td>10</td>
<td>4 rpt CTs</td>
<td><strong>Parotid</strong>: $D_{\text{mean}} \uparrow 4%$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>SC</strong>: $D_{\text{2\text{cm}}} \uparrow 4.5%$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>CTV</strong>: No change</td>
</tr>
<tr>
<td>Nishi et al. (2013)</td>
<td>20</td>
<td>Rpt CT</td>
<td><strong>GTV</strong>: $D_{98\text{%}} \uparrow 1% ; \text{(primary)}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>**D_{98\text{%}} \uparrow 0.3% ; \text{(nodal)}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>Parotid</strong>: $D_{\text{mean}} \uparrow 20%$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>SC</strong>: $D_{2\text{cm}} \uparrow 5%$</td>
</tr>
<tr>
<td>Castelli et al. (2015)</td>
<td>15</td>
<td>Weekly CT</td>
<td><strong>Parotid</strong>: $D_{\text{mean}} \uparrow 3.7; \text{Gy}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(59% parotids)</td>
</tr>
</tbody>
</table>

**Abbreviations**: SC = spinal cord; BS = brainstem; L = left; R = right; V = volume; ipsi = ipsilateral.
patients. They reported that 59% of parotid glands were over-irradiated by a mean dose of 4 Gy (up to 10 Gy).

This is in contrast to Ho et al. (2012), who investigated 10 consecutive oropharyngeal cancer patients treated over 6 weeks and imaged with weekly cone beam CT (CBCT) imaging. Although they reported significant volume changes to both parotids of at least 25%, they found no significant changes to the cumulative parotid dose as a result of the observed anatomic changes, even in patients who lost more than 10% of their initial weight. They also failed to identify any statistically significant changes in the maximum dose to the spinal cord or brainstem. Robar et al. (2007) found in their study of 15 consecutive patients that the change in mean dose to the parotids was just 2.6%, and that the mean change in maximum dose to the spinal cord and brainstem was <2%. However, they reported large deviations in dose for some patients.

Although there are inconsistencies in the magnitude of dosimetric changes, taken together, the literature suggests that the cumulative delivered dose to the parotid glands tends to be greater than planned, and that this increase is caused by anatomic changes experienced during radiotherapy. The magnitude of the dosimetric changes depend on local dose gradients and the proximity to the primary target volume, and could explain some of the variations in the reported cumulative doses in the literature, which tend to use small sample sizes.

1.5 Adaptive radiotherapy

The anatomic changes that often occur during head and neck radiotherapy lead to progressive dosimetric changes such that a treatment planned solely on a CT obtained before treatment is sub-optimal. IGRT, in which rigid couch shifts are applied to align the position of the radiation isocentre with that intended, cannot
Figure 1.5: An offline adaptive radiotherapy workflow. The dose distribution is calculated on an on-treatment image and compared to the planned dose distribution; if necessary, a new treatment plan is produced.

fully account for these changes due their non-rigid nature (Schwartz, 2012). Adaptive radiotherapy is a formal process of maintaining the intended plan quality throughout the course of treatment, by creating a new treatment plan in response to anatomic and dosimetric changes. First introduced almost 20 years ago by Yan et al. (1997), ART has not been widely implemented due to its technically challenging nature (Castadot et al., 2011; Zhao et al., 2011).

ART requires a formal approach to handling changes that occur during treatment, and a general strategy is illustrated in Figure 1.5. The process starts with deformable image registration (DIR) between the planning CT and the on-treatment image, resulting in a deformation vector field (DVF), mapping the spatial correspondence between equivalent voxels in the two images. The DVF is used to propagate structures from the planning CT onto the on-treatment image,
and the dose distribution for the new patient anatomy is calculated. The inverse
DVF can then be used to warp the fractional dose into the anatomy of the pCT,
summing the fractional doses in a common frame of reference. This cumulative
dose distribution or the fractional dose distribution is then compared to the
planned dose distribution, and if they differ by some pre-determined amount that
is deemed to be clinically significant, then a new treatment plan is implemented.

On-treatment imaging is an important component of ART. It is used for computing the dose distribution at a particular treatment fraction, and for DIR between the on-treatment image and the pCT to enable automatic segmentation and dose accumulation. CBCT imaging is often used for on-treatment imaging, but it is subject to poor image quality due to increased scatter, as well as image artefacts (Siewerdsen and Jaffray, 2001). There has been much interest in improving CBCT image quality, which is beyond the scope of this thesis, but reviews of scatter correction techniques have previously been published (Rührnschopf and Klingenberg, 2011; Rührnschopf and and Klingenberg, 2011). Improving CBCT image quality should enable accurate Hounsfield Unit (HU) calibration and, therefore, dose calculation directly on CBCT images. An alternative is to use DIR to warp the pCT into the anatomy represented by the CBCT, directly mapping the HU from the pCT into the new anatomy observed at a specific treatment fraction, enabling dose calculation at that fraction (Veiga et al., 2014). Although the accuracy of this technique will depend strongly on the accuracy of the DIR algorithm, this approach shows promise for enabling accurate dose calculation based on updated patient anatomy.

The general ART workflow represented in Figure 1.5 can be implemented either online, for which changes are observed at a particular fraction and implemented before treatment of that fraction, or offline, where an adapted plan is created between fractions, combining multiple on-treatment images. This section summarises some of the approaches reported in the literature for head and neck
ART.

1.5.1 ART strategies

ART strategies can be performed in three timescales: offline, in which a new treatment plan is implemented for future fractions based on images from the previous fraction; online, in which a new treatment plan is implemented based on images from the current fraction; and real-time, in which a treatment plan is continually updated based on images obtained simultaneously with treatment. Some of the different approaches suggested in the literature are discussed below.

An online ART strategy for head and neck radiotherapy was suggested by Ahunbay et al. (2009). Based on an aperture-morphing technique for IMRT, a new plan is created at each fraction, accounting for anatomic changes as well as random non-rigid setup errors that cannot be mitigated with IGRT. They performed a retrospective planning study for five patients and reported severe under-dosage of the PTV and over-dosage of the spinal cord with IGRT, which was removed with their online ART technique. It should be noted, however that their analysis considered dosimetric changes to the PTV; they did not report whether under-dosage of the CTV also occurred.

An alternative approach to online ART was introduced by Ramakrishnan et al. (2012). Instead of adapting beam apertures, they described a dynamic programming approach whereby the fraction size is altered depending on the proximity of the target to an OAR: if the target is nominally closer to the OAR, the fraction size is reduced; if the target is further from the OAR, the fraction size is increased. They demonstrated this principle in a planning study, and showed that the cumulative dose to an OAR can be reduced by using this technique. However, the radiobiological implications of this technique are not known, and its application to the head and neck, in which there can be several OARs in close proximity to the target, would likely be non-trivial.
Online ART is technically challenging and labour-intensive, and is probably not necessary in the head and neck, in which anatomic changes generally occur over a period of weeks (Schwartz and Dong, 2011). The majority of ART techniques for the head and neck have therefore focused on offline correction strategies.

Wu et al. (2009) retrospectively investigated the dosimetric benefit of different ART strategies in a small cohort of 11 patients. They found that replanning at midcourse reduced the parotid dose by 3% relative to no replanning, two equally-spaced replans improved parotid sparing by 5%, and weekly replanning improved parotid sparing by 6%. These results assumed that it took one week to implement the replan. Instant implementation of a replan (online ART) resulted in parotid sparing of 8%, which almost completely compensated for the mean parotid mean dose increase of 10% when no ART was used. These results imply that offline replanning twice during the course of treatment is likely to be sufficient.

Schwartz et al. (2012) performed a prospective trial to investigate whether ART is safe and whether it provides any clinical benefits. They performed daily CT-on-rails image guidance for each patient, and calculated the fractional dose at least weekly. Plan adaptation was performed at least once for each patient, using no CTV-to-PTV margin for the adapted plan. Patients who received a single replan received a lower parotid dose of 3%–4% compared to IGRT alone. As this was a prospective trial, they adapted plans when significant changes were observed, and they found that the median time for replanning was fraction 16 (30–33 fractions overall). Eight patients required a second replan, providing sparing of 4% and 9% to the contralateral and ipsilateral parotids, respectively, with median replanning points at fractions 11 and 22.

Zhao et al. (2011) also retrospectively investigated ART for 33 nasopharyngeal patients who had been previously replanned due to tumour and/or nodal shrinkage and weight loss at their institution. They reported that the mean replanning
fraction for these patients was fraction 15, which corresponded to approximately half way through treatment. This agrees with the findings of Schwartz et al. (2012).

Another ART strategy was described by Ahn et al. (2011), who described their protocol in which every head and neck patient is routinely rescanned midway through treatment. Contours are mapped onto the new scan and edited by a clinician, before the original plan is copied onto the new scan. The dosimetry is assessed and it is determined whether a new plan would be beneficial. They base the assessment on several factors, including target coverage and OAR doses. For example, if the spinal cord exceeds tolerance a new plan may be required. However, they assess previous CBCT scans to determine if the cord position is reproducible before they decide to replan. Whilst they do not explicitly mention the proportion of patients that are actually replanned, it is clear that this is a time-consuming process and would not be a practical solution in the majority of centres.

Whilst the previous studies have investigated the potential benefit of one or two replans during treatment, Zhang et al. (2016) investigated 63 replanning scenarios for 13 oropharyngeal cancer patients who received weekly CT imaging. In a treatment planning study, they calculated the benefit of differing number of replans at different time points with respect to target coverage and parotid gland mean dose. The mean reduction in parotid mean dose relative to standard IGRT was 3.3 Gy, and 94 % of this benefit was obtained from three replans. They suggested that the optimal replanning strategy was to replan in weeks 1, 3 and 5; this resulted in a reduction of mean dose to the contralateral parotid of > 5 Gy for 31 % of patients.

With regards to the optimum timing for ART, Kager et al. (2015) investigated the effect of hydration that is often performed as part of cisplatin-based chemotherapy on the parotid volume. Although the overall parotid volume changes were
in line with patients treated with radiotherapy alone, they reported that for pa-
tients treated with concurrent chemotherapy, the parotids temporarily increased
in volume by up to 11 %, which coincided with the start of a chemotherapy cycle
and the administration of saline solution for hydration. Whilst this temporary
increase in volume was not clinically significant overall, they suggested that it
could have implications for ART if a replan coincided with the parotids’ being in
the temporarily enlarged state.

The majority of studies investigating ART strategies have used a new treat-
ment plan based on a repeat CT. However, a CT scan provides a snapshot of
the patient anatomy, which can contain systematic deformations. van Kranen
et al. (2013) argue that using a repeat CT scan for replanning will introduce new
systematic deformations. They suggested a new adaptive strategy, in which the
pCT is modified based on the average anatomy derived from daily imaging at the
eyear stages of treatment. They found that residual misalignments between bony
and soft tissue landmarks identified on the modified CT and the daily CBCT
images were reduced by 40 % (bony) and 19 % (soft tissue) for a single inter-
vention at fraction 10, and by 61 % (bony) and 33 % (soft tissue) for weekly
interventions, relative to no intervention. The reduction in systematic deform-
ations could enable a reduction in CTV-to-PTV margin, and van Kranen et al.
(2016) followed up their initial study by investigating the effect this would have
on OAR and target coverage during head and neck radiotherapy. Simulating on-
line image guidance, they investigated CTV-to-PTV margins of 5 mm, 3 mm and
0 mm, and reported a reduction in mean dose to OARs of approximately 1 Gy
per mm reduction in margin. However, the use of a 0 mm CTV-to-PTV margin
also resulted in reduced CTV coverage, requiring ART to restore the intended
dose. For these patients with reduced CTV coverage, a single replan based on
the pCT modified according to the average anatomy from the previous 10 CBCT
images restored the intended cumulative CTV coverage for four out of the six
patients who required ART. The two patients for whom CTV coverage was not restored likely required further plan adaptation.

These studies have investigated potential ART strategies in terms of replanning frequency for head and neck IMRT. The majority of techniques have focussed on offline ART due to the incremental anatomic changes that are typical during head and neck radiotherapy. ART strategies are justified based on the dosimetric changes that occur as a result of anatomic changes, and so strategies based on head and neck IMRT are not necessarily applicable to other treatment modalities, such as VMAT. The majority of planning studies highlighting the potential benefit of ART have used IMRT, and so ART strategies derived from these results are not necessarily directly applicable to VMAT. Potential differences in robustness to anatomic changes between IMRT and VMAT have not been explicitly explored for head and neck radiotherapy.

1.5.2 Predicting anatomic changes

Although ART can mitigate some of the dosimetric changes that can occur as a result of anatomic changes during treatment, it has the potential to be extremely labour-intensive. Whilst the dosimetric benefit for some patients is clear, ART is probably not necessary for all patients (Marzi et al., 2012). The likely spread of dosimetric benefit afforded by ART means that identification of the patients for whom ART will be of benefit is important for increasing the efficiency of an ART workflow. Such patients are likely to be those who experience extreme anatomy changes throughout treatment, and so predicting anatomic changes has the potential to inform the selection of these patients.

Weight loss is common during radiotherapy, and can be associated with volumetric changes of structures in the head and neck and increased difficulty in patient setup (Barker et al., 2004; Broggi et al., 2010; Ahn et al., 2011). Predicting weight loss in individual patients was explored by Yan et al. (2013), who
found that patients with a pre-treatment body mass index $\geq 25$ lost weight at a faster rate than patients with a body mass index $< 25$. They also reported a greater rate of CTV volume reduction for patients with a larger body mass index and recommended that these patients should be considered for ART.

Another commonly-reported anatomic change that occurs during radiotherapy is GTV volume reduction, and Barker et al. (2004) reported that the rate of volume loss of the GTV correlated with its initial volume. Yock et al. (2014) developed and tested different statistical models for predicting GTV volume changes based on the initial tumour volume for oropharyngeal cancer patients, and suggested that such models could be used to identify patients who were likely to experience large volume changes during treatment. These patients could then be selected for ART, and the predicted volume change could be used to infer the number and timing of replans.

One of the most commonly-reported anatomic change during head and neck radiotherapy is that of the parotids. Predicting the parotid volume change throughout treatment has therefore been the focus of several investigations, with dose (Vásquez Osorio et al., 2008; Wang et al., 2009; Broggi et al., 2010; Fiorino et al., 2011; Sanguineti et al., 2013), initial parotid volume (Broggi et al., 2010) and age (Broggi et al., 2010; Sanguineti et al., 2013) having been found to correlate with change in parotid volume by the end of treatment.

Broggi et al. (2010) investigated of parotid changes for 87 patients in four separate institutions and reported many factors that correlated with volumetric changes in the parotid. They focused on pre-treatment predictive factors and reported that the two strongest predictors for absolute volumetric parotid changes were the mean dose to the parotid and the initial parotid volume. For the relative parotid volume changes, the strongest predictors were age and the volume receiving 40 Gy ($V_{40Gy}$), although the predictive power of their relative volumetric change model was lower than that for absolute volumetric change (Spearman’s...
rank correlation coefficients of 0.22 and 0.58 for the absolute and relative changes, respectively). Fiorino et al. (2011) analysed a subset of these patients using the Jacobian of the deformation vector field to assess the shrinkage of individual voxels in the parotids, and reported a correlation between the parotid shrinkage and the low-dose bath ($V_{10Gy} - V_{15Gy}$).

Sanguineti et al. (2013) also investigated parotid changes throughout radiotherapy in a study of 85 patients. They noted that age correlated with parotid volume loss, with younger patients losing more parotid volume, but that sex, tumour location, tumour stage and concurrent chemotherapy did not.

Early variation in image texture, which quantifies the spatial relationship between intensities in an image, has also been investigated for predicting changes in parotid volume. Several textural parameters, such as the mean Houndsfield unit (HU) value, variance, entropy, homogeneity and fractal dimension have been found to correlate with parotid shrinkage (Fiorino et al., 2012; Scalco et al., 2013). Image texture analysis quantifies subtle changes in the distribution of voxel intensities, and so repeat CT imaging would probably be required (Fiorino et al., 2012), which is often impractical on a routine basis.

Whilst several studies have investigated correlations between different pre-treatment parameters and dosimetric or anatomic changes, relatively few have incorporated them into a model to predict the requirement for ART. Brouwer et al. (2016) created a linear regression model to predict the need for ART based on the planned mean parotid dose. Training their model on 113 patients, they proposed that parotids with a mean dose $>22.2$ Gy should be replanned at some point during treatment. However, upon validation with an independent cohort using receiver operator characteristic (ROC) analysis, their model achieved an area under the curve of 0.53. The use of this model on this cohort of patients would have spared 24% of patients from ART (assuming all patients would have otherwise had plan adaptation); however, with a positive predictive value of 19%,
the number of false positives would still result in a large amount of unnecessary replanning.

Similarly, Castelli et al. (2016) created a model to predict the necessity of ART for a particular patient based on parameters from the initial planning CT or changes in the first week of treatment. In a cohort of 20 patients, they created a linear regression model and nomogram that predicted the increase in cumulative dose to the parotids based on the volume of the primary CTV receiving 70 Gy, the change in this volume between the original CT and the first week of treatment, and the change in parotid mean dose between the pCT and the first week of treatment. Using leave-one-out cross validation, they reported a sensitivity of 80% and a specificity of 60% when predicting whether the cumulative parotid dose would increase or decrease. However, they acknowledge that validation of the model in a larger, independent dataset is required.

1.5.3 Deformable image registration and automatic segmentation

One of the key elements of any ART strategy is accurate deformable image registration (DIR), which can be used for automatic segmentation and dose accumulation. DIR enables a spatial transformation from one image, often termed the floating image, to another image, often termed the fixed image. This provides a one-to-one mapping between points in the two images, and is illustrated in Figure 1.6. The floating image is shown in Figure 1.6a and the fixed image is shown in Figure 1.6b. The spatial transformation, which is commonly represented by a deformation vector field (DVF), can then be applied to the floating image, deforming it into the coordinate system of the fixed image (Figure 1.6c). An overlay of the floating and fixed images is shown in Figure 1.6d.

DIR is an optimisation problem that aims to determine the transformation,
Figure 1.6: Deformable image registration between a floating image (a), and a fixed image (b). The deformation vector field is used to deform the floating image into the coordinate system of the fixed image (c). An overlay of the floating image and the fixed image is shown in (d).

$h$, that minimises the difference between a fixed image, $A$, and a transformed floating image, $B(h)$. The DIR cost function, $D$, takes the general form

$$D = S(A, B(h)) + \alpha R,$$

where $S$ is the measure of similarity between the fixed image and the transformed floating image, $R$ is the regularisation term used to penalise unrealistic deformations, and $\alpha$ is a constant that controls the relative weights of the similarity and regularisation components.

Similarity metrics used in DIR algorithms can be broadly separated into feature- and intensity-based. Intensity-based algorithms seek to minimise image intensity differences in individual voxels or image patches, such as in the popular Demons algorithm (Thirion, 1998). Some of the most commonly-used intensity
based metrics include the sum of squared differences (SSD), normalised cross-correlation (NCC), and mutual information (MI). SSD is based on the assertion that equivalent structures should have identical image intensities in the floating and fixed images, and so registration seeks to match voxels with equivalent intensity values (Oliveira and Tavares, 2014). This works well for images from the same modality (e.g., CT-CT), but when there is an intensity offset between the floating and fixed images, such as for inter-modality registration, this metric is not suitable (Nithiananthan et al., 2011). However, image preprocessing, such as histogram intensity matching, can alleviate this problem somewhat and improve inter-modality image registration (Hou et al., 2011). Inter-modality registration is more often handled using metrics such as NCC (Silva et al., 2011), which measures the correlation between two images, and MI, which maximises the amount of information, quantified by entropy, the two images contain about each other (Pluim et al., 2003; Veiga et al., 2014).

Intensity-based DIR can be limited by image artefacts and in regions of low contrast. Feature-based DIR uses known features in the fixed and floating images to determine the transformation. These features are often anatomic landmarks or contours that have been identified in both the fixed and floating images prior to registration. Manual identification of these features can be time-consuming and are not practical for routine ART (Castadot et al., 2008), but automated methods of identifying features in images can improve efficiency (Vásquez Osorio et al., 2012). Furthermore, hybrid approaches that combine intensity-based and feature-based DIR methods have also been suggested (Gu et al., 2013; Weistrand and Svensson, 2015).

One of the key uses for DIR within the context of ART is for automatic segmentation. Outlining structures for radiotherapy treatment planning is extremely time-consuming, particularly in the head and neck where there is a large number of complex structures (Harari et al., 2010). ART requires regular dosimetric
Figure 1.7: Atlas-based automatic segmentation. An atlas consists of an image and corresponding contours (left). Contours are propagated from the atlas to a new image (right) according to the results of non-rigid registration between the atlas and new image. In this example, the atlas is the pCT and the new image is a CBCT of the same patient obtained in the sixth week of treatment. The parotids are blue, the mandible green, the spinal cord red and the GTV orange.

assessment of a treatment plan on updated patient anatomy and the generation of a new treatment plan if changes are sufficiently large. For this it is necessary for the structures of interest to be outlined. It is not feasible to repeatedly outline structures manually, and methods of automatic segmentation are essential (Zhang et al., 2007; Muren and Thwaites, 2013).

Automatic segmentation techniques can be broadly separated into model-based and atlas-based methods. Model-based algorithms use a physical model to guide the segmentation based on prior knowledge about the features and shape of the organ to be segmented (Heimann and Meinzer, 2009). For example, Al-Mayah et al. (2010) successfully used a biomechanical-based DIR technique for automatic segmentation of structures in the head and neck. Whilst typically robust to image artefacts (Heimann and Meinzer, 2009), the segmentations produced by model-based techniques are limited by the shapes and features of the structures included in the model, meaning their performance is very sensitive to the data used to build the model (Sharp et al., 2014).
Atlas-based segmentation uses a different approach whereby prior knowledge is used in the form of an atlas, which is an image with corresponding segmentations (Han et al., 2008). Atlas-based segmentation can be used for the initial segmentation of the pCT, for which the atlas is registered to the pCT and the contours associated with the atlas are propagated to the pCT. The performance of atlas-based segmentation of the pCT is strongly dependent on the registration accuracy, the quality of the segmentation within the atlas itself, and the anatomic similarity between the atlas and patient image (Aljabar et al., 2009). Selecting the most appropriate atlas from a database according to image similarity and using an atlas based on an average patient anatomy have been shown to improve segmentation performance relative to an arbitrary choice of a single atlas (Rohlfing et al., 2004; Wu et al., 2007; Commowick et al., 2009). Performance can be further improved by using a multiple-atlas approach in which the atlas is registered to each patient image, and a voxel-by-voxel voting system is used to classify each individual voxel (Rohlfing et al., 2004).

Atlas-based segmentation can also be used for segmentation of on-treatment images. The concept is illustrated in Figure 1.7, which shows contours delineated on a pCT propagated onto a CBCT obtained during the sixth week of treatment for the same patient. The pCT is registered to the CBCT, and the DVF describing the displacement of each voxel in the pCT to its corresponding position in the CBCT can be used to propagate the structures in the pCT into the coordinate system of the CBCT (Zhang et al., 2007; Castadot et al., 2008). This is known as intra-patient segmentation or contour propagation. Contours on the pCT are propagated onto an on-treatment image, often a CBCT, and several algorithms have been evaluated for this use (Table 1.3). Alternative strategies for contour propagation have been suggested, and Hvid et al. (2016) investigated a method in which contours were propagated from pCT onto a CBCT via a CBCT obtained on the first treatment fraction. That is, contours were propagated from
pCT onto the first CBCT and then directly onto another CBCT. They reported that this method resulted in poorer geometric accuracy, assessed by comparison of propagated contours with manually adjusted contours, than direct propagation from pCT to CBCT. Similarly, Godley et al. (2013) compared three different methods of contour propagation in the male pelvis. They compared a direct pCT-to-CBCT propagation, propagation from the CBCT image from the previous treatment fraction, and a multiple-atlas approach using CBCT images from several previous fractions. They found that the multiple-atlas approach was superior to the other methods, with propagation from the CBCT from the previous fraction slightly better than a direct propagation approach.

Validation of automatic segmentation before clinical implementation is essential, and is generally performed by measuring the geometric agreement between automatically-generated contours and manually-drawn ‘ground truth’ contours. Many different metrics exist for quantifying the accuracy of automatic segmentation, and they can be broadly separated into moment-based, overlap-based and surface-based metrics (Sharp et al., 2014). The simplest form of metrics are those based on moment statistics, such as the centre of mass (Hardcastle et al., 2013) or the volume of the structure (Greenham et al., 2014). Whilst simple to calculate, they do not necessarily represent segmentation accuracy, particularly for more complex shapes.

Overlap-based metrics quantify the amount of overlap between two structures, and include the popular Dice similarity coefficient (DSC) (Dice, 1945). DSC is given by

\[
DSC = \frac{2(V_a \cap V_b)}{|V_a| + |V_b|},
\]

where \(V_a\) and \(V_b\) are the volumes of two structures to be compared (see Figure 1.8). DSC varies between 0, in which there is no overlap, and 1, which indicates perfect
Figure 1.8: The Dice similarity coefficient quantifies the spatial overlap of two structures with volumes $V_a$ and $V_b$. Image adapted from Beasley et al. (2016).

agreement between the two volumes. Other variants of DSC exist, such as the Jaccard index, but DSC is the most commonly-reported metric in the literature. However, although DSC is relatively simple to interpret, it is highly dependent on the volumes of the structures being outlined, and does not quantify the distance between the two volumes (Sharp et al., 2014; Sykes, 2014).

Surface-based metrics quantify the distance between two surfaces, and are often based on the distance-to-agreement (DTA). Also known as Hausdorff distance, contour distance or surface distance, they are not well standardised; for the purposes of this discussion, DTA will be used to refer to these metrics. For two surfaces $A$ and $B$ characterised by points $A = \{a_1, a_2, \ldots, a_n\}$ and $B = \{b_1, b_2, \ldots, b_m\}$, DTA is calculated for a point on surface $A$ from

$$\text{DTA}(a, B) = \min_{b \in B} \|a - b\|.$$ 

This is repeated for each point on surface $A = \{a_1, a_2, \ldots, a_n\}$ and a DTA histogram is built up. This is illustrated in Figure 1.9. For each point on surface $A$, the minimum distance to any point on surface $B$ is computed (Figure 1.9a) and the distribution of distances is stored as a histogram (Figure 1.9b). From this histogram, several different metrics can be obtained. These include the mean and
Figure 1.9: For each point on surface A (red), the shortest distance to surface B (blue) is computed (a) and can be combined into a DTA histogram (b).

median (Qazi et al., 2011; Hardcastle et al., 2013; Thomson et al., 2014), maximum (Thomson et al., 2014) and 95th percentile (Huger et al., 2014). Whilst these metrics provide a quantitative measure of segmentation accuracy, it can be difficult to interpret the clinical relevance of discrepancies (Sykes, 2014).

Quantification of segmentation accuracy is usually performed relative to reference structures manually drawn by an experienced clinician; this is typically called the ‘ground truth’. However, manual delineations are subject to inter-and intra-observer variation (Nelms et al., 2012), and so benchmarking automatic segmentation algorithms relative to a contour drawn by a single observer is subject to uncertainty. It can therefore be more informative to compare automatically-generated structures against multiple delineations of the same structure from different clinicians. Comparison can then either be performed relative to a STAPLE volume, which is a consensus structure formed from all delineations (Warfield et al., 2004), or the inter-observer variation.
The observed accuracy of automatically-generated contours is limited by inter-observer variation, and an alternative technique is to qualitatively score the acceptability of a contour by rating it according to a scoring system. Several different systems have been reported, scoring acceptability out of two (Stapleford et al., 2010), three (Hardcastle et al., 2013), and four (Thor et al., 2011). Whilst such scoring techniques implicitly determine whether a contour is acceptable, scoring is subjective and requires an understanding of where the greatest degree of accuracy is required for treatment planning. A similar method was performed by Hoang Duc et al. (2015), who rated automatically generated contours on a three-point scale for clinical acceptability, but the expert reviewers were blinded to the source of the contours (i.e. automatically generated or manually created); this enabled a more objective comparison of the acceptability of the manual and automatically generated contours.

If automatically-generated contours are to be used for treatment planning or dosimetric assessment for ART, it is important that the dose to these structures is correct (Nelms et al., 2012). This has prompted some studies to determine the acceptability of automatically-generated structures with respect to dosimetric discrepancies compared to the ground truth (Tsuji et al., 2010; Eiland et al., 2014).

Table 1.3 provides an overview of the performance of different automatic segmentation algorithms, from which the variety of metrics used for assessing accuracy is clear. Validation of an automatic segmentation algorithm prior to clinical use is essential, but there is no consensus as to the best approach or the best metrics to use (Sharp et al., 2014). Automatic segmentation is essential for ART, and so standardised reporting of automatic segmentation performance, using the most appropriate metrics, is important.
Table 1.3: Overview of automatic segmentation performance.

<table>
<thead>
<tr>
<th>Structure</th>
<th>Reference</th>
<th>Modality</th>
<th>Metrics used</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parotid</td>
<td>Zhang et al. (2007)</td>
<td>CT-CBCT</td>
<td>DSC</td>
<td>~ 0.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>DT</td>
<td>2.03 mm</td>
</tr>
<tr>
<td></td>
<td>Thomson et al. (2014)</td>
<td>Atlas-CT</td>
<td>DSC</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mean DTA</td>
<td>1.6 mm</td>
</tr>
<tr>
<td></td>
<td>Varadhan et al. (2013)</td>
<td>CT-CT</td>
<td>DSC</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>HD</td>
<td>5.7 mm</td>
</tr>
<tr>
<td></td>
<td>Macchia et al. (2012)</td>
<td>Atlas-CT</td>
<td>DSC</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Sensitivity</td>
<td>0.78</td>
</tr>
<tr>
<td>Brainstem</td>
<td>Zhang et al. (2007)</td>
<td>CT-CBCT</td>
<td>DSC</td>
<td>~ 0.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>DT</td>
<td>1.79 mm</td>
</tr>
<tr>
<td></td>
<td>Varadhan et al. (2013)</td>
<td>CT-CT</td>
<td>DSC</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>HD</td>
<td>11.9 mm</td>
</tr>
<tr>
<td></td>
<td>Macchia et al. (2012)</td>
<td>Atlas-CT</td>
<td>DSC</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Sensitivity</td>
<td>0.74</td>
</tr>
<tr>
<td>Submandibular</td>
<td>Thomson et al. (2014)</td>
<td>Atlas-CT</td>
<td>DSC</td>
<td>0.75</td>
</tr>
<tr>
<td>glands</td>
<td></td>
<td></td>
<td>Mean DTA</td>
<td>0.6 mm</td>
</tr>
<tr>
<td></td>
<td>Huger et al. (2014)</td>
<td>CT-CBCT</td>
<td>DSC</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>95%-HD</td>
<td>3.1 mm</td>
</tr>
<tr>
<td>Nodes</td>
<td>Zhang et al. (2007)</td>
<td>CT-CBCT</td>
<td>DSC</td>
<td>~ 0.75</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>DT</td>
<td>3.15 mm</td>
</tr>
<tr>
<td></td>
<td>Stapleford et al. (2010)</td>
<td>Atlas-CT</td>
<td>DSC</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mean SD</td>
<td>3.3 mm;</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Max SD</td>
<td>17.4 mm</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>∆Volume</td>
<td>103 mL</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Sensitivity</td>
<td>85 %</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>FPR</td>
<td>32 %</td>
</tr>
<tr>
<td>Larynx</td>
<td>Thomson et al. (2014)</td>
<td>Atlas-CT</td>
<td>DSC</td>
<td>~ 0.60</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mean DTA</td>
<td>~ 6 mm</td>
</tr>
<tr>
<td></td>
<td>Varadhan et al. (2013)</td>
<td>CT-CT</td>
<td>DSC</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>HD</td>
<td>5.7 mm</td>
</tr>
<tr>
<td></td>
<td>Macchia et al. (2012)</td>
<td>Atlas-CT</td>
<td>DSC</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Sensitivity</td>
<td>0.87</td>
</tr>
</tbody>
</table>

Abbreviations: DSC = Dice similarity coefficient; DT = distance transformation; DTA = distance-to-agreement; HD = Hausdorff distance; SD = surface distance; FPR = false positive rate.
Automatic segmentation is not generally considered sufficiently reliable to fully replace the clinician. Manual review of propagated contours is therefore considered necessary (Acharya et al., 2016); however, this is also time-consuming and limits the implementation of ART. Methods for inferring automatic segmentation accuracy when there is no manually-drawn ground truth would facilitate the implementation of ART, but are currently lacking. Techniques based on machine learning have been suggested for assessing segmentation accuracy in the absence of a ground truth (Kohlberger et al., 2012; McIntosh et al., 2013; Chen et al., 2015), and show promise as a potential tool for quality control (QC) of automatic segmentation.

In addition to contour propagation, DIR can be used for dose accumulation, in which doses to individual voxels are tracked during treatment and summed on a single reference image, typically the pCT. Whilst beyond the scope of this thesis, dose accumulation will play an important role in adaptive radiotherapy and so a brief discussion is provided here. By tracking the dose that is delivered to on-treatment images it is possible to determine deviations between the planned and delivered dose, and modify the treatment plan accordingly. However, dose accumulation is challenging, as accurate DIR is required inside contours as well as at contour boundaries. This means that methods for assessing automatic segmentation accuracy, in which accuracy at high contrast features such as anatomic boundaries or landmarks, are not suitable for assessing dose accumulation accuracy. Yeo et al. (2012) showed in their study using a deformable gel phantom that accurate registration in high contrast regions does not necessarily result in accurate registration in low contrast regions. This was further highlighted when they compared the performance of 12 algorithms and again found that algorithms that performed similarly for high contrast features performed differently in low contrast regions (Yeo et al., 2013). Alternative methods of measuring DIR accuracy in the context of dose accumulation have therefore been suggested, such
as ones based on the physical characteristics of the underlying DVF (Varadhan et al., 2013).

In addition to the challenges of assessing a DIR algorithm for dose accumulation, the appearance or disappearance of tissue during treatment, such as with tumour progression or regression, leads to difficulties when summing doses from different time points during treatment Schultheiss et al. (2012). Nevertheless, dose accumulation has the potential to enable plan adaptation based on the dose actually delivered to the patient, and tools are being developed to integrate it into the clinical workflow (Park et al., 2016).

1.6 Predictive factors of radiotherapy toxicities

DIR and automatic segmentation enable measurement of the anatomic changes that occur during head and neck radiotherapy, and ART mitigates the dosimetric effects of these changes. This is important because the dose to different OARs is related to the risk of many toxicities in head and neck radiotherapy. The risk of xerostomia is related to the dose and irradiated volume of the parotid gland, with a review of published clinical data suggesting that in order to maintain at least 25% of baseline salivary function it is necessary to ensure that one gland is spared to a mean dose of $< 20$ Gy, or that both glands are spared to a mean dose of $< 25$ Gy (Deasy et al., 2010). The PARSORT trial confirmed that reducing the dose to the parotid glands reduces the incidence of xerostomia, and parotid-sparing IMRT is now considered standard for head and neck radiotherapy (Nutting et al., 2011).

Dysphagia is another common side effect of radiotherapy and several clinical factors, such as stage, primary site and baseline performance status are risk factors (Denaro et al., 2013). Dose-response relationships have been identified for the larynx and pharyngeal constrictor muscles (Levendag et al., 2007; Rancati
Another late side effect of radiotherapy, trismus is characterised by difficulties in mouth opening. Risk factors include surgery, age, tumour location and dose to the muscles of mastication (Hsieh et al., 2014). The muscles of mastication include the masseter, lateral and ipsilateral pterygoid, temporomandibular joint (TMJ) and temporalis, and there is evidence of a dose-response relationship for at least some of these structures.

Hsiung et al. (2008) measured the maximum incisor-to-incisor distance (MID) before radiotherapy and 5 and 12 months post-radiotherapy for 17 patients who received parotid-sparing IMRT for nasopharyngeal cancer. They compared the MID of these patients with historical data and reported reduced incidence of trismus in patients treated with IMRT. They suggested that this reduced incidence of trismus was caused by the likely lower doses received by at least some of the muscles of mastication due to the parotid-sparing technique; however, they did not report doses to the individual muscles of mastication.

Hsieh et al. (2014) graded trismus using the Common Terminology Criteria for Adverse Events (CTCAE) v3.0 at 6, 12 and 24 months in 22 patients who received image-guided IMRT. Separating patients into two groups depending on whether CTCAE grade improved or worsened over time, they reported statistically significant differences in dose delivered to the lateral pterygoid muscle and parotid gland for the two groups. However, no statistically significant differences were found between the two groups for the dose delivered to the medial pterygoid muscle, masseter muscle or the temporalis muscle.

In contrast, van der Molen et al. (2013) reported that the mean dose to the masseter and the pterygoid muscles were statistically significant predictors of trismus at 10 weeks post radiotherapy, defined as an MID ≤ 35 mm, in their analysis of 55 patients who received IMRT. Whilst they also obtained MID measurements 12 months post radiotherapy, the incidence of trismus was too low for statistical
Similar results were found by Lindblom et al. (2014) in their study investigating the correlation between the mean dose to the muscles of mastication and the incidence of trismus. They investigated 121 patients, 112 of whom received 3DCRT, and measured trismus using the EORTC H&N35 Quality of Life questionnaire at 3, 6, 12, 24 and 60 months, as well as MID at 66 months. They reported statistically significant correlations between MID and the quality of life questionnaire, and between the mean dose and MID for all structures investigated. The strongest correlation was found for the ipsilateral masseter, with the ipsilateral structures in general providing stronger correlations that the corresponding contralateral structures.

A much larger study was recently performed by Rao et al. (2016), in which 798 patients who received head and neck IMRT were graded according to CTCAE v4.0 at several time points after completion of radiotherapy. Using an endpoint of any occurrences of Grade 1 or greater during the follow-up period, they investigated potential relationships with the dose delivered to the masseter, medial and lateral pterygoid, and temporalis. They reported that the strongest correlates were the ipsilateral masseter and ipsilateral medial pterygoid. Furthermore, they calculated that the mean dose corresponding to a 10% risk of trismus was 40 Gy and 65 Gy for the ipsilateral masseter and ipsilateral medial pterygoid, respectively.

It is clear from these studies that the delivered dose to the mastication structures is likely related to the incidence of trismus after radiotherapy. However, there is no consensus on which structures are the most sensitive, although the two most recent and largest studies have identified the ipsilateral masseter as having the strongest relationship with trismus (Lindblom et al., 2014; Rao et al., 2016). Pauli et al. (2016) have also tried to address this lack of consensus and, in agreement with Lindblom et al. (2014) and Rao et al. (2016), reported that
the dose to the ipsilateral masseter proved the best predictor for trismus. However, trismus is a complex disorder and Rao et al. (2016) acknowledge that one of the limitations of their study is the use of delineated structures; structures not included in the analysis could potentially contribute.

1.7 Clinical benefit of ART

Dose-response relationships for the different structures in the head and neck imply that reducing the dose to certain OARs will reduce the incidence of corresponding side effects. ART mitigates the dosimetric changes that can occur during head and neck radiotherapy, and it is thought that the reduction in cumulative dose to the parotids will translate into a reduction in the risk of xerostomia. Marzi et al. (2012) calculated the increased xerostomia risk resulting from anatomic and dosimetric changes during head and neck radiotherapy in 15 patients using a Lyman-Kutcher-Burman (LKB) NTCP model. Although they reported a modest increase in both dose and NTCP for the patient population, they identified significant increases in NTCP on an individual patient basis. Similarly, Castelli et al. (2015) estimated the potential reduction in xerostomia risk from weekly replanning for 15 patients using an LKB NTCP model. They reported that weekly replanning could reduce the mean dose to the parotids by 5.1 Gy and the risk of xerostomia by 11%. However, as this was a treatment planning study, this estimated reduction in the risk of xerostomia could not be validated.

Hunter et al. (2013) investigated the clinical impact of increased dose to the parotids as a result of anatomic changes during head and neck radiotherapy. They investigated 18 patients who underwent salivary tests before, during and after treatment (up to 24 months). They accumulated the dose from daily CBCT images and calculated the dose actually received by the parotid glands and investigated whether this provided better correlation with toxicity, compared to
the planned dose. Although both the planned and delivered doses were correlated with salivary function, the delivered dose did not provide better correlation than the planned dose. However, they only observed an increase in parotid dose of <1 Gy (2.8 %), which is smaller than other values reported in the literature (see Table 1.2). This would make any differences between the correlation of planned and delivered dose with salivary function difficult to identify with such a small sample size.

The above study had a small sample size and did not explicitly investigate the benefits of ART, as patients were not replanned, but the results imply that ART may not have been able to improve salivary function for all patients. Schwartz et al. (2012) investigated the clinical benefits of an adaptive strategy in a prospective clinical trial. They performed ART on 22 patients and reported dosimetric improvements for their ART strategy relative to daily online IGRT. They found no significant reduction in acute toxicity relative to their standard technique, but reported that “early chronic toxicity results suggest encouraging post-treatment functional recovery”. However, they stressed that these results are preliminary.

A study investigating the clinical benefit of ART was also performed by Zhao et al. (2011), who retrospectively compared the clinical outcomes of 33 patients who had been replanned in their institution against 66 control patients who were judged to have had undergone “obvious anatomical changes before fraction 20”, but had not been replanned. They followed up the patients for 40 months and reported no difference in survival or toxicity for early-stage patients. However, for late stage patients (T3 or above), a significant improvement in 3-year relapse-free survival was observed for those who had received a replan. Additionally, late side effects were reduced for patients with high nodal staging (N≥2). The findings suggest that adaptive planning may have provided a clinical benefit for specific patients, although it should be noted that the study was retrospective and could therefore be subject to biases, particularly in the selection of control patients.
These studies provide evidence that ART enables reduction of radiotherapy-related xerostomia in at least some patients, but it is likely that this benefit will be modest on a population level. In parallel with dosimetric changes, the patients who will likely benefit the most from ART will be those that experience the largest anatomic changes during treatment.

The dosimetric consequences of anatomic changes during radiotherapy depend on several factors, including the location of the tumour relative to the OARs and local dose gradients in the treatment plan (Castadot et al., 2010). There are several treatment-related sequelae for head and neck radiotherapy, and as more advanced radiotherapy treatments, such as proton beam therapy and magnetic resonance-guided radiotherapy, become available, the potential to further reduce toxicities will be explored. This will be accomplished by reducing the dose to sensitive structures, resulting in local steep dose gradients.

Similarly to parotid-sparing IMRT, any potential dosimetric benefits of sparing sensitive structures could be lost if the entire treatment is based solely on a pre-treatment CT scan. Yang et al. (2013) reported that patients who received a replan during radiotherapy for nasopharyngeal cancer suffered fewer complications, including speech problems, dry mouth and reduced mouth opening, compared to patients who were offered a replan but refused. Patients who received a replan also reported a higher general quality of life after treatment compared to patients who did not. However, the reasons for patients refusing a replan were not reported, and the authors acknowledge that the non-randomised nature of the study was a limitation.

In order to optimise ART, it is necessary to consider the different toxicities that occur as part of treatment. Accurate knowledge of dose-response relationships for different toxicities will help determine which patients will benefit most from adaptive radiotherapy. However, radiotherapy related toxicities can be complex, and it is not always clear which anatomic structures should be considered
1.8 Rationale and aims

The rationale for this thesis is the optimisation of ART for head and neck cancer, both in terms of clinical implementation and clinical benefit. This includes investigating methods of optimising the efficiency of ART to enable routine clinical adoption, and methods of optimising the clinical benefit of ART by identifying which structures are important for a particular endpoint. To this end, the following aims have been identified:

1. To determine the relative plan robustness to weight loss during radiotherapy of VMAT and IMRT.

2. To determine which metrics should be used for assessing automatic segmentation accuracy in the head and neck.

3. To develop a method of performing patient-specific quality control of contour propagation accuracy.

4. To develop a statistically robust image-based data mining technique for identifying structures exhibiting a dose-response relationship with a continuously variable toxicity outcome.

5. To determine which structures in the head and neck have a dose-response relationship with trismus.

1.9 Thesis structure

Permission has been granted for submission of this thesis in alternative format. Whilst the projects included within the thesis comprise a coherent set of experiments, they are well-suited to individual investigations. This section outlines
the structure of this thesis and how the individual aims identified in the previous section have been addressed.

Chapter 2 describes an investigation into potential differences in plan robustness to weight loss during head and neck radiotherapy between IMRT and VMAT, and addresses aim 1. There is much literature on the dosimetric changes that occur during head and neck IMRT, but relatively little for head and neck VMAT. Differences might exist between the two techniques due to the fundamental difference in delivery characteristics. ART workflows are based on the dosimetric changes that occur during treatment, and so differences in the robustness of IMRT and VMAT to anatomic changes could mean that the frequency and number of required replans are different. With the well-established delivery time efficiency benefits of VMAT relative to IMRT, VMAT is gaining in popularity. The present study was the first to measure the relative plan robustness of IMRT and VMAT to weight loss during radiotherapy, and has been published in Medical Dosimetry. It is presented here with minor modifications to some of the wording and formatting to improve clarity and to ensure consistency throughout the thesis.

The suitability of different metrics often used for validating automatic segmentation algorithms was determined in Chapter 3, addressing aim 2. This was performed by measuring the correlation between geometric discrepancies quantified by different metrics with the resulting dosimetric discrepancies. A suitable metric was defined as one whose result correlated with the dosimetric discrepancies. This work addresses the lack of consensus in the literature on which metrics should be used for validating automatic segmentation algorithms. Lack of consistency in reporting algorithm performance makes comparison of different algorithms difficult, and the results of this study should inform the selection of metrics used when automatic segmentation algorithms are assessed. This chapter has been published in the Journal of Applied Clinical Medical Physics, and is presented here in its published form, subject to changes in formatting.
Chapter 4 also concerns automatic segmentation, but focuses on the need for automated methods of monitoring contour propagation performance. Patient-specific contour propagation accuracy is currently assessed manually, which is a barrier to routine ART implementation. The chapter describes a novel method of inferring accuracy without ground truth that can be used to perform patient-specific quality control for contour propagation, and addresses aim 3. This method was validated on a set of head and neck cancer patients by testing its ability to detect artificially created errors. The proposed workflow has potential as a flag for manual review of propagated contours that are likely to be incorrect, and could facilitate routine adoption of ART. This chapter has been published in Physics in Medicine and Biology and is presented in its published form, subject to changes in formatting.

Optimising ART requires consideration of different potential clinical benefits afforded by plan adaptation as well as efficient use of automation to enable it to be routinely implemented within a clinical setting. ART workflows have generally focussed on reducing the dose to the parotids, reducing the incidence of xerostomia. However, other sequelae exist in head and neck radiotherapy. Radiotherapy and ART may have the potential to reduce the probability of different toxicities, particularly with the increasing application of advanced radiotherapy techniques such as proton beam therapy and magnetic resonance-guided radiotherapy. In order to assess the potential for reducing toxicities it is necessary to identify accurate dose-response relationships. Image-based data mining is a powerful technique for generating dose-response hypotheses from previously treated patients, but it is subject to a multiple comparisons problem. Permutation testing can correct for this, but it is currently limited to categorical clinical endpoints. In Chapter 5, a novel method of performing permutation testing in image-based data mining for continuously variable clinical endpoints is introduced. Addressing aim 4, it
enables statistically robust image-based data mining to be performed for any clinical endpoint in radiotherapy. This will enable new dose-response hypotheses to be generated and allow for more comprehensive study of radiotherapy related toxicities. In this chapter, the technique is also used to identify structures exhibiting a dose-response relationship with trismus, which addresses aim 5. This chapter is written as a journal article, but has not yet been submitted for publication. Ongoing work to validate the findings in an independent patient cohort will be added to the paper prior to submission to a peer-reviewed journal.
2 Relative plan robustness of IMRT and VMAT for weight loss

Published (subject to some rewording and reformatting) in Medical Dosimetry, 2016, 41, pp. 154–8

Authors

David J Thomson\(^1,2,\ast\), William J Beasley\(^2,3,\ast\), Kate Garcez\(^1\), Lip W Lee\(^1\), Andrew J Sykes\(^1\), Carl G Rowbottom\(^2,3\), Nicholas J Slevin\(^1,2\)

\(^\ast\)These authors contributed equally to this work

\(^1\)Department of Clinical Oncology, The Christie NHS Foundation trust, Manchester, UK
\(^2\)Division of Molecular and Clinical Cancer Sciences, School of Medical Sciences, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester, UK
\(^3\)Christie Medical Physics and Engineering (CMPE), The Christie NHS Foundation Trust, Manchester, UK

Author contributions

I, along with my co-authors, developed the methodology for comparing the relative plan robustness of IMRT and VMAT for weight loss in the head and neck. I created the treatment plans, which included generation of the automated planning protocol for the two techniques, and I applied the original treatment plans to the new anatomy to determine the dosimetric changes. DT identified suitable patients and manually segmented the images. I performed the statistical analysis of the dosimetric changes that resulted from weight loss, and DT and I generated the conclusions. I partly authored the paper, modifying an initial draft produced by DT. All authors read and approved the manuscript.
Abstract

Introduction: Inter-fractional anatomic alterations may have a differential effect on the dose delivered by step-and-shoot intensity modulated radiotherapy (IMRT) and volumetric modulated arc therapy (VMAT). The increased degrees of freedom afforded by rotational delivery may increase plan robustness, defined as change in target volume coverage and doses to organs at risk (OARs). However, this has not been evaluated for head and neck cancer.

Materials and methods: A total of 10 patients who required repeat computed tomography (CT) simulation and replanning during head and neck IMRT were included. IMRT and VMAT plans were generated from the original planning scan. The initial and second CT scans were fused and contours transferred, reviewed and modified. The plans were applied to the second CT scan and doses recalculated without repeat optimisation. Differences between IMRT and VMAT for changes in target volume coverage and doses to OARs between first and second CT scans were compared by Wilcoxon signed rank test.

Results: There were clinically relevant dosimetric changes between the first and the second CT scans for both techniques, including increased mean doses to the parotid glands. However, there were no significant differences between IMRT and VMAT for change in any target coverage parameter or dose to any OARs between the first and the second CT scans.

Conclusions: For patients with head and neck cancer who required replanning mainly due to weight loss, there were no significant differences in plan robustness between step-and-shoot IMRT and VMAT. This information is useful with increased clinical adoption of VMAT.
2.1 Introduction

Intensity modulated radiotherapy (IMRT) for locally advanced head and neck cancer affords complex dose shaping and highly conformed treatment volumes, which compared with 3D-conformal radiotherapy may improve target coverage, organ at risk (OAR) sparing, and functional outcomes (Xia et al., 2000; Nutting et al., 2011). However, the creation of steep dose gradients increases susceptibility to inter-fractional volumetric, shape and positional changes (Castadot et al., 2010).

In studies that have assessed the cumulative dosimetric effect of anatomic changes, the main difference was increased mean dose to the parotid glands, with maintained clinical target volume (CTV) coverage (O’Daniel et al., 2007; Robar et al., 2007; Lee et al., 2008a; Wu et al., 2009; Bhide et al., 2010; Castadot et al., 2011; Ho et al., 2012). There are provisional clinical data to support adaptive replanning to account for inter-fractional dosimetric alterations (Schwartz et al., 2012), and it is thought that approximately 20 % to 30 % of patients may benefit (Grégoire et al., 2012). This group is not well defined but expected to be determined by patient (nutritional status), tumour (site, bulk, stage and proximity to OARs), and treatment factors. This includes planning target volume (PTV) margin, dose gradient and treatment technique; e.g., IMRT or volumetric modulated arc therapy (VMAT). Inter-fractional variables, for example weight loss, fraction number, tumour shrinkage, parotid shift to the mid sagittal plane, and skin separation at various portions of the head and neck may help select patients. However, in a study that correlated positional and anatomic changes with dosimetric parameters, no single factor predicted the need for replan (Ahn et al., 2011).

VMAT is a rotational technique that uses a conventional linear accelerator to deliver highly conformal radiotherapy by variation in gantry speed, beam shape,
and dose rate. Compared with IMRT, VMAT results in reduced treatment delivery time, decreased monitor unit requirements and maintained or improved PTV coverage, conformity, and OAR sparing (Palma et al., 2010). These factors have led to increased adoption of VMAT (Teoh et al., 2011).

The relative dosimetric effect of IMRT and VMAT on treatment plan robustness, defined as inter-fractional change in target volume coverage and doses to OARs, has not been assessed for head and neck cancer. There may be differences between the techniques due to differential effect of asymmetric anatomic alterations or because of variations in treatment conformity. This study sought to address whether there is a difference, which would be important to inform frequency of verification imaging and comparisons between adaptive IMRT studies.

At our institution, patients with head and neck cancer (except early glottic cancer) receive IMRT or VMAT. They are treated supine, positioned using markers on the thermoplastic shell and orthogonal laser beams, and receive image-guided radiotherapy using kilovoltage cone-beam computed tomography (CBCT) imaging for the first three fractions and weekly thereafter. The verification action level is 3 mm, and following any required positional shift, there is daily CBCT imaging for a further three days. The requirement for adaptive replanning is assessed where there are notable anatomic changes (e.g., weight loss, fluid distribution and tumour shrinkage), concerns on volumetric imaging for decreased CTV coverage or increased doses to OARs, or an ill-fitting immobilisation mask. For this, a dose calculation is performed on the CBCT images by performing bulk density overrides to account for gross changes relative to the planning CT. Replanning, with a repeat CT scan, is considered when deviations greater than 2 % are observed in the cumulative dose to relevant clinical structures and deemed clinically important.

We instigated this retrospective study to investigate the relative plan robustness between step-and-shoot IMRT and VMAT in patients who required repeat
CT imaging and replanning for head and neck cancer. In theory, the increased
degrees of freedom afforded by rotational delivery may increase plan robustness.
However, this has not yet been evaluated for head and neck cancer.

2.2 Method

2.2.1 Patient selection

A retrospective chart review at our institution identified 10 patients with head
and neck cancer who received IMRT or VMAT from October 2009 to January
2013 and required repeat CT simulation and replanning during treatment.

2.2.2 Treatment planning

Patients were immobilised in a custom-made 5-point fixation head, neck and
shoulder thermoplastic mask. A radiotherapy planning CT scan was performed
in the treatment position with a slice thickness of 3 mm. CT data were transferred
to the Pinnacle Treatment Planning System (version 9.6). For each patient, target
volumes and OARs were manually contoured on the planning CT scan and for
quality assurance independently reviewed by a second radiation oncologist.

The gross tumour volume (GTV) was defined as the primary tumour and
involved lymph nodes defined as those ≥10 mm in short axis diameter or with
the presence of a necrotic centre. The CTV was the GTV and areas deemed at
risk of microscopic disease. A total of three CTVs were delineated: CTV1 was
defined as the GTV with an isotropic expansion of 10 mm (edited for natural
barriers to disease spread including bone, air, and fascia); CTV2 was defined
as the remainder of the involved subsite and nodal levels, or following a neck
dissection, pathologically-involved nodal levels; and CTV3 was defined as nodal
levels deemed at lower risk of microscopic disease spread: ipsilateral Ib-Vb, VIIb;
contralateral II-IVa; bilateral VIIa for hypopharyngeal primary cancers and ipsilateral VIIa for oropharyngeal cancers; and, where IVa was involved, IVb and Vc. A total of three PTVs (PTV1, PTV2, and PTV3) were defined by an isotropic expansion of 4 mm from CTV1, CTV2, and CTV3, respectively, and edited 5 mm from the skin surface.

Mean doses of 66 Gy, 60 Gy and 54 Gy in 30 daily fractions were prescribed to PTV1, PTV2 and PTV3, respectively. A 6 MV simultaneous integrated boost technique was used. Each patient was retrospectively planned with both IMRT (7 equally spaced beams) and VMAT (two 360° arcs) techniques, resulting in an IMRT and VMAT plan for each patient. A user-independent automated planning method was used to generate the plans for both techniques (Boylan and Rowbottom, 2014). This ensured that treatment plans were free from planner bias, allowing a fair comparison between the two techniques. Planning objectives for OARs were set to be identical for the IMRT and VMAT plans for each patient, ensuring that plans of equivalent quality were produced.

For the PTVs, it was ensured that the minimum dose to 95 % of the volume (D95%) was greater than 95 % of the prescribed dose, and that the minimum dose to 1 cm³ was greater than 90 % of the prescription dose. The minimum dose to 1 cm³ of the target volume was used in preference to D98% (minimum dose to 98 % of the target volume) as this is independent of the volume of the region of interest and in accordance with our standard departmental policy. Planning objectives for the OARs are given in Table 2.1.

The first and repeat (second) CT scans were fused with alignment to bony landmarks (C2/C3 vertebrae) and soft tissue in proximity to the tumour. Target volume and OAR contours were then transferred from the first to the second CT scans. These were reviewed and modified by the same radiation oncologist who performed the initial contouring on the first CT scan. To reduce bias, the revised contours were independently assessed by a second radiation oncologist.
Table 2.1: Organ at risk treatment planning objectives.

<table>
<thead>
<tr>
<th>Organ at risk</th>
<th>Dose constraint</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spinal cord</td>
<td>Max 1 cm³ ≤ 45 Gy</td>
</tr>
<tr>
<td>Brainstem</td>
<td>Max 1 cm³ ≤ 50 Gy</td>
</tr>
<tr>
<td>Contralateral parotid</td>
<td>Mean &lt; 26 Gy</td>
</tr>
<tr>
<td>Larynx</td>
<td>Mean &lt; 45 Gy</td>
</tr>
<tr>
<td>Oral cavity</td>
<td>Mean &lt; 45 Gy</td>
</tr>
</tbody>
</table>

CTV1 was not adjusted, except to account for natural barriers to disease spread including bone, air and fascia; this is in keeping with standard practice not to adapt the therapeutic target volume to tumour response. CTV2, CTV3, and OAR contours were amended to account for anatomic and positional alterations (i.e. redefined to conform to anatomic definitions/boundaries). The PTVs were regrown, again constrained 5 mm from the skin surface to account for the dose buildup region. The initial IMRT and VMAT plans (from the first CT scan) were applied to the second CT scan and doses to target volumes and OARs were recalculated without repeat optimisation.

2.2.3 Dosimetric comparisons

For both the IMRT and VMAT plans, dose-volume histograms (DVHs) were generated for target volumes and OARs. The changes in key dosimetric parameters to target volumes and OARs between first and second CT scans for IMRT and VMAT plans were compared by Wilcoxon matched pairs signed rank test. To account for multiple statistical comparisons and possible increase in type I error rate, differences were considered statistically significant at $p < 0.01$. 
2.3 Results

2.3.1 Patient characteristics

Table 2.2 includes patient, tumour and treatment characteristics. Of the 10 patients, there were 7 oropharyngeal, 2 hypopharyngeal, and 1 laryngeal squamous cell carcinoma, all of which were locally advanced disease. A total of 4 patients underwent neck dissection before radiotherapy. The median time to replanning was fraction 20, and mean weight loss from start of radiotherapy to replanning was 8.9 %, with a standard deviation of 2.8 %.

2.3.2 Dosimetric comparisons

Mean changes in volumes of CTV1, CTV2, and CTV3 between first and second CT scans are shown in Table 2.3. The mean D95% for PTV1, PTV2 and PTV3 deteriorated between first and second CT scans for both IMRT and VMAT plans (Table 2.4). All except one IMRT plan did not meet the target D95% for PTV coverage for the second CT scan. The mean D95% for CTV1, CTV2 and CTV3 was maintained between first and second CT scans, but for CTV2 and CTV3 the mean D_min was reduced for both IMRT and VMAT plans (Table 2.4).

There were no significant differences between VMAT and IMRT plans for change in dosimetric parameters between the first and second CT scans for any target volume (Table 2.5). The mean differences for change in D95% were 0.1 Gy, 0.3 Gy and 0.5 Gy for CTV1, CTV2 and CTV3, respectively. The mean differences in D_min were 0.6 Gy, 0.6 Gy and 0.7 Gy for CTV1, CTV2 and CTV, respectively. These were not statistically significant.

For the OARs, there were no clinically relevant changes in doses to spinal cord, brainstem, larynx, or oral cavity between the first and second CT scans for either
Table 2.2: Patient, tumour and treatment characteristics.

<table>
<thead>
<tr>
<th>Sex</th>
<th>Age</th>
<th>ECOG PS</th>
<th>Primary site</th>
<th>Stage</th>
<th>Chemotherapy</th>
<th>Replan fraction</th>
<th>Weight loss / %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>53</td>
<td>2</td>
<td>OPX</td>
<td>T4N2c</td>
<td>Synchronous carboplatin</td>
<td>21</td>
<td>12.3</td>
</tr>
<tr>
<td>Female</td>
<td>59</td>
<td>1</td>
<td>OPX</td>
<td>T2pN2b</td>
<td>None</td>
<td>13</td>
<td>3.4</td>
</tr>
<tr>
<td>Male</td>
<td>58</td>
<td>0</td>
<td>LX</td>
<td>T2pN2c</td>
<td>Synchronous cisplatin</td>
<td>21</td>
<td>12.8</td>
</tr>
<tr>
<td>Male</td>
<td>66</td>
<td>1</td>
<td>HPX</td>
<td>T3N2b</td>
<td>Induction docetaxel, cisplatin, 5-fluorouracil synchronous cisplatin</td>
<td>4</td>
<td>8.4</td>
</tr>
<tr>
<td>Male</td>
<td>48</td>
<td>0</td>
<td>OPX</td>
<td>T3N2b</td>
<td>Induction docetaxel, cisplatin, 5-fluorouracil synchronous cisplatin</td>
<td>23</td>
<td>10.4</td>
</tr>
<tr>
<td>Male</td>
<td>45</td>
<td>1</td>
<td>OPX</td>
<td>T2N2c</td>
<td>Induction docetaxel, cisplatin, 5-fluorouracil synchronous cisplatin</td>
<td>22</td>
<td>9.9</td>
</tr>
<tr>
<td>Male</td>
<td>53</td>
<td>2</td>
<td>HPX</td>
<td>T2pN2b</td>
<td>None</td>
<td>19</td>
<td>9.6</td>
</tr>
<tr>
<td>Male</td>
<td>65</td>
<td>1</td>
<td>OPX</td>
<td>T4pN2b</td>
<td>None</td>
<td>22</td>
<td>8.1</td>
</tr>
<tr>
<td>Female</td>
<td>45</td>
<td>1</td>
<td>OPX</td>
<td>T2N2c</td>
<td>Induction docetaxel, cisplatin, 5-fluorouracil synchronous cisplatin</td>
<td>4</td>
<td>8.2</td>
</tr>
<tr>
<td>Male</td>
<td>73</td>
<td>2</td>
<td>OPX</td>
<td>T3N2c</td>
<td>Synchronous cetuximab</td>
<td>21</td>
<td>6.2</td>
</tr>
</tbody>
</table>

Abbreviations: ECOG PS = Eastern Cooperative Oncology Group Performance Status; OPX = oropharynx; LX = larynx; HPX = hypopharynx; T = tumour; N = node; p = pathological stage (post-neck dissection).
Table 2.3: Changes in clinical target volumes between first (CT1) and repeat (CT2) CT scans.

<table>
<thead>
<tr>
<th></th>
<th>Mean CT1 volume / cm³</th>
<th>Mean CT2 volume / cm³</th>
<th>Mean volume change (SD) / %</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTV1</td>
<td>275</td>
<td>247</td>
<td>-7.7 (7.2)</td>
</tr>
<tr>
<td>CTV2</td>
<td>311</td>
<td>290</td>
<td>-4.4 (9.8)</td>
</tr>
<tr>
<td>CTV3</td>
<td>143</td>
<td>139</td>
<td>-6.7 (15.1)</td>
</tr>
</tbody>
</table>

Abbreviations: SD = standard deviation.

IMRT or VMAT plans (Table 2.6). However, the mean doses to the ipsilateral and contralateral parotid glands increased by 3.1 Gy (7.7 %) and 2.5 Gy (10.4 %), respectively, for IMRT plans and 3.5 Gy (8.6 %) and 2.8 Gy (11.9 %) for VMAT plans. For one patient who was replanned at fraction 22, the summed cumulative dose to the contralateral parotid gland exceeded the mean dose constraint of 26 Gy by 2.9 Gy and 2.8 Gy for the IMRT and VMAT plans, respectively. There were no statistically significant differences between IMRT and VMAT plans for change in doses to any OARs between the first and the second CT scans (Table 2.7).

2.4 Discussion

This study included 10 patients with locally advanced head and neck cancer who required repeat CT simulation and replanning during a course of IMRT. It evaluated the relative change in target volume coverage and doses to OARs between step-and-shoot IMRT and VMAT plans resulting from weight loss during head and neck radiotherapy. The results showed that although there was some deterioration in target coverage and an increase in dose to the parotid glands between the initial and repeat CT scans, no statistically significant differences were observed in these changes between IMRT and VMAT.

In recent years, there have been a number of planning comparative studies between IMRT and VMAT (Bertelsen et al., 2010; Wiehle et al., 2011; Lu et al.,
Table 2.4: Changes in target volume coverage for intensity modulated radiotherapy (IMRT) and volumetric modulated arc therapy (VMAT) plans between first (CT1) and repeat (CT2) CT scans.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>IMRT CT1 Mean (SD) / Gy</th>
<th>IMRT CT2 Mean (SD) / Gy</th>
<th>Difference / Gy</th>
<th>VMAT CT1 Mean (SD) / Gy</th>
<th>VMAT CT2 Mean (SD) / Gy</th>
<th>Difference / Gy</th>
</tr>
</thead>
<tbody>
<tr>
<td>PTV1 D95%</td>
<td>63.7 (0.6)</td>
<td>62.8 (2.0)</td>
<td>-0.9</td>
<td>64.3 (0.5)</td>
<td>63.4 (2.0)</td>
<td>-0.9</td>
</tr>
<tr>
<td>PTV2 D95%</td>
<td>57.9 (0.3)</td>
<td>54.2 (2.5)</td>
<td>-3.7</td>
<td>58.2 (0.3)</td>
<td>54.3 (3.5)</td>
<td>-3.8</td>
</tr>
<tr>
<td>PTV3 D95%</td>
<td>52.7 (0.5)</td>
<td>49.7 (2.7)</td>
<td>-3.0</td>
<td>52.4 (0.6)</td>
<td>49.5 (2.5)</td>
<td>-2.9</td>
</tr>
<tr>
<td>CTV1 D95%</td>
<td>64.9 (0.4)</td>
<td>64.7 (1.3)</td>
<td>-0.2</td>
<td>65.0 (0.3)</td>
<td>64.9 (1.5)</td>
<td>-0.1</td>
</tr>
<tr>
<td></td>
<td>63.0 (1.8)</td>
<td>62.1 (2.6)</td>
<td>-0.9</td>
<td>62.6 (2.1)</td>
<td>62.4 (3.1)</td>
<td>-0.3</td>
</tr>
<tr>
<td>CTV2 D95%</td>
<td>59.8 (0.4)</td>
<td>59.0 (0.8)</td>
<td>-0.8</td>
<td>59.6 (0.4)</td>
<td>59.1 (0.9)</td>
<td>-0.5</td>
</tr>
<tr>
<td></td>
<td>57.8 (1.0)</td>
<td>53.9 (3.9)</td>
<td>-3.9</td>
<td>57.6 (1.1)</td>
<td>54.2 (4.6)</td>
<td>-3.4</td>
</tr>
<tr>
<td>CTV3 D95%</td>
<td>53.6 (0.4)</td>
<td>52.6 (1.0)</td>
<td>-0.9</td>
<td>53.0 (0.8)</td>
<td>52.6 (1.0)</td>
<td>-0.5</td>
</tr>
<tr>
<td></td>
<td>52.9 (0.5)</td>
<td>50.4 (2.5)</td>
<td>-2.6</td>
<td>52.1 (1.4)</td>
<td>50.2 (2.2)</td>
<td>-1.9</td>
</tr>
</tbody>
</table>

**Abbreviations:** D95% = minimum dose to 95% of the treatment volume; D_min = minimum dose to 1 cm³; SD = standard deviation.
Table 2.5: Differences between the changes in target volume coverage for intensity modulated radiotherapy (IMRT) and volumetric modulated arc therapy (VMAT).

<table>
<thead>
<tr>
<th>Parameter / Gy</th>
<th>VMAT vs IMRT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Difference / Gy</td>
</tr>
<tr>
<td>PTV1 D95%</td>
<td>0.0</td>
</tr>
<tr>
<td>PTV2 D95%</td>
<td>-0.1</td>
</tr>
<tr>
<td>PTV3 D95%</td>
<td>0.1</td>
</tr>
<tr>
<td>CTV1 D95%</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>D_min</td>
</tr>
<tr>
<td>CTV2 D95%</td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td>D_min</td>
</tr>
<tr>
<td>CTV3 D95%</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>D_min</td>
</tr>
</tbody>
</table>

Abbreviations: D95% = minimum dose to 95 % of the treatment volume; D_{min} = minimum dose to 1 cm^3.

2012). Although potentially subject to biases including experience of planners, treatment planning systems and optimisation algorithms, they have suggested that VMAT offers improved target volume conformity and equivalent or improved sparing of OARs relative to IMRT. This, coupled with improved treatment efficiency has led to increased adoption of VMAT (Teoh et al., 2011).

The relative dosimetric effect of weight loss or gain between IMRT and VMAT for patients with prostate cancer has previously been evaluated. In a study of 10 patients, the effect was modelled by variation in the source-to-surface distance (SSD) (Pair et al., 2013). Comparisons were made between IMRT using a 9-beam arrangement, and VMAT using two 300° arcs (excluding posterior gantry angles for technical reasons). Alteration in the SSD of 1 cm for IMRT and VMAT plans correlated with a respective 2.9 % and 3.6 % change in target mean dose. This difference may have arisen because the VMAT plans did not include posterior
Table 2.6: Changes in doses to organs at risk for intensity modulated radiotherapy (IMRT) and volumetric modulated arc therapy (VMAT) between first (CT1) and repeat (CT2) CT scans.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>IMRT CT1 Mean (SD) / Gy</th>
<th>IMRT CT2 Mean (SD) / Gy</th>
<th>Difference / Gy</th>
<th>VMAT CT1 Mean (SD) / Gy</th>
<th>VMAT CT2 Mean (SD) / Gy</th>
<th>Difference / Gy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spinal cord</td>
<td>D&lt;sub&gt;max&lt;/sub&gt; 39.6 (1.3)</td>
<td>40.6 (1.7)</td>
<td>1.0</td>
<td>D&lt;sub&gt;max&lt;/sub&gt; 37.4 (2.5)</td>
<td>38.0 (3.5)</td>
<td>0.6</td>
</tr>
<tr>
<td>Brainstem</td>
<td>D&lt;sub&gt;max&lt;/sub&gt; 43.2 (4.4)</td>
<td>42.8 (5.3)</td>
<td>-0.3</td>
<td>D&lt;sub&gt;max&lt;/sub&gt; 42.3 (2.6)</td>
<td>42.1 (4.3)</td>
<td>-0.3</td>
</tr>
<tr>
<td>Parotid, ipsi</td>
<td>D&lt;sub&gt;mean&lt;/sub&gt; 40.4 (11.9)</td>
<td>43.6 (12.0)</td>
<td>3.1</td>
<td>D&lt;sub&gt;mean&lt;/sub&gt; 40.8 (12.4)</td>
<td>44.3 (12.1)</td>
<td>3.5</td>
</tr>
<tr>
<td>Parotid, contra</td>
<td>D&lt;sub&gt;mean&lt;/sub&gt; 24.0 (2.7)</td>
<td>26.6 (4.2)</td>
<td>2.5</td>
<td>D&lt;sub&gt;mean&lt;/sub&gt; 23.6 (1.9)</td>
<td>26.4 (4.7)</td>
<td>2.8</td>
</tr>
<tr>
<td>Larynx</td>
<td>D&lt;sub&gt;mean&lt;/sub&gt; 62.2 (4.3)</td>
<td>63.1 (4.4)</td>
<td>0.9</td>
<td>D&lt;sub&gt;mean&lt;/sub&gt; 62.2 (4.2)</td>
<td>63.3 (4.4)</td>
<td>1.0</td>
</tr>
<tr>
<td>Oral cavity</td>
<td>D&lt;sub&gt;mean&lt;/sub&gt; 44.6 (8.8)</td>
<td>45.0 (9.0)</td>
<td>0.4</td>
<td>D&lt;sub&gt;mean&lt;/sub&gt; 43.8 (9.1)</td>
<td>44.2 (9.5)</td>
<td>0.4</td>
</tr>
</tbody>
</table>

*Abbreviations:* Ipsi = ipsilateral; contra = contralateral; D<sub>max</sub> = maximum dose to 1 cm³; D<sub>mean</sub> = mean dose; SD = standard deviation.
Table 2.7: Differences between the changes in organ at risk doses for intensity modulated radiotherapy (IMRT) and volumetric modulated arc therapy (VMAT).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>/ Gy</th>
<th>VMAT vs IMRT</th>
<th>Difference / Gy</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spinal cord D&lt;sub&gt;max&lt;/sub&gt;</td>
<td>-0.4</td>
<td>0.33</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brainstem D&lt;sub&gt;max&lt;/sub&gt;</td>
<td>0.1</td>
<td>0.88</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parotid, ipsi D&lt;sub&gt;mean&lt;/sub&gt;</td>
<td>0.4</td>
<td>0.21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parotid, contra D&lt;sub&gt;mean&lt;/sub&gt;</td>
<td>0.3</td>
<td>0.22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Larynx D&lt;sub&gt;mean&lt;/sub&gt;</td>
<td>0.1</td>
<td>0.80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oral cavity D&lt;sub&gt;mean&lt;/sub&gt;</td>
<td>0.0</td>
<td>0.65</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Abbreviations: Ipsi = ipsilateral; contra = contralateral; D<sub>max</sub> = maximum dose to 1 cm<sup>3</sup>; D<sub>mean</sub> = mean dose.

Gantry angles. In a similar study, 7-field IMRT was compared with single 360° arc VMAT for 5 patients (Chow and Jiang, 2013). Weight loss was simulated by decreasing the body contour by 0.5 cm to 2.0 cm from the surface in the anterior and lateral directions. There were increases in D99% to both the PTV and CTV of 4.0 % and 2.7 % per cm, and the D30% to the rectum of 4.0 % and 2.2 % per cm for IMRT and VMAT plans, respectively.

This study is the first to determine whether inter-fractional anatomic alterations have a differential effect on the dose delivered by IMRT and VMAT for head and neck cancer. In theory, the increased degrees of freedom afforded by rotational delivery may increase plan robustness. However, the increased conformity of VMAT may increase susceptibility to positional and anatomic alterations. Our data showed that IMRT with 7 equally spaced beams and VMAT with two 360° arcs resulted in similar plan robustness with no statistically significant differences between techniques in changes in dosimetric parameters for target volume coverage or doses to OARs.

A total of 9 patients were replanned because of weight loss, with one patient
replanned because of anatomic alterations that resulted in a shift of the spinal cord towards the high dose region. However, the main limitation of this study was that the dosimetric changes that triggered replanning were not evaluable. For the purpose of this study, it would have been preferred if patients had also been rescanned in their initial immobilisation mask. However, to account for this, we used rigid registration and bone alignment between the first and second CT scans to control for positional uncertainties. We demonstrated dosimetric changes between the first and the second CT scans for both IMRT and VMAT; however, the aim of this study was to determine the relative changes in dosimetric parameters between IMRT and VMAT.

The patients in this study were highly selected and included only those who required replanning during treatment. Therefore, the absolute dosimetric differences between the first and second CT simulations are not generalisable to all patients, and we do not present cumulative dosimetric data. However, it is of interest that even in this selected group with marked anatomic changes that required adaptive replanning, there were no clinically relevant changes in D95% for CTV coverage or, except parotid glands, doses to OARs seen for either technique. This suggests that the benefit from adaptive replanning is to recover the increased delivered dose to the parotid glands. As this was a retrospective study that included small patient numbers, we did not perform secondary analyses to investigate the relationship between dosimetric changes and anatomic factors.

The results of this study have shown that the dosimetric changes that occur as a result of anatomic changes during head and neck radiotherapy are equivalent for IMRT and VMAT. Plan adaptation strategies are based on typical dosimetric changes, and so these results indicate that the number and frequency of replans would be equivalent for IMRT and VMAT.
2.5 Conclusion

There is no significant difference in plan robustness between IMRT and VMAT plans in head and neck cancer patients who required replanning due to weight loss. The frequency of plan adaptation for these patients is therefore equal for IMRT and VMAT.
3 The suitability of common metrics for assessing parotid and larynx automatic segmentation accuracy

Published in Journal of Applied Clinical Medical Physics, 2016, 17(2), pp. 41–9

Authors

William J Beasley\textsuperscript{1,2}, Alan McWilliam\textsuperscript{1}, Adam Aitkenhead\textsuperscript{1,2}, Ranald I Mackay\textsuperscript{1,2}, Carl G Rowbottom\textsuperscript{1,3}

\textsuperscript{1}Division of Molecular and Clinical Cancer Sciences, School of Medical Sciences, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester, UK
\textsuperscript{2}Christie Medical Physics and Engineering (CMPE), The Christie NHS Foundation Trust, Manchester, UK
\textsuperscript{3}The Clatterbridge Cancer Centre, NHS Foundation trust, Bebington, UK

Author contributions

I, along with my co-authors, developed the strategy for measuring the suitability of the different metrics. I performed the experimental work, which included creating treatment plans, automatic segmentation using SPICE, and creation of the STAPLE volumes. I modified Matlab scripts for measuring segmentation accuracy that were initially created by AA. I performed the statistical analyses and generated the conclusions along with all co-authors. I wrote the paper, subject to minor modifications from co-authors.
Abstract

Contouring structures in the head and neck is time-consuming and automatic segmentation is an important part of an adaptive radiotherapy workflow. Geometric accuracy of automatic segmentation algorithms has been widely reported, but there is no consensus as to which metrics provide clinically meaningful results. This study investigated whether geometric accuracy (as quantified by several commonly used metrics) was associated with dosimetric differences for the parotid and larynx, comparing automatically generated contours against manually-drawn ground truth contours. This enabled the suitability of different commonly used metrics to be assessed for measuring automatic segmentation accuracy of the parotid and larynx. Parotid and larynx structures for ten head and neck patients were outlined by five clinicians to create ground truth structures. An automatic segmentation algorithm was used to create automatically generated normal structures, which were then used to create volumetric modulated arc therapy plans. The mean doses to the automatically generated structures were compared with those of the corresponding ground truth structures, and the relative difference in mean dose was calculated for each structure. It was found that this difference did not correlate with the geometric accuracy provided by several metrics, notably the Dice similarity coefficient, which is a commonly used measure of spatial overlap. Surface-based metrics provided stronger correlation, and are therefore more suitable for assessing automatic segmentation of the parotid and larynx.
3.1 Introduction

Intensity modulated radiotherapy (IMRT) and volumetric modulated arc therapy (VMAT) are capable of creating highly conformal treatment plans, with steep dose gradients providing efficient organ at risk (OAR) sparing (Cozzi et al., 2004; Vanetti et al., 2009). IMRT has been shown to benefit patients in the head and neck (Chao et al., 2001; Saarilahti et al., 2005, 2006), with the PARSPECT trial demonstrating reduced incidence of xerostomia in patients treated with parotid-sparing IMRT relative to those treated with conformal radiotherapy (Nutting et al., 2011). However, in order to realise the benefits afforded by IMRT, accurate delineation of targets and normal structures is essential (Stapleford et al., 2010).

Contouring in the head and neck is time-consuming and labour-intensive (Miles et al., 2005; Harari et al., 2010), but automatic segmentation has shown potential to reduce inter-observer variation and improve efficiency by reducing the time required for outlining (Chao et al., 2007; Stapleford et al., 2010; Walker et al., 2014). This is of particular benefit to adaptive radiotherapy (ART), and there has therefore been much interest in automatic segmentation, with several algorithms having been assessed for accuracy (Zhang et al., 2007; Al-Mayah et al., 2010; Stapleford et al., 2010; Thomson et al., 2014; Walker et al., 2014).

In such studies, the accuracy of automatic segmentation algorithms has been assessed by measuring the geometric agreement between automatically generated structures and ‘ground truth’ structures provided by manual delineation. A wide variety of metrics has been reported in the literature, and can be broadly separated into volume-based and surface-based metrics. Volume-based metrics, such as the Dice similarity coefficient (DSC), which measures the spatial overlap of two volumes (see Figure 3.1), and the conformity index (CI), which measures the relative difference in volumes, are commonly used (Dice, 1945; Simmat et al., 2012; Veiga et al., 2014). Whilst these metrics are relatively simple to understand,
Figure 3.1: DSC and DTA. DSC measures the spatial overlap between two volumes, and DTA describes the shortest distance between two surfaces for a specific point.

they are difficult to interpret and are sensitive to the volumes of the structures being assessed (Sharp et al., 2014; Sykes, 2014). Surface-based metrics provide a quantitative measure of the concordance of two surfaces, and are typically based on distance-to-agreement (DTA). DTA is calculated by computing the minimum distance from a point on a reference surface to any point on a target surface (see Figure 3.1), which is repeated for all points on the reference surface. From this a DTA-histogram can be produced. Several different metrics can be derived from this DTA-histogram, and some of the most commonly-reported include the mean- and maximum-DTA (Hoffmann et al., 2014; Thomson et al., 2014) and the 95%-Hausdorff distance (95%-HD) (Hou et al., 2011; Huger et al., 2014), which is defined as the 95th percentile of the DTA-histogram. Although these metrics provide a measure of the distance between two structures, they too can be difficult to translate into clinical relevance (Sykes, 2014).

To parallel the term ‘geometric accuracy’, which quantifies the spatial agreement of two different structures, we introduce the term ‘dosimetric accuracy’ to quantify the difference in dose between two structures within a given dose distribution. In the case of automatic segmentation, the goal is to create automatically generated structures with high geometric accuracy relative to the ground truth.
Similarly, within the context of treatment planning and evaluation, in which an automatically generated contour might be used for treatment planning, it is also important that the dose reported to an automatically generated contour agrees with the dose reported to the corresponding ground truth structure (Nelms et al., 2012). Erroneous dose reporting may ultimately lead to a sub-optimal plan.

With such a variety of spatial metrics available, there is no consensus as to the most suitable metric for assessing geometric accuracy (Kumar et al., 2012; Sharp et al., 2014). As both the geometric and dosimetric accuracy are important for treatment planning and evaluation, it can be argued that suitable spatial metrics are those that provide results related to dosimetric accuracy (Tsuji et al., 2010). A geometrically accurate contour, as measured with a suitable spatial metric, should therefore be reflected in a small dosimetric difference, and vice versa.

In the present study, the geometric and dosimetric accuracy are measured for the parotid and larynx in head and neck VMAT treatment planning. The relationship between the geometric and dosimetric accuracy is measured, thus identifying suitable spatial metrics.

### 3.2 Method

Five clinicians outlined the parotids and larynx for ten head and neck cancer patients. These contours were created as part of a recent study at our institution assessing the geometric accuracy of a commercial automatic segmentation algorithm. The contouring has previously been described (Thomson et al., 2014), but is briefly outlined here. Contouring was performed according to locally-agreed protocols, and all observers contoured the structures independently, with access to the same clinical information. The observers were free to adjust the windowing and level according to personal preference. For each structure, the five clinician contours were combined into a single ground truth contour using the simultaneous
truth and performance level estimation (STAPLE) algorithm, which computes a probabilistic estimate of the ground truth from multiple segmentations of the same structure (Warfield et al., 2004). The resulting STAPLE contours were used as the reference standard (i.e. the ground truth) against which automatic contours were compared.

3.2.1 Dosimetric and geometric accuracy

For each patient, dual-arc 6 MV Elekta VMAT plans were retrospectively created using the Philips Pinnacle³ v9.6 treatment planning system (Philips Radiation Oncology Systems, Andover, MA), according to standard departmental protocols (see Table 3.1). Planning target volumes (PTVs) were created from a uniform 4 mm expansion of relevant clinical target volumes (CTVs), which had been drawn at the point of initial treatment; automatic segmentation of target volumes was not investigated. Automatically generated normal structures were created using the Philips Smart Probabilistic Contouring Engine (SPICE) software. These automatically generated contours were then used directly in the plan optimisation, with a 5 mm uniform margin applied to the spinal cord and brainstem to create planning organ at risk volumes (PRVs). The STAPLE contours for the parotids and larynx were imported into the treatment plan and the mean doses to these structures (the ‘true’ doses) were compared to those of the corresponding automatically generated contours. The mean dose was used as this is the dosimetric parameter of interest when assessing a treatment plan for these structures. The dosimetric accuracy was then defined as the percentage difference between the mean dose to the automatically generated and ground truth structures, relative to the dose to the ground truth structure.

In addition to measuring dosimetric accuracy for the automatically generated structures, the difference in mean dose to the individual clinician contours relative to the true dose (dose to the STAPLE contour) was also measured for each patient.
Table 3.1: OAR dose constraints used for creating the VMAT plans.

<table>
<thead>
<tr>
<th>Organ at risk</th>
<th>Dose constraint</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spinal cord PRV</td>
<td>Max $&lt; 48$ Gy</td>
</tr>
<tr>
<td></td>
<td>Max $1$ cm$^3 &lt; 45$ Gy</td>
</tr>
<tr>
<td>Brainstem PRV</td>
<td>Max $&lt; 54$ Gy</td>
</tr>
<tr>
<td></td>
<td>Max $1$ cm$^3 &lt; 50$ Gy</td>
</tr>
<tr>
<td>Contralateral parotid</td>
<td>Mean $&lt; 26$ Gy</td>
</tr>
<tr>
<td>Larynx</td>
<td>Mean $&lt; 45$ Gy</td>
</tr>
<tr>
<td>Oral cavity</td>
<td>Mean $&lt; 45$ Gy</td>
</tr>
</tbody>
</table>

This provided a measurement of the dosimetric inter-observer variation in mean dose for each patient, ultimately defining the range within which the dose to the automatically generated structure is acceptable.

A number of commonly used metrics were used to measure the geometric accuracy of the automatically generated contours relative to the ground truth structures, using an in-house MATLAB script. Two volume-based metrics were investigated: the conformity index (CI), which is the ratio of the volumes of the two structures; and DSC, which is a measure of the spatial overlap of two structures, defined as $DSC = 2(V_1 \cap V_2)/(|V_1|+|V_2|)$ (see Figure 3.1). The centroid separation, which is the magnitude of the distance between the centres of mass of two structures, was also measured.

The other metrics were based on the surface-agreement of two structures, and an in-house MATLAB script was used to calculate a DTA-histogram for each structure-pair. DTA is defined for a particular point on a reference surface, A, as the shortest distance to any point on surface B (see Figure 3.1). This is performed for each point on surface A, and a cumulative DTA-histogram is created. From this DTA-histogram, the mean and maximum DTA were measured, along with the 95%-Hausdorff distance (95%-HD), measuring the 95th percentile of the cumulative DTA-histogram.
3.2.2 Relationship between geometric and dosimetric accuracy

The correlation between the dosimetric accuracy and the different metrics was measured using the Pearson product-moment correlation coefficient; the strength of the correlation indicated the strength of the relationship between the geometric and dosimetric accuracy.

3.3 Results

3.3.1 Dosimetric and geometric accuracy

The mean dosimetric accuracy was measured to be $-4.8 \pm 3.4\%$ and $-8.4 \pm 2.3\%$ (dose to the automatically generated contours lower than that to the STAPLE contours) for the parotids and larynx, respectively. The uncertainties were estimated from the mean standard deviations in the inter-observer variation in mean dose, which provides an estimate of the uncertainty in the dose delivered to the ground truth contours. Figures 3.2 and 3.3 show boxplots of the dosimetric inter-observer variation for the individual parotid and larynx structures, respectively. The boxes indicate the interquartile range and the whiskers indicate the maximum and minimum range of variation in mean dose to the five clinician-drawn structures relative to the STAPLE contour (equal to the inter-observer variation). The dosimetric accuracy of the automatically generated contours is also indicated for each structure by the black circles; it can be seen that the dose to the automatically generated contour was outside the dosimetric inter-observer variation for sixteen out of the twenty parotid glands and nine of the ten larynx contours. Note that parotid 13 (Figure 3.2) had a dosimetric accuracy of $+43\%$. This was caused by the gland’s being in a region of low dose (mean doses of 376 cGy and 514 cGy to the STAPLE and automatically generated contour, respectively), resulting in a large relative difference in mean dose between the SPICE and STAPLE contours.
Figure 3.2: Dosimetric inter-observer variation for the parotids. Boxplot showing the inter-observer variation in dosimetric accuracy relative to the STAPLE contours for the parotid glands. Red boxes indicate right hand parotid glands and blue boxes indicate left hand glands. The boxes indicate the inter-quartile range, the whiskers indicate the minimum and maximum variation, and the horizontal lines indicate the median accuracy of the five clinician contours. The mean dosimetric accuracy of the automatically generated contours is indicated by the circles.

Similarly, parotid 14 also received a low mean dose (approximately 500 cGy), so the inter-observer variation was relatively large for this gland.

3.3.2 Relationship between geometric and dosimetric accuracy

Table 3.2 shows the correlation coefficients between the dosimetric accuracy and the various metrics. There was no correlation between the volume-based metrics (DSC and CI) and dosimetric accuracy for the parotids. Metrics based on surface
Figure 3.3: Dosimetric inter-observer variation for the larynx. Boxplot showing the inter-observer variation in dosimetric accuracy relative to the STAPLE contours for the larynx. The boxes indicate the inter-quartile range, the whiskers indicate the minimum and maximum variation, and the horizontal lines indicate the median accuracy of the five clinician contours. The mean dosimetric accuracy of the automatically generated contours is indicated by the circles.
Table 3.2: Correlation coefficients between the different metrics and the dosimetric accuracy. Statistical significance at p < 0.05 is indicated by *.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Parotid</th>
<th>Larynx</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSC</td>
<td>-0.35</td>
<td>-0.59*</td>
</tr>
<tr>
<td>CI</td>
<td>-0.33</td>
<td>0.58*</td>
</tr>
<tr>
<td>Centroid separation</td>
<td>0.82*</td>
<td>0.50</td>
</tr>
<tr>
<td>Max DTA</td>
<td>0.55*</td>
<td>0.60*</td>
</tr>
<tr>
<td>Mean DTA</td>
<td>0.69*</td>
<td>0.64*</td>
</tr>
<tr>
<td>95%-HD</td>
<td>0.61*</td>
<td>0.63*</td>
</tr>
</tbody>
</table>

agreement (DTA) showed statistically significant (p < 0.05) correlations with dosimetric accuracy, with mean DTA and 95%-HD showing strong correlation. The strongest correlate for the parotid was found to be the centroid separation. This can be seen in Figure 3.4, which shows scatter plots of the dosimetric accuracy as a function of centroid separation (left hand plot), for which correlation was strong and statistically significant, and DSC (right hand plot), for which correlation was weak and not statistically significant (p > 0.05).

Centroid separation did not correlate with dosimetric accuracy for the larynx, and weak correlation was observed for the volume-based metrics. Strong correlation was observed for the surface-based metrics.

3.4 Discussion

Automatic segmentation will be an essential component of treatment planning and ART, and it is important to assess the geometric accuracy of automatic segmentation algorithms before clinical implementation. There are many widely-accepted spatial metrics for assessing geometric accuracy, but there is no consensus as to the most appropriate metrics to use. For treatment planning and evaluation, it is also important that the dose to an automatically generated structure agrees with the dose to its corresponding ground truth structure. This work
proposes that an appropriate spatial metric is one that correlates with dosimetric accuracy, and aims to identify spatial metrics suitable for assessing automatic segmentation accuracy of the parotid and larynx in the head and neck.

The results have indicated that several commonly used geometric metrics do not correspond to dosimetric accuracy and are not suitable for assessing automatic segmentation performance for certain OARs in the head and neck. Notably, it has been shown that DSC is a poor surrogate for dosimetric accuracy for the parotids. This is highlighted by the fact that the mean parotid DSC in this study was 0.77, which is generally considered to be clinically acceptable (Mattiucci et al., 2013; Thomson et al., 2014), but the mean dose to the automatically generated structures was outside the range of dosimetric inter-observer variation for sixteen out of the twenty parotid glands investigated.

Although DSC and CI did not correlate with dosimetric accuracy for the parotids, the remaining metrics provided statistically significant correlations. The centroid separation, which is the magnitude of the distance between the centres
of mass of the automatically generated and STAPLE contours, provided the strongest correlation. This is likely explained by the fact that the parotids are often in close proximity to the target volume and in the region of a unidirectional steep dose gradient, such that global differences in organ position have a large effect on dosimetric accuracy. In contrast, the centroid separation did not correlate with dosimetric accuracy for the larynx. This was likely caused by the fact that the larynx is often in the region of several dose gradients, and so a global shift of position does not necessarily change the mean dose.

The surface-based metrics correlated with dosimetric accuracy for both the parotids and larynx, although the correlation for max DTA was weaker than for mean DTA and 95%-HD, probably due to the fact that a discrepancy in a single point on a surface does not necessarily have a large effect on the mean dose to that structure. Nevertheless, the correlation of the surface-based metrics with dosimetric accuracy suggests that these metrics are suitable for assessing automatic segmentation accuracy.

Whilst there have been many studies reporting the geometric accuracy of various automatic segmentation algorithms (Al-Mayah et al., 2010; Stapleford et al., 2010; Qazi et al., 2011; Thomson et al., 2014), there have been relatively few that have investigated the dosimetric effect of automatic segmentation uncertainties in head and neck IMRT. Tsuji et al. (2010) investigated the dosimetric accuracy in sixteen patients treated with head and neck IMRT. They compared the doses delivered to automatic and manual contours, and reported that the dosimetric differences were significant for the targets, but minor for the OARs. In contrast, Eiland et al. (2014) reported significant dosimetric differences between automatic and manual contours in seven head and neck IMRT plans, and concluded that automatic segmentation cannot yet replace manual delineation for treatment planning. This is in agreement with our findings, which show that
the dose delivered to automatic contours is generally outside the range of inter-observer variation.

The above studies measured the difference between dosimetric parameters for different structures in the head and neck for automatic and manual contours, using a single observer to define the ground truth. However, in our study, five clinicians outlined each structure, providing the inter-observer variation in dose for individual patients. This enabled a more realistic assessment of the acceptability of automatic contours, as the dose to an automatically generated structure could be compared to the inter-observer variation for the specific patient in question. Additionally, the use of a STAPLE volume provided a better estimate of the ground truth than for a single observer (Warfield et al., 2004).

Metrics suitable for assessing the geometric accuracy of automatic segmentation algorithms should be related to dosimetric accuracy, and there is no consensus as to which metrics are suitable for use in head and neck VMAT. Tsuji et al. (2010) investigated the relationship between geometric and dosimetric accuracy for two metrics: DSC and the overlap index (OI), which measures the proportion of the manual contour within the automatic contour. Although they reported correlation between GTV dosimetric agreement and the OI, there was no correlation for the OARs. This supports the results obtained in the present study, where the volume-based metrics did not correlate with differences in mean dose; however, the authors did not investigate surface-based metrics. Nelms et al. (2012) did investigate a surface-based metric, the ‘linear penalty’, which is a modified DTA giving more weight to larger contour discrepancies. Although they reported that this metric was related to dosimetric accuracy, the linear penalty is not commonly used; nevertheless, this supports our findings that surface-based metrics are suitable for measuring automatic segmentation accuracy.

In the present study, a single commercial automatic segmentation algorithm
was used to generate automatic contours. It should be emphasised that the re-
relationship between the geometric and dosimetric accuracy would be independent
of the specific automatic segmentation algorithm used.

The present study used a small dataset of ten head and neck patients to assess
the relationship between dosimetric and geometric accuracy for the parotid and
larynx. These structures are considered parallel organs, and the results cannot
be extrapolated to serial organs in the head and neck, such as the spinal cord and
brainstem. The location of these structures relative to typical dose gradients, as
well as the fact that the dosimetric parameter of interest is the maximum dose,
means that further work is required to determine which metrics are suitable for
assessing such serial organs.

Similarly, extrapolation of the results presented here to other treatment sites
should be performed with caution. By using the mean dose to quantify dosimet-
ric accuracy, a complex three dimensional dose distribution has been collapsed
into a single dosimetric parameter, disregarding any positional information about
the dose distribution. However, this was mitigated in our study by planning all
patients with the same head and neck VMAT class solution, such that the dose
distributions of all ten patients were similar. This means that the results presen-
ted here apply to the parotid and larynx when used for head and neck VMAT
treatment planning. For example, although it might be expected that metrics
useful for the parotid might also be useful for the rectum, as they are both in
close proximity to a target volume and are in a region of a single steep dose
gradient, further work would be required to verify this.

This study has assessed the relationship between geometric and dosimetric
accuracy for several spatial metrics commonly used for assessing automatic seg-
mentation accuracy. Specifically, the suitability of these metrics has been assessed
for the parotid and larynx in head and neck VMAT treatment planning. The res-
ults have indicated that the suitability of a spatial metric is dependent on the
structure. In particular, the common volume-based metrics, such as DSC and OI, are not related to dosimetric accuracy for the parotid, and only weakly related to dosimetric accuracy for the larynx. The surface-based metrics were related to dosimetric accuracy and therefore suitable for assessing automatic segmentation accuracy for both the parotids and larynx.

3.5 Conclusion

There are several spatial metrics available for assessing automatic segmentation accuracy, and there is no consensus on which metrics should be used. For treatment planning and evaluation, both geometric and dosimetric accuracy are important, as inaccurate contours can result in a sub-optimal treatment plan. A suitable spatial metric should therefore be related to dosimetric accuracy. This study has measured the relationship between geometric and dosimetric accuracy for several commonly used metrics in head and neck VMAT treatment planning. We found that this relationship is structure-dependent, and that there was no statistically significant relationship between volume-based metrics and dosimetric accuracy for the parotids, with only a weak correlation for the larynx. The surface-based metrics correlated with dosimetric accuracy for both structures, indicating that these metrics are more suitable measures of automatic segmentation accuracy of the parotid and larynx in the head and neck.
4 An automated workflow for patient-specific quality control of contour propagation

Published in Physics in Medicine and Biology, 2016, 61, pp. 8577–86

Authors

William J Beasley\textsuperscript{1,2}, Alan McWilliam\textsuperscript{1}, Nicholas J Slevin\textsuperscript{3}, Ranald I Mackay\textsuperscript{1,2}, Marcel van Herk\textsuperscript{1}

\textsuperscript{1}Division of Molecular and Clinical Cancer Sciences, School of Medical Sciences, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester, UK
\textsuperscript{2}Christie Medical Physics and Engineering (CMPE), The Christie NHS Foundation Trust, Manchester, UK
\textsuperscript{3}Department of Clinical Oncology, The Christie NHS Foundation trust, Manchester, UK

Author contributions

I, along with my co-authors, developed the workflow to assess contour propagation accuracy. I developed the methodology for testing the workflow and performed the experimental work, which included writing the Python scripts for the automated contour propagation, generation of noisy images, and building the logistic regression model. I performed the statistical analyses and wrote the paper, subject to minor changes from co-authors.
Abstract

Contour propagation is an essential component of adaptive radiotherapy, but current contour propagation algorithms are not yet sufficiently accurate to be used without manual supervision. Manual review of propagated contours is time-consuming, making routine implementation of real-time adaptive radiotherapy unrealistic. Automated methods of monitoring the performance of contour propagation algorithms are therefore required. We have developed an automated workflow for patient-specific quality control of contour propagation and validated it on a cohort of head and neck patients, on which parotids were outlined by two observers. Two types of error were simulated: mislabelling of contours; and introducing noise in the scans before propagation. The ability of the workflow to correctly predict the occurrence of errors was tested, taking both sets of observer contours as ground truth, using receiver operator characteristic analysis. The area under the curve was 0.90 and 0.85 for the observers, indicating good ability to predict the occurrence of errors. This tool could potentially be used to identify propagated contours that are likely to be incorrect, acting as a flag for manual review of these contours. This would make contour propagation more efficient, facilitating the routine implementation of adaptive radiotherapy.
4.1 Introduction

Intensity modulated radiotherapy (IMRT) and volumetric modulated arc therapy (VMAT) have become an integral part of modern radiotherapy (Cozzi et al., 2004; Palma et al., 2008; Vanetti et al., 2009). Their ability to create highly conformal dose distributions has enabled the dose to organs at risk (OARs), and therefore radiation-induced toxicities, to be reduced (Nutting et al., 2011; Gulliford et al., 2012). Accurate delineation of OARs is therefore essential to fully realise the benefits afforded by IMRT and VMAT.

However, significant anatomic changes can occur during radiotherapy (Muren et al., 2003; Fiorino et al., 2005; Hong et al., 2005), and the steep dose gradients present in IMRT and VMAT mean that treatment plans can be particularly sensitive to such changes. As a result, a treatment plan based on anatomy imaged before treatment is often sub-optimal (Hansen et al., 2006; Bhide et al., 2010; Nishi et al., 2013). Adaptive radiotherapy (ART), in which a treatment plan is modified in response to anatomic changes, can be used to mitigate resulting dosimetric consequences (Yan et al., 1997).

On-treatment imaging is an integral component of ART, and accurate delineation of structures on these images is important. Manual delineation is impractical, and so automatic segmentation is essential to enable ART to be introduced into routine clinical use (Zhang et al., 2007; Muren and Thwaites, 2013). The most common form of on-treatment automatic segmentation is contour propagation (Thor et al., 2011; Hardcastle et al., 2013; Kumarasiri et al., 2014). For this, an on-treatment image is non-rigidly registered with the planning computed tomography (pCT) image, and the resulting deformation vector field is used to deform the contours from the pCT to the new anatomy of the on-treatment image.

Initial validation of a contour propagation algorithm is essential before clinical use, and several algorithms have been evaluated (Thor et al., 2011; Hou et al.,
2011; Hardcastle et al., 2013). However, it is generally considered that they cannot yet replace manual contour review. Furthermore, initial validation cannot guarantee the absence of contour propagation failures in routine clinical use. For contour propagation to be implemented into routine use, patient-specific quality control (QC) of contour propagation is essential. This would facilitate routine implementation of ART.

We present here an automated workflow for real-time patient-specific QC of contour propagation. This workflow monitors contour propagation quality, identifying situations in which the propagated contours are more likely to be subject to large uncertainties. It is validated on a cohort of head and neck cancer patients with two different sources of propagation error.

4.2 Method

4.2.1 Automated workflow for contour propagation QC

Figure 4.1 illustrates the automated workflow for contour propagation QC. The workflow requires two sets of ground truth contours: one on the pCT, and one on an image taken on the first treatment fraction; in this case a cone beam CT (CBCT) (A). Note that the second ground truth can originate from contours propagated from the pCT. At treatment fraction $n$, contours are propagated from pCT onto CBCT$n$; it is these structures on which QC is performed (B). For this, these structures are propagated back onto CBCT1 (C), such that there are now two sets of structures on CBCT1. The concordance of these structures on CBCT1 is measured using the Dice similarity coefficient (DSC) and the mean distance-to-agreement (DTA). These “consistency metrics” are then used to infer the accuracy of the contours propagated onto CBCT$n$.

This workflow was implemented using an in-house Python (v 2.7) script and
Figure 4.1: Illustration of the contour QC workflow. Consistency metrics are calculated from the concordance of structures on CBCT1. These consistency metrics are used to infer the quality of contour propagation onto CBCTn.
ADMIRE (ADMIRE v 1.11, Elekta AB, Stockholm, Sweden), an automatic segmentation algorithm. The Python script was used to run the propagations in ADMIRE via a batch file, and to calculate the consistency metrics.

4.2.2 Image and contour data

Ten head and neck cancer patients who received weekly CBCT imaging as part of a previous study at our institution (Ho et al., 2012) were included in the study. Two observers (GT1 and GT2) independently contoured the parotids on the pCT and each weekly CBCT for each patient; these structures were taken as the ground truth. ADMIRE was used to propagate these ground truth parotids from the pCT onto each CBCT, and the accuracy of the propagations was measured with DSC and mean DTA. In addition to the accuracy of the propagated contours, the inter-observer variation was estimated from the concordance of the two sets of ground truth structures.

4.2.3 Workflow validation: error scenario I

The ability of the automated workflow to detect gross propagation errors was tested by copying contours to incorrect images for a subset of patients for a single observer (GT1). Propagated contours on CBCTs 3-6 were copied onto CBCT2, such that the contours on CBCT2 originated from a different image set. The automated workflow was performed on these structures and the consistency metrics were measured. The ability of the uncertainty metrics to identify these errors was investigated.

4.2.4 Workflow validation: error scenario II

For the second error scenario, Gaussian noise was added to the CBCT images (CBCT2–6) using an in-house Python script. Gaussian kernels with standard deviations of 20 HU, 100 HU, 300 HU, 500 HU and 1000 HU were used. Structures
from pCT were propagated onto these noisy images, and the propagation accuracy was measured by comparison with the manual delineations as before. The automated workflow was performed and the consistency metrics were evaluated for each propagation.

For this error scenario, the definition of an error was based on the accuracy of the propagation, using a threshold to determine whether a propagation error had occurred. This threshold was calculated for DSC and mean DTA relative to the ground truth CBCT contour, flagging anything further than three standard deviations from the mean inter-observer variation on pCT. A propagated contour with an accuracy that exceeded this error threshold for either DSC or mean DTA was therefore classified as an error. Note that the threshold was one-sided, such that only propagated contours with discrepancies larger than the mean inter-observer variation were classified as errors; any propagated contours with deviations smaller than the mean inter-observer variation were not classified as errors.

A logistic regression model was trained to detect these errors using the consistency metrics alone. This model was implemented using the Python library ‘scikit-learn’, an open-source machine learning library (Pedregosa et al., 2011). The data were pre-processed to ensure that the consistency metrics had a mean centred on zero and unity standard deviation. Stratified three-fold cross validation was used to split the data into training and test datasets; the model was trained on the training set and was then used to predict whether an error had occurred in the test dataset. The error predictions were compared with the known errors, and receiver operating characteristic (ROC) analysis was performed. The area under the curve (AUC), which provides a measure of the model performance, was calculated. This was performed for each iteration of the stratified three-fold cross validation, and the mean ROC curve and AUC were used to summarise the model performance. This was performed for both sets of ground truth contours.
Table 4.1: Mean accuracy and standard deviation of the propagated structures relative to the ground truth structures, and inter-observer variation for the CBCT images and pCT. Note that inter-observer variations calculated excluding the three patients with large discrepancies between observers are denoted with *.

<table>
<thead>
<tr>
<th>Propagation</th>
<th>DSC</th>
<th>Mean DTA / mm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Propagation accuracy (GT1)</td>
<td>0.82 ± 0.02</td>
<td>1.64 ± 0.26</td>
</tr>
<tr>
<td>Propagation accuracy (GT2)</td>
<td>0.79 ± 0.06</td>
<td>1.96 ± 0.43</td>
</tr>
<tr>
<td>Inter-observer variation (CBCT)</td>
<td>0.74 ± 0.05</td>
<td>3.52 ± 1.49</td>
</tr>
<tr>
<td>Inter-observer variation (CBCT)*</td>
<td>0.75 ± 0.06</td>
<td>3.05 ± 1.13</td>
</tr>
<tr>
<td>Inter-observer variation (pCT)</td>
<td>0.84 ± 0.03</td>
<td>2.20 ± 1.18</td>
</tr>
<tr>
<td>Inter-observer variation (pCT)*</td>
<td>0.86 ± 0.02</td>
<td>1.57 ± 0.21</td>
</tr>
</tbody>
</table>

independently.

4.3 Results

4.3.1 Propagation performance

The mean accuracy and standard deviations of the propagated contours, as well as the inter-observer variation, are shown in Table 4.1. Note that there were consistent discrepancies between the parotid contours drawn by the two observers for three patients (on pCT and all CBCTs); the inter-observer variations after excluding these patients are denoted with * in the table. It can be seen from the table that the accuracy and standard deviations of the propagated contours are better than those of the inter-observer variation, indicating good performance of ADMIRE for contour propagation.

4.3.2 Workflow validation: error scenario I

A plot of the consistency metrics from error scenario I is shown in Figure 4.2. The errors, illustrated as red stars, are clearly separate from the standard propagations with no errors (shown as green circles). This implies that the consistency metrics
can identify these types of gross error using a simple threshold.

4.3.3 Workflow validation: error scenario II

For error scenario II, noise was added to CBCT2–6 for each patient to reduce propagation performance and induce propagation errors. As expected, the accuracy of the contour propagation algorithm reduced with increasing levels of noise, as shown in Figure 4.3, albeit surprisingly slowly.

The automated workflow was applied to these noisy images, and a logistic regression model was trained to predict the occurrence of contour propagation errors. A contour propagation error was defined as any propagated contour with an accuracy (as quantified by DSC or mean DTA) more than three standard deviations from the mean inter-observer variation on the pCT (Table 4.1). Due
Figure 4.3: Propagation accuracy reduced with increasing noise.
to large differences between observers on three patients, these were excluded when calculating the error threshold. An error was therefore defined as any propagated contour with a mean DTA greater than 2.20 mm, or with a DSC less than 0.80.

Using these thresholds to define an error, stratified three-fold cross validated ROC analysis was performed using a logistic regression model, the results of which are illustrated in Figure 4.4. Shown are the resulting ROC curves for both sets of ground truth contours. The red dotted line represents the random guess line. The AUC for the two observers, GT1 and GT2, were 0.90 and 0.85, respectively. It is clear that the model works well for detecting these errors, although the performance was better for GT1 than for GT2. This is likely caused by the larger intra-observer variation in GT2, which can be inferred from the larger standard deviation in the propagation accuracy for GT2 in Table 4.1.

4.4 Discussion

Accurate automatic contour propagation is essential for clinical implementation of ART (Zhang et al., 2007). Although pre-clinical evaluation of an algorithm is important, successful contour propagation cannot be guaranteed in routine clinical use. It is therefore important to perform patient-specific QC on propagated contours, and we have introduced here an automated workflow for QC of contour propagation. This workflow uses consistency metrics, which quantify the consistency of contour propagation over multiple registrations, to monitor propagation performance.

We have shown that this workflow can be used for patient-specific QC of contour propagation, identifying simulated propagation errors. Automatic segmentation is not yet able to fully replace manual delineation, and to facilitate its introduction into clinical use, tools for monitoring propagation performance are important. To the best of the authors’ knowledge, no such tools have been
Figure 4.4: ROC curves for the three-fold validated logistic regression model for GT1 (a) and GT2 (b). The solid black line shows the average curve, and the red dotted line shows the random guess line. The AUC was 0.90 for GT1 and 0.85 for GT2.
The concept of automated contour review is not in itself a new concept, however. Machine learning has been suggested as a technique for detecting contouring errors (Kohlberger et al., 2012; McIntosh et al., 2013; Chen et al., 2015). Kohlberger et al. (2012) used linear and non-linear regression models to predict the underlying ground truth accuracy of manually-drawn contours using 42 intensity-based and geometric features. McIntosh et al. (2013) extended this concept with a random forest model that also included features describing the relative position of structures. They reported that their model could predict manual contouring errors with an AUC of 0.75. Chen et al. (2015) reported a similar method, which was implemented in a graphical user interface to highlight potentially erroneous contours.

These techniques show promise in identifying manually defined or automatically generated contouring errors. However, they are designed for baseline contour assessment, and do not explicitly test contour propagation. Monitoring the quality of the initial contours is important, as errors at baseline would be propagated onto on-treatment images. Such tests could therefore complement a workflow for monitoring contour propagation quality.

Contour propagation accuracy is closely linked to deformable image registration (DIR) accuracy. There are many metrics for measuring DIR uncertainty, and the automated workflow described here is similar in concept to the distance-to-discordant metric, described by Saleh et al. (2014). This metric measures the geometric uncertainty in each voxel over multiple non-rigid registrations, and they reported that it was related to known registration errors. Such methods could be useful for validating dose warping techniques, but the registration accuracy inside or outside a contour is not necessarily important for monitoring contour propagation accuracy. So tests for monitoring DIR quality may be unnecessarily strict for testing contour propagation accuracy.
The automated error detection workflow described here has been validated on a set of ten head and neck cancer patients, for whom two independent observers outlined the parotids on weekly CBCT images. The parotids were considered in the present study as they are known to undergo anatomic changes during radiotherapy, and the resulting increase in cumulative mean dose is often the reason for ART (Wu et al., 2009). In addition, contour propagation accuracy of the parotids was reasonable, allowing us to gradually introduce errors and test their effect. However, anatomic changes for the parotid are gradual, and so further work is required to assess the workflow for structures that experience different types of motion, such as the rectum or bladder, in which anatomic changes between fractions are more unpredictable.

The workflow has been implemented here with ADMIRE. This algorithm was chosen as its command line interface meant that it could be easily integrated with an in-house Python program, enabling automation of the QC process. Implementation of the workflow with an alternative algorithm should be possible, but its suitability should be carefully evaluated. Both inter- and intra-modality registration is required for the workflow (Figure 4.1), and any algorithm should be capable of performing both accurately. For example, an algorithm with poor CBCT-CBCT contour propagation performance would create more false positives due to errors introduced when creating the consistency metrics (step C in Figure 4.1).

Contour propagation accuracy was quantified with DSC and mean DTA, although many other metrics exist for measuring propagation accuracy. However, there is no consensus on which are the most appropriate to use. DSC and mean DTA have been used here as they are commonly-reported metrics, and there is evidence that mean DTA is related to discrepancies in the mean dose for the parotids (Beasley et al., 2016). However, inclusion of additional geometric metrics, as well as intensity-based metrics, could potentially improve the ability of the
workflow to detect propagation failures.

One of the difficulties in assessing an automatic contouring algorithm is the uncertainty in the ground truth contours. Manual segmentations are inherently subject to intra- and inter-observer variation (Nelms et al., 2012), which limits the perceived contour propagation accuracy. This uncertainty in the ground truth contours also limits the ability of the automated workflow described here to detect propagation failures. Uncertainty in the propagation accuracy translates into uncertainty in the designation of propagation failures as defined by a threshold on DSC and mean DTA. A larger uncertainty in the ground truth causes a larger uncertainty in the definition of propagation errors, limiting the ability to detect errors. This was apparent in our data when the logistic regression model was tested on the second observer (GT2). The propagation accuracy was lower and the variance higher for GT2 than for GT1, implying that there was a larger amount of uncertainty in GT2. Indeed, when the model was applied to GT2, the AUC was lower than for GT1. This uncertainty in the ground truth contours therefore limits our ability to test whether a model can correctly identify propagation errors.

Propagation errors were defined as any propagated contour with discrepancies in DSC or mean DTA more than three standard deviations from the mean inter-observer variation on pCT. The inter-observer variation on pCT was chosen because this represented the ‘true’ inter-observer variation; as the same observers outlined the parotids on both pCT and CBCT, the larger inter-observer variation on CBCT was likely caused by the poorer image quality of the CBCT images. However, the exact choice of error threshold does not affect the overall functionality of the logistic regression model, as training the model implicitly accounts for the choice of threshold used to classify an error.

The proposed workflow was validated using simulated errors, as the performance of the contour propagation software was acceptable for the available patients.
Two types of error were simulated: gross errors, in which the contours were effectively not propagated according to the underlying deformation vector field (error scenario I); and errors resulting from uncertainty in the propagation due to noise in the images (error scenario II). Although these errors are not necessarily realistic, they tested different sources of potential failure. Error scenario I simulated a grossly incorrect contour, and it would be expected that any error detection workflow would be capable of detecting such errors. In error scenario II, noise was added to the CBCT images; although noise of this type is not necessarily realistic, it introduces uncertainty into the registration and so the ability of the workflow to detect propagation uncertainty was tested. Nevertheless, further work is required to verify the workflow on clinically-observed errors.

Automatic contour propagation is essential for ART, but manual review of propagated contours is often necessary. A recent study described a workflow for online plan adaptation using magnetic resonance image guided radiotherapy (Acharya et al., 2016). Although they reported a reasonable time for plan adaptation (median time of 26 minutes), it was necessary for a clinician to manually review the propagated contours at each treatment fraction. This would be a barrier for routine ART implementation. The automated workflow we have presented here is a potential solution to this problem, as it could be used to highlight contours that are likely to be incorrect, acting as a flag for manual review by a clinician. For this, the model parameters would be optimised to obtain an appropriate balance between false positives (incorrect prediction of an error), resulting in unnecessary contour review, and false negatives (incorrect prediction of the absence of an error), resulting in an incorrect contour going unnoticed. This would therefore ensure that only contours with potential errors would be manually reviewed, improving ART efficiency.

Patient-specific QC of contour propagation is important to facilitate routine
implementation of ART, and the automated workflow described here shows potential as a tool for patient-specific QC of contour propagation, enabling ART to become more feasible.

4.5 Conclusion

Contour propagation is an essential component of ART, but unreliable propagation limits its routine clinical implementation. There are currently no tools to aide patient-specific QC of contour propagation. An automated workflow for patient-specific QC of contour propagation, based on consistency metrics calculated from multiple registrations, has been presented and tested on a set of ten head and neck patients with simulated propagation errors. This workflow shows potential as a tool for quality control of contour propagation, and could help facilitate the clinical implementation of adaptive radiotherapy.

Acknowledgement

ADMIRE was supplied by Elekta as part of a research agreement.
5 Permutation testing for image-based data mining

Authors

William J Beasley\textsuperscript{1,2}, Maria Thor\textsuperscript{3}, Andrew Green\textsuperscript{1}, Alan McWilliam\textsuperscript{1}, Ranald I Mackay\textsuperscript{1,2}, Nicholas J Slevin\textsuperscript{4}, Ziad Saleh\textsuperscript{3}, Nancy Y Lee\textsuperscript{5}, Nadeem Riaz\textsuperscript{3}, Cherry L Estilo\textsuperscript{6}, Caroline E Olsson\textsuperscript{7}, Caterina Finizia\textsuperscript{8}, Niclas Pettersson\textsuperscript{9}, Joseph O Deasy\textsuperscript{3}, Marcel van Herk\textsuperscript{1}

\textsuperscript{1}Division of Molecular and Clinical Cancer Sciences, School of Medical Sciences, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester, UK
\textsuperscript{2}Christie Medical Physics and Engineering (CMPE), The Christie NHS Foundation Trust, Manchester
\textsuperscript{3}Department of Medical Physics, Memorial Sloan Kettering Cancer Centre, New York, USA
\textsuperscript{4}Department of Clinical Oncology, The Christie NHS Foundation trust, Manchester, UK
\textsuperscript{5}Department of Radiation Oncology, Memorial Sloan Kettering Cancer Centre, New York, USA
\textsuperscript{6}Department of Surgery, Memorial Sloan Kettering Cancer Centre, New York, USA
\textsuperscript{7}Department of Radiation Physics, Institute of Clinical Sciences, The Sahlgrenska Academy at the University of Gothenburg, Gothenburg, Sweden
\textsuperscript{8}Department of Otorhinolaryngology, Institute of Clinical Sciences, The Sahlgrenska Academy at the University of Gothenburg, Gothenburg, Sweden
\textsuperscript{9}Department of Radiation Medicine and Applied Sciences, University of San Diego, La Jolla, USA

Author contributions

The concept of performing image-based data mining and permutation testing for continuously variable data was based on work previously performed by MvH for binary data. I performed the experimental work using simulated data to test the method. The statistic images for image-based data mining were created using code developed by AG, and I modified Lua scripts originally created by
MvH for permutation testing with an extreme statistic. I developed the Python code used for measuring the suprathreshold cluster size. My co-authors and I devised the strategy for applying this methodology to head and neck radiotherapy patients with respect to radiation-induced trismus. The data were supplied by The Sahlgrenska Academy at the University of Gothenburg, and I performed the experimental work and analysis whilst at Memorial Sloan Kettering Cancer Centre. I wrote the paper, subject to minor modifications from AM, NS, RM and MvH. This has not been published; however, it will be combined with work performed by MT, as well as future validation analyses, before submission to a peer-reviewed journal for publication.
Abstract

Image-based data mining enables detailed dose-response relationships to be identified from previously treated patients. However, it is subject to a multiple comparisons problem that can lead to false positives and incorrect identification of regions exhibiting a dose-response relationships. Permutation testing can be used to correct for this multiple comparisons problem for binary outcome data, but a technique of performing this correction for a continuously variable clinical endpoint has not previously been demonstrated. We have developed a method of correcting image-based data mining with continuously variable outcome data for multiple comparisons using permutation testing, and demonstrated it with simulated data and a cohort of head and neck cancer patients with respect to radiation-induced trismus. Two different methods permutation testing were compared using simulated data and it was found that methods based on the maximum size of clusters above a threshold were more powerful than those based on an extreme statistic. This technique was demonstrated on a cohort of head and neck cancer patients treated with radiotherapy. A region exhibiting a statistically significant dose-response relationship with the maximum mouth opening at 6 months post-radiotherapy was identified. This region overlapped with the ipsilateral masseter muscle, indicating that the dose to this structure is related to the severity of trismus. Sparing this structure during radiotherapy could therefore reduce the probability of developing trismus.
5.1 Introduction

The aim of radiotherapy is to deliver a high radiation dose to the tumour whilst minimising the dose to nearby sensitive structures; this often requires a balance between the probability of curing the disease and the potential for causing radiation-induced toxicities. Dose-response relationships, which characterise the relationship between the dose delivered to a particular anatomic structure and the probability of a particular clinical endpoint occurring, are essential tools for optimising the balance between the probability of cure and toxicity, and therefore radiotherapy. Dose-response relationships are typically derived from dose volume histogram (DVH) or dose-surface histogram (DSH) analyses, in which complex three dimensional (3D) dose distributions are collapsed into a single dosimetric parameter (such as mean dose), which is then related to a clinical endpoint. However, such analyses lose potentially informative spatial information, and also require prior hypotheses about the structures responsible for a dose-response relationship.

Image-based data mining is an alternative technique that can be used for interrogating dose distributions from previously treated patients to identify dose-response relationships with a particular clinical endpoint. Witte et al. (2010) mapped 352 prostate patients into a common frame of reference and calculated the average dose distributions for patients with and without treatment failure. By performing a voxel-wise t-test on the difference between these average images, they identified a region in the lymph nodes in which a lower dose was associated with an increased risk of treatment failure.

Heemsbergen et al. (2010) used a very similar technique on a group of 557 prostate cancer patients who had been randomised between 68 Gy or 78 Gy as part of a clinical trial. By mapping anatomic points on bladder wall based on the distance from the prostate and their angle relative to the centre of the
prostate, they created average dose maps of patients with and without urinary obstructions. They reported that patients who suffered urinary obstructions after 2 years received a higher dose to the trigonal area than those who did not. Palorini et al. (2016) reported a similar method for performing a pixel-wise analysis of bladder dose surface maps for prostate radiotherapy. Improta et al. (2016) applied this method to a group of 539 patients with respect to change in international prostate symptom score (IPSS) by the end of radiotherapy. They found that the dose to the trigonal region was associated with changes in IPSS $\geq 10$ and $\geq 15$.

Acosta et al. (2013) also performed a voxel-based approach for prostate radiotherapy, but used it for identifying regions associated with rectal bleeding. Performing global non-rigid registration to spatially normalise 105 patients to a common frame of reference, and creating average dose distributions for patients with and without rectal bleeding, they identified a region on the anterior wall of the rectum to which patients with rectal bleeding received a dose 6 Gy higher on average than those who did not suffer rectal bleeding.

Image-based data mining is appealing due to its ability to spatially localise organs or sub-volumes exhibiting potential dose-response relationships. However, it is subject to a multiple comparisons problem due to the large number of voxels in medical images (Nichols and Holmes, 2002). This means that the null hypothesis can be incorrectly rejected at some voxels, leading to false positives and the potential to incorrectly infer the presence of a structure exhibiting a dose-response relationship. Correction for this multiple comparisons problem is therefore essential.

Permutation testing has been suggested as a method of addressing this problem in image-based data mining and was introduced by Chen et al. (2013) in the context of radiotherapy. It is based on the principle that, for a given statistic map calculated from a set of images as part of an image-based mining analysis, the labelling of the images (i.e. the clinical endpoint associated with a particular
patient) is arbitrary under the null hypothesis. That is, the observed statistic map would be similar whatever the image labels. Evidence against the null hypothesis is therefore obtained by permuting the image labels and counting the proportion of permutations in which the the result is more extreme than that with the original labelling.

The use of permutation testing in neuroimaging is well-established, having been introduced in 1996 by Holmes et al. (1996). A detailed discussion of the theory of permutation testing with respect to neuroimaging has been provided by Nichols and Holmes (2002), who discuss the applications to functional magnetic resonance imaging and positron emission tomography. With regards to radiotherapy, permutation testing has been performed in only a small number of image-based data mining studies. Chen et al. (2013) introduced the concept for radiotherapy, and applied it to the data from Witte et al. (2010), confirming that their observations were statistically significant. The same group applied the method of permutation testing to an image-based data mining study of 475 prostate cancer patients with respect to acute gastrointestinal toxicity after radiotherapy (Wortel et al., 2015). Using permutation testing, they reported that the local dose to the rectum and anal walls were significantly different between patients with and without various gastrointestinal toxicities.

Recently, Palma et al. (2016) performed an image-based data mining study investigating radiation-induced lung damage in 98 patients treated for Hodgkin lymphoma. Spatially normalising all patients to a single reference frame, they calculated the average dose distributions for patients with and without radiation-induced lung damage. After permutation testing, they identified a statistically significant dose-response relationship, with patients who suffered radiation-induced lung damage received a higher dose to parenchymal regions than those who did not.

Image-based data mining in radiotherapy has currently only been applied to
categorical clinical endpoints in which patients can be split into two distinct
groups. However, clinical endpoints in radiotherapy can also be quantified by
a continuous variable. Image-based data mining and permutation testing with
continuously variable data has been applied to neuroimaging for identifying brain
regions connected with behavioural measures (Han et al., 2013), but have not been
applied in the context of radiotherapy.

For example, trismus is a common side effect of radiotherapy, affecting between
5 % and 38 % of patients (Dijkstra et al., 2004). It is characterised by reduced
mouth opening and risk factors include surgery, age, tumour location and dose to
the muscles of mastication (Hsieh et al., 2014; Lindblom et al., 2014; Pauli et al.,
2016; Rao et al., 2016). However, there is no consensus on which muscles exhibit
the strongest dose-response relationship and should therefore be preferentially
spared during radiotherapy.

We introduce here a means of performing permutation testing for image-based
data mining when the clinical endpoint is a continuous variable. We demonstrate
this principle with simulated data and compare two methods of performing per-
mutation testing in this context. Finally, we develop a methodology for applying
this technique within a clinical setting and apply it to a cohort of head and neck
patients with respect to radiation-induced trismus.

5.2 Method

5.2.1 Simulated data

In order to demonstrate the principle of permutation testing for image-based
data mining with a continuously variable clinical endpoint, simulated data were
created. Image sets, consisting of \( N \) two dimensional (2D) 128 x 128 images with
corresponding outcome data, were created. For each individual image within an
image set, a central region was assigned a uniform pixel value that increased
Figure 5.1: Simulated data consisted of a set of images whose central region increased uniformly, correlated to some outcome measure (a). Noise was added and the image was blurred with a Gaussian filter (b).

in a monotonic fashion across the image set. Noise, sampled from a Gaussian distribution, was added to the images, and each image was then blurred with a Gaussian kernel, simulating the spatial correlations between neighbouring voxels in dose distributions. An example of a simulated image with a 20x20 central correlation region is shown in Figure 5.1, with the image before (Figure 5.1a) and after (Figure 5.1b) the addition of the noise and subsequent blurring. Outcome data were simulated as a set of monotonically increasing integers from 0 to $N$, to which Gaussian noise was added to simulate different strengths of correlation within the central region. An image set therefore consisted of $N$ images, each with a corresponding outcome label.

5.2.1.1 Test statistic

The image-based data mining technique proposed here uses Spearman’s rank correlation as the test statistic. The procedure is illustrated in Figure 5.2, which shows an image set consisting of six individual images in the same frame of reference. For each pixel in the frame of reference, Spearman’s rank correlation
Figure 5.2: Image-based data mining for continuous variables. An image set consists of individual images with associated outcome data labels. Spearman’s rank correlation coefficient is calculated between the signal at each pixel and the corresponding outcome data associated with that image. This produces a map of correlation coefficients; in this example, the central pixel is strongly correlated with the outcome data, whereas the outer pixels are not.

Coefficient is calculated between the signal from all pixels in the same location and their corresponding outcome data label. These correlation maps were created using an in-house toolkit implemented in C++. In addition to the correlation map, an uncorrected p-value was calculated at each voxel using a two-tailed t-test, providing an uncorrected p-value map.

5.2.1.2 Permutation testing

To correct for multiple comparisons, permutation testing was performed. A permutation test requires several steps: 1) generate a statistic image; 2) form a null hypothesis; 3) determine the summary statistic; 4) permute the labels; 5) calculate the permutation distribution; 6) calculate a corrected p-value. These steps are explained below.
Statistic image. Spearman’s rank correlation coefficient was chosen as the statistic image, as it is non-parametric, making no assumptions about the underlying distribution of the dose at each voxel and its relation to outcome. Calculation of the statistic image is illustrated in Figure 5.2, with an example of the statistic image for the simulated data is shown in Figure 5.3b.

Null hypothesis. The null hypothesis for a permutation test is that there is no relationship between the signal (dose) at any voxel and the labelling (outcome data). So the observed statistic image (the correlation map) would be equally likely regardless of the ordering of the outcome data.

Summary statistic. A summary statistic is used in a permutation test to summarise evidence against the null hypothesis. The two most common forms of summary statistic are the extreme value in the statistic image, and the size of the largest cluster above a fixed threshold (Nichols and Holmes, 2002). The extreme value in the statistic image is taken to be the maximum or minimum value in the image; in this case, the maximum or minimum value of the correlation coefficient. The size of the largest cluster above a fixed threshold is an alternative summary statistic that is calculated by thresholding the statistic image at some fixed value to create a binary image, and then identifying clusters of neighbouring voxels. The size of the largest cluster in the image is known as the maximum suprathreshold cluster size \((STCS_{\text{max}})\) and is used to characterise evidence against the null hypothesis in the same way as the extreme value. Both types of summary statistic were investigated in the present study: the maximum correlation coefficient in the image \((R_{\text{max}})\); and the maximum suprathreshold cluster size \((STCS_{\text{max}})\).

Relabelling. In order to build evidence against the null hypothesis, the labels are permuted and a statistic image is calculated for each relabelling. This
randomisation was performed 1000 times, as this has been shown to provide an acceptable correction (Edgington, 1969).

**Permutation distribution and statistical significance.** The summary statistic is calculated for each relabelling to create the permutation distribution, as illustrated in Figure 5.3a for simulated data using $R_{\text{max}}$ as the summary statistic. The summary statistic from the original labelling ($R_0$) is also indicated. Under the null hypothesis, a summary statistic equal to $R_0$ would be equally likely for any labelling order, but it can be seen that after 1000 permutations $R_0$ is the most extreme summary statistic that was obtained. This provides strong evidence against the null hypothesis. The (corrected) p-value of the test is calculated from the proportion of the permuted summary statistics greater than or equal to $R_0$. In the example above, the corrected p-value of the test is $\frac{1}{1000}$, as no permuted summary statistic was larger than $R_0$. Furthermore, individual pixels can be classified as statistically significant at level $\alpha$ if they have a value in the statistic image greater than the critical value, $R_{\text{crit}}$, where $R_{\text{crit}}$ is the $c+1$ largest member of the permutation distribution calculated from $P$ permutations, where $c = \alpha P$. For example, for $\alpha = 0.05$, pixels with a signal greater than the 50th largest member of the permutation distribution are statistically significant. The corrected p-value for each pixel is the proportion of the permutation distribution greater than or equal to its value in the observed image. This is illustrated in Figure 5.3b, which shows a correlation map with statistically significant pixels ($\alpha = 0.05$) shown by the solid black line. Also shown in the figure by the dotted line is the statistically significant region calculated from the uncorrected p-value.
5.2.1.3 Comparison of permutation testing techniques

The simulated data were used to demonstrate permutation testing for a continuously variable clinical endpoint. Correlation images were calculated using an in-house toolkit, and permutation testing was implemented with Lua and Python scripts using 1000 permutations. The two summary statistics, $R_{max}$ and $STCS_{max}$, were compared, and their relative sensitivity to different sample sizes ($N$) and outcome data noise was investigated.

5.2.2 Application to radiation-induced trismus

A methodology for permutation testing and image-based data mining with continuously variable outcome data was developed and applied to a cohort of head and neck cancer patients with respect to radiation-induced trismus. Trismus was quantified by the maximum incisor-to-incisor distance (MID) at six months post-radiotherapy, which quantifies the ability of a patient to open their mouth.
5.2.2.1 Patient data

A total of 95 patients treated for head and neck cancer as part of a prospective trial investigating potential advantages of a mouth exercise intervention as part of radiotherapy were included in the analysis (Pauli et al., 2013). A summary of the available patient details is shown in Table 5.1. Patients received a mixture of 3D conformal radiotherapy and IMRT, delivering 64.6 Gy – 72.6 Gy to the primary target volume in 33 – 38 fractions using a simultaneous integrated boost technique or a two-phased approach. To account for the different radiotherapy prescriptions, dose distributions were converted into biologically effective dose (BED). The BED at voxel \( i \) is given by

\[
BED_i = nd_i \left( 1 + \frac{d_i}{\alpha/\beta} \right),
\]

where \( n \) is the number of fractions, \( d_i \) is the dose per fraction at voxel \( i \), and the constant \( \alpha/\beta \) was taken to be 3 across the entire patient. For patients treated in two phases, the BED distributions of individual phases were summed. For each patient the MID at baseline and at six months post-radiotherapy were available. Patients were excluded if they received an intervention before the 6 month MID measurement, or if they had an MID < 35 mm at baseline, which could indicate tumour-related trismus.

5.2.2.2 Workflow

The methodology for image-based data mining for continuously variable outcome data is illustrated in Figure 5.4. The first step of the process was to spatially normalise the BED distributions of each patient into a common frame of reference. For this, the CT images of each patient were non-rigidly registered to a reference patient, manually selected from the cohort, and the resulting deformation vector field (DVF) was applied to the corresponding BED image. For the deformable
Diagnosis
  Oral cavity 11
  Oropharynx 49
  Nasopharynx 13
  Other 22

Treatment approach
  Simultaneous integrated boost 47
  Two phase 45

High dose region
  Left 34
  Right 20
  Bilateral 41

Table 5.1: Summary of the patients included in the analysis.

image registration (DIR), air cavities were first masked from the CT images by setting the minimum value for voxels within the patient to 1000 HU. Rigid registration was performed using an automatically generated clip box covering the muscles of mastication and the mandible, before non-rigid registration using a fast free-form deformation method based on cubic B-splines (Modat et al., 2010) was implemented in NiftyReg, an open-source DIR platform. The spatially normalised BED distributions were combined into a four dimensional (4D) array and a Spearman’s rank correlation coefficient ($R_s$) map was calculated, using the MID at six months post-radiotherapy as the outcome variable.

In addition to this bulk analysis, this process was repeated for patients in whom the dose distribution in the region of the muscles of mastication was primarily on one side of the patient; the muscles of mastication could then be split into ipsilateral and contralateral muscles. The CT images and BED distributions for patients in which the ipsilateral muscles were on the right hand side of the patient were reflected through the sagittal plane, inverting the left and right coordinates. This ensured that all ipsilateral structures were on the left hand side
Figure 5.4: Methodology for image-based data mining. Images were non-rigidly registered to the reference image, and the dose distributions warped to a common frame of reference. The dose distribution was combined into a 4D array and a voxel-wise correlation was performed with continuously-variable outcome data to produce a map of correlation coefficients for each voxel. This map performs the input for the permutation testing as discussed in the main text.
of the patients, and all contralateral structures were on the right hand side. Spatial normalisation was then performed as above. This analysis was performed in two ways. First, only patients whose muscles of mastication could be classified as ipsilateral and contralateral were included; for this method the term ‘ipsi-contra analysis’ will be used. In the second, patients with bilateral disease were also included; for this the CT images and BED distributions were included both with and without being reflected through the sagittal plane. This ensured that all structures in which the dose was greatest (for the bilateral patients both structures received equally high doses) were on the left hand side of the patients, and the term ‘ipsilateral analysis’ will be used here. The left hand side of the patient was therefore the ipsilateral side, but the right hand side was a mixture of ipsilateral and contralateral, and so correlations on this side were not valid.

Permutation testing was performed for the bulk, ipsi-contra and ipsilateral analyses, using the minimum correlation in the image \( R_{\text{min}} \) and \( \text{STCS}_{\text{max}} \) as the summary statistics. Using 1000 permutations, statistical significance was considered at \( \alpha = 0.05 \).

5.3 Results

5.3.1 Simulated data

Figure 5.5 shows the correlation maps of the simulated data for differing sample sizes \( (N) \). Shown by the dotted black, solid black, and yellow contours are the regions identified as statistically significant by the uncorrected p-value, permutation testing using \( R_{\text{max}} \), and permutation testing using \( \text{STCS}_{\text{max}} \), respectively. In addition to the central region that was designed to correlate with the simulated outcome data, several other regions also show a statistically significant correlation for the uncorrected p-value, even for large \( N \). These are false positives.

124
Figure 5.5: The statistically significant region identified with an uncorrected p-value (dotted line), through permutation testing based on $R_{\text{max}}$ (solid black line) and permutation testing based on $\text{STCS}_{\text{max}}$ (solid gold line) for increasing $N$.

Permutation testing with both $R_{\text{max}}$ and $\text{STCS}_{\text{max}}$ successfully corrected for multiple comparisons, identifying just the central region as statistically significant. For the smallest sample size, permutation testing with $\text{STCS}_{\text{max}}$ identified more of the central region than permutation testing with $R_{\text{max}}$, indicating that $\text{STCS}_{\text{max}}$ is potentially more sensitive for small sample sizes.

The effect of noise in the outcome data was also investigated, and Figure 5.6 shows the correlation image ($N = 50$) for differing amounts of noise in the outcome data. Again, the central region was correctly identified with the uncorrected p-value (dotted line), but several false positive regions were also identified. These false positives were removed by permutation testing, and both forms of permutation testing (solid black lines and yellow lines for $R_{\text{max}}$ and $\text{STCS}_{\text{max}}$, respectively) correctly identified the central regions as statistically significant when no noise was applied to the outcome data (Figure 5.6a). As more noise was applied to the outcome data, the permutation tests became more conservative, with the permutation test based on $R_{\text{max}}$ failing to identify any statistically significant pixels for the noisiest outcome data (Figure 5.6c). The permutation test based
Figure 5.6: The statistically significant region identified with an uncorrected p-value (dotted line), through permutation testing based on $R_{\text{max}}$ (solid black line) and permutation testing based on STCS$_{\text{max}}$ (solid gold line) for increasing levels of noise in the output data; i.e. increasingly weaker correlations.

on STCS$_{\text{max}}$ successfully identified the central region for all levels of noise, indicating that it is more sensitive than a permutation test using $R_{\text{max}}$ when the correlation with the outcome data is weak.

For permutation testing using STCS$_{\text{max}}$, a threshold is applied to the statistic image, and the size of the largest resulting cluster is used as the summary statistic. Figure 5.7 illustrates the effect of different thresholds for an image set with $N = 50$. When a threshold of 0.25 was used, the central region was not identified as being statistically significant; instead a false positive region was identified (Figure 5.7a). A more appropriate threshold of 0.40 restored the central region as a statistically significant cluster (Figure 5.7b). This highlights the importance of choosing an appropriate threshold when using STCS$_{\text{max}}$, particularly when it is likely that the region of interest is small.
Figure 5.7: The statistically significant region calculated from STCS max is shown in gold, and the uncorrected p-value is shown by the dotted line. The choice of threshold is important, as it can result in false positives.

5.3.2 Application to radiation-induced trismus

Next, the methodology was applied to a cohort of head and neck cancer patients. Of the 95 patients with available data, five were excluded as they had baseline MID < 35 mm. The bulk analysis was performed on the remaining 90 patients. Of these 90 patients, 50 patients were used for the ipsi-contra analysis, and all 90 were used for the ipsilateral analysis.

Figure 5.8 shows the voxel-wise correlation maps for the bulk (Figure 5.8a), ipsi-contra (figure 5.8b), and ipsilateral (Figure 5.8c) analyses on a representative axial slice. There were large regions in which BED was negatively correlated with MID at six months post-radiotherapy for all analysis techniques, but the strongest correlation was observed for the ipsi-contra analysis, with $R_{\text{min}} = -0.40$, $R_{\text{min}} = -0.58$ and $R_{\text{min}} = -0.41$ for the bulk, ipsi-contra and ipsilateral analyses, respectively. The negative correlations mean that a higher dose leads to a reduced mouth opening ability, and the strongest correlation for all analyses was on the
Figure 5.8: Voxel-wise correlation maps on a representative axial slice of the reference patients for the bulk analysis (a), the ipsi-contra analysis (b), and the ipsilateral analysis (c). The masseters are shown in red, the medial pterygoids in blue, the lateral pterygoids in yellow, and the temporalis in magenta. The right hand side (as viewed) of the ipsi-contra and ipsilateral analyses (b and c) is the ipsilateral side, and the left hand side (as viewed) of the ipsi-contra analysis (b) is the contralateral side.

Figure 5.9: Logarithmic plot of the uncorrected p-value for the bulk (a), ipsi-contra (b), and the ipsilateral analysis (c). The masseters are shown in red, the medial pterygoids in blue, the lateral pterygoids in yellow, and the temporalis in magenta. The right hand side (as viewed) of the ipsi-contra and ipsilateral analyses (b and c) is the ipsilateral side, and the left hand side (as viewed) of the ipsi-contra analysis (b) is the contralateral side.
Table 5.2: Corrected p-values from permutation testing and the minimum and maximum correlation coefficients in the statistic image for the different analyses.

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Corrected p-value $R_{\text{min}}$</th>
<th>STCS$_{\text{max}}$</th>
<th>Min R</th>
<th>Max R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bulk</td>
<td>0.048</td>
<td>0.028</td>
<td>-0.40</td>
<td>0.21</td>
</tr>
<tr>
<td>Ipsi-contra</td>
<td>0.015</td>
<td>0.036</td>
<td>-0.58</td>
<td>0.37</td>
</tr>
<tr>
<td>Ipsilateral</td>
<td>0.002</td>
<td>0.003</td>
<td>-0.41</td>
<td>0.17</td>
</tr>
</tbody>
</table>

left hand side of the patient, corresponding to the ipsilateral side for the ipsi-contra and ipsilateral analyses.

Figure 5.9 shows the corresponding logarithmic plot of the uncorrected p-value map for the three analysis types. For the bulk analysis (Figure 5.9a) there was a large region of low p-value on the left hand side of the patient, with smaller regions on the right hand side. For the ipsi-contra analysis (Figure 5.9b) there was a smaller region of statistical significance on the ipsilateral side of the patient, but not on the contralateral side. For the ipsilateral analysis (Figure 5.9c) there were large regions with low p-value on both sides of the patient; however, only the ipsilateral side (left hand side of the reference patient) is valid, as the right hand side of the reference patient consists of a mixture of ipsilateral and contralateral structures.

Permutation testing was performed on the correlation images using $R_{\text{min}}$ and STCS$_{\text{max}}$ as the summary statistics. The corrected p-values from permutation testing are summarised in Table 5.2, from which it can be seen that the correlation images for all analyses were statistically significant at $\alpha = 0.05$ for both type of permutation test.
Figure 5.10: Statistically significant voxels ($\alpha = 0.05$) identified using permutation testing and STCS$_\text{max}$ are shown by the red regions for the bulk analysis (a), the ipsi-contra analysis (b), and the ipsilateral analysis (c). The masseters are shown in red, the medial pterygoids in blue, the lateral pterygoids in yellow, and the temporalis in magenta. The right hand side (as viewed) of the ipsi-contra and ipsilateral analyses (b and c) is the ipsilateral side, and the left hand side (as viewed) of the ipsi-contra analysis (b) is the contralateral side.

For permutation testing using STCS$_\text{max}$, thresholds of $-0.35$, $-0.45$ and $-0.35$ were used to create the clusters for the bulk, ipsi-contra and ipsilateral analyses, respectively. Figure 5.10 shows statistically significant clusters ($\alpha = 0.05$) identified from permutation testing. Note that the figure shows only the statistically significant clusters identified using STCS$_\text{max}$; although statistically significant voxels were identified with $R_{\text{min}}$, only a small number were statistically significant and so are not shown here. For the three analyses, a single statistically significant cluster was identified. For the bulk analysis (Figure 5.10a) this was on the left hand side of the reference patient, close to the masseter. For the ipsi-contra analysis (Figure 5.10b) the cluster was located on the ipsilateral side, in close proximity to the masseter. For the ipsilateral analysis (Figure 5.10c) a statistically significant cluster was identified on the ipsilateral side of the image, now overlapping with the masseter.
Figure 5.11: Axial slices of the correlation map (a), mean dose (b), and standard deviation (c) for the ipsilateral analysis. The masseters are shown in red, the medial pterygoids in blue, the lateral pterygoids in yellow, and the temporalis in magenta. Profiles through the region identified as having a statistically significant relationship with MID at six months are shown in (d). The correlation coefficient is shown by the solid black line, and the normalised mean dose ($\mu$) and normalised standard deviation ($\sigma$) are indicated by the solid and dotted red lines, respectively.
Figures 5.11 and 5.12 show transverse and sagittal slices, respectively, of the correlation maps (Figures 5.11a and 5.12a), mean doses over all patients (Figures 5.11b and 5.12b), and standard deviations of the dose over all patients (Figures 5.11c and 5.12c) from the ipsilateral analysis. Line profiles were taken across these images, indicated by the red lines, and the results are plotted in Figures 5.11d and 5.12d for representative transverse and sagittal slices, respectively. It can be seen that the regions of high correlation were positioned on the edge of the regions of highest mean dose, with the correlation falling in regions where the mean dose was high or low. The standard deviation was relatively flat across the profiles. Furthermore, it can be seen in Figure 5.12 for the sagittal profile that there was a reduction in the standard deviation corresponding to the point at which the correlation reduced. This could explain why the correlating cluster was not observed on more inferior slices.

5.4 Discussion

Image-based data mining is a powerful technique that enables dose-response relationships to be identified from the dose distributions of previously treated patients. However, the large number of voxels in typical medical images present a multiple comparisons problem that can result in false positives. We have introduced here a novel method of correcting for multiple comparisons with permutation testing when the clinical endpoint is continuously variable. We have demonstrated this technique with simulated data and showed that cluster-based techniques are more powerful than techniques based on an extreme statistic. Furthermore, upon application to a cohort of head and neck cancer patients, we have shown that the ipsilateral masseter has a dose-response relationship with trismus.

There has been increasing interest in image-based data mining in radiotherapy
Figure 5.12: Sagittal slices of the correlation map (a), mean dose (b), and standard deviation (c) for the ipsilateral analysis. The masseters are shown in red, the medial pterygoids in blue, the lateral pterygoids in yellow, and the temporalis in magenta. Profiles through the region identified as having a statistically significant relationship with MID at six months are shown in (d). The correlation coefficient is shown by the solid black line, and the normalised mean dose ($\mu$) and normalised standard deviation ($\sigma$) are indicated by the solid and dotted red lines, respectively.
in recent years due to its ability to spatially localise dose-response relationships. Witte et al. (2010) split 352 prostate patients into two groups corresponding to patients with and without regional failure, and analysed the difference in dose distribution between the two groups. They found that patients with regional recurrence received on average a lower dose outside the target volume, in the lymphatic region where regional cancer spread might be expected. Similar voxel-based analyses have also been performed with respect to radiotherapy-related toxicities. Such analyses have been performed for prostate radiotherapy with respect to urinary obstructions (Heemsbergen et al., 2010), rectal bleeding (Acosta et al., 2013), proctitis (Wortel et al., 2015) and international prostate symptom score (Improtta et al., 2016; Palorini et al., 2016), as well as for Hodgkin’s lymphoma with respect to radiation-induced lung damage (Palma et al., 2016). In addition, a similar technique has been applied to the parotids for identifying sub-volumes responsible for for radiotherapy-related xerostomia in head and neck cancer (van Luijk et al., 2015).

Correcting for multiple comparisons is essential for image-based data mining, and this was highlighted with the simulated data, for which there was a central region with a known dose-response relationship. Without correcting for multiple comparisons, several regions with false positive dose-response relationships were observed. These could potentially lead to incorrect hypotheses.

There are several methods of correcting for multiple comparisons. One such technique is the Bonferroni correction, in which the p-value required for statistical significance is obtained by dividing the nominal significance level by the number of statistical tests. In reality, however, the Bonferroni test is too conservative, particularly for radiotherapy in which there is a strong correlation between neighbouring voxels in dose distributions. An alternative approach that has been proposed for voxel-based analyses is one based on random field theory (Friston et al., 1994). This assumes that each voxel in a statistic image has a parametric
distribution under the null hypothesis, against which each voxel can be tested for significance. For this, it is common to assume that the voxel values follow a Gaussian distribution. However, this is not always the case, and non-parametric approaches based on permutation test theory have been developed that do not require any prior assumptions about the distribution of voxel values in the image set (Holmes et al., 1996).

The concept of permutation testing for correction of multiple comparisons is well-established within the field of neuroimaging (Friston et al., 1994; Holmes et al., 1996); however, its application to the field of radiotherapy was only recently introduced by Chen et al. (2013). They calculated a test statistic based on the voxel-wise mean dose difference between two patient groups, normalised by the voxel-wise standard deviation. This test statistic, \( T \), represented a normalised dose difference map between the two groups, and calculation of \( T_{\text{max}} \) over 1000 permutations enabled correction for multiple comparisons.

Subsequent applications of permutation testing have been based on this \( T \)-statistic (Wortel et al., 2015; Palma et al., 2016), but they have only been applied to categorical endpoints in which patients were split into two groups. The present study is the first application of permutation testing in radiotherapy for continuously variable outcome data.

The choice of 1000 randomisations for the permutation testing was chosen as it has been shown that this is sufficient to yield an effective approximate permutation test (Edgington, 1969). However, the simulation used in the present study is suited to investigating the effect of different numbers of randomisations on the performance of permutation testing for image-based data mining. Such an experiment could be used to investigate any potential differences in the minimum number of randomisations required for the different summary statistics used.

Image-based data mining relies on accurate DIR to ensure that all patients are spatially normalised to the same frame of reference. This ensures that correlation
coefficients are calculated over voxels that correspond to the same spatial location in the dose images. In the present study, the results of DIR were carefully assessed by visual inspection of the bony anatomy in the deformed CT images for each patient. More sophisticated techniques for analysing the registration performance are available, such as quantifying the concordance of structures (Brock, 2010), or measuring the uncertainty in the DVF (Saleh et al., 2014). However, visually, no large discrepancies in the deformed CT images were observed for the present study. Uncertainty in the registration is inevitable, but this is mitigated somewhat by the high degree of spatial correlation between doses to neighbouring voxels in radiotherapy dose distributions. Further work is planned to assess the impact of registration uncertainties on the results. This will be performed by quantifying the registration uncertainties by means of measuring the variation in the centre of mass of the deformed structures, and then blurring the dose distributions accordingly. Image-based data mining using these blurred dose distributions will then enable incorporation of the registration uncertainties into the results.

Similarly, the choice of reference patient is important for the spatial normalisation process, as DIR is less accurate when large deformations are present (Yeo et al., 2012). In the present study, the reference patient was chosen manually by visual inspection. Automatic selection of a reference patient that is most like the other patients in the sample have been suggested (Acosta et al., 2013). An alternative approach would be to use group-wise registration, in which all images are used for registration simultaneously, removing the need for selection of a reference patient that could bias the registration results (Ehrhardt et al., 2011). This will be investigated as part of future work. However, DIR was considered successful in all patients used for the present study, and any potential improvements in overall registration accuracy are likely to be small.
Permutation testing revealed a cluster of voxels exhibiting a statistically significant dose-response relationship. Analysis of all patients in the cohort (bulk analysis) revealed a cluster in close proximity to the masseter on the left hand side of the patient. The asymmetric correlation map is likely explained by the fact that there were 33 patients in which the ipsilateral side was judged to be on the left hand side of the patient, but only 17 patients in which the ipsilateral side was on the right hand side (40 patients had bilateral disease at the level of the muscles of mastication). This implied that the correlation was driven by the structure receiving the highest dose (i.e. the dose to the ipsilateral side).

This was confirmed with the ipsi-contra analysis, in which the dose distributions of patients with right-sided tumours were flipped such that the left hand side of the correlation map was the ipsilateral side, and the right hand side was the contralateral side. Strong correlation was only observed on the ipsilateral side, confirming that the ipsilateral structures were related to MID six months post-radiotherapy. For this analysis, the patients for whom there was no clear distinction between ipsilateral and contralateral structures were excluded. For the final analysis (ipsilateral analysis), these patients were included such that both sets of structures (left and right hand side) were analysed. This increased the power of the analysis as more patient data were included, and confirmed a cluster on the ipsilateral side of the patient, overlapping with the masseter. Flipping images in this way is a novel technique for increasing the power of image-based data mining experiments when separation into ipsilateral and contralateral structures is warranted.

The cluster identified here was not confined solely to the ipsilateral masseter; much of the cluster was outside this structure. However, there are no anatomic structures in this region that are likely to have a dose-response relationship with trismus, and so it is unlikely that this cluster was caused by an unexpected anatomic structure. The correlation outside the masseter was more likely caused
by the fact that neighbouring voxels within dose distributions tend to be highly 
correlated, as well as uncertainty in the spatial normalisation process. In addition, 
the cluster identified here was present on four CT slices (2 cm), at the edge of 
the mean dose distribution in the superior-inferior direction. It did not extend 
throughout the entire masseter, but this was likely caused by the fact that there 
was less variation in dose in more inferior regions of the masseter.

The present study developed a novel method of correcting image-based data 
mining for multiple comparisons when the clinical endpoint is continuously vari-
able. Two different summary statistics were used for permutation testing: one 
based on an extreme statistic ($R_{\text{max}}$); and one based on the largest cluster of 
voxels exceeding some threshold ($\text{STCS}_{\text{max}}$). Both methods successfully accoun-
ted for multiple comparisons, but $\text{STCS}_{\text{max}}$ was found to be more sensitive than 
$R_{\text{max}}$ for both the simulated and clinical data. Although the localising ability of 
$\text{STCS}_{\text{max}}$ is lower than that of $R_{\text{max}}$ (Nichols and Holmes, 2002), it has been shown 
here that it is more powerful for detecting dose-response relationships for clinical 
endpoints in which there are confounding variables. Due to the complex nature 
of dose-response relationships in radiotherapy and the often weak correlations, 
permutation testing using $\text{STCS}_{\text{max}}$ should be considered for image-based data 
mining experiments in radiotherapy. The results from the simulated data showed 
that careful selection of the threshold is important, and so improvements to the 
metric to incorporate threshold-free approaches would remove the sensitivity of 
image-based data mining approaches to threshold selection. Such approaches 
have been suggested (Smith and Nichols, 2009), and may prove useful for future 
image-based data mining studies.

The results of this study have indicated that the ipsilateral masseter has a 
dose-response relationship with trismus. Whilst validation in an external cohort is 
essential, this indicates that reducing the dose to this structure during radiother-
apy should reduce the severity of trismus. Further work is required to determine
which dosimetric parameters should be considered, and treatment planning studies will enable the potential for reducing the dose to the ipsilateral masseter to be assessed.

The methodology introduced here enables image-based data mining with any radiotherapy outcome data, and has potential applications to many radiotherapy-related toxicities, as well as outcomes. This will enable a better understanding of dose-response relationships in radiotherapy, which will help to identify specific structures to which the dose should be minimised in order to reduce a particular toxicity. This knowledge will help to optimise adaptive radiotherapy (ART), which is the management of anatomic changes that often occur during radiotherapy and the resulting dosimetric deviations from the intended treatment plan. ART is labour-intensive, and by better understanding the relationship between delivered dose and a specific toxicity, decisions for plan adaptation can be based on predicted risk of complication due to changes in dose. This will enable better patient selection, enabling ART to focus on patients who will likely receive the most benefit.

5.5 Conclusion

Image-based data mining is a powerful technique, capable of identifying regions with a dose-response relationship with a clinical endpoint. Permutation testing corrects for multiple comparisons, and has been previously applied to categorical outcome data. This study is the first to extend permutation testing for continuously variable radiotherapy outcome data, enabling more dose-response relationships to be explored.

Permutation testing based on the maximum size of clusters above a threshold ($\text{STCS}_{\text{max}}$) was shown to be more powerful than that based on an extreme statistic ($R_{\text{max}}$), particularly when the outcome data are affected by confounding variables.
Permutation testing using cluster-based techniques is therefore recommended for image-based data mining in radiotherapy.

Upon application to head and neck patients with respect to trismus, a cluster of voxels exhibiting a statistically significant dose-response relationship was identified in the ipsilateral masseter. Validation is required, but these results suggest that the ipsilateral masseter is an important mastication structure, and that sparing this structure during radiotherapy could lead to a reduction in the incidence and severity of radiation-induced trismus.

Permutation testing allows statistically robust dose-response relationship hypotheses to be generated, and can now be applied to any categorical or continuously variable clinical endpoint. Improved understanding of dose-response relationships in radiotherapy will ultimately enable improvement in the therapeutic ratio.
6 Discussion

The rationale for this thesis was the optimisation of adaptive radiotherapy (ART) for head and neck cancer, both in terms of clinical implementation and clinical benefit. Several recent planning studies have shown that the conventional approach of applying radiotherapy based on patient anatomy before the start of treatment is not optimal. ART can mitigate anatomic changes leading to changes in cumulative dose through replanning, resulting in a reduction in the severity and incidence of radiotherapy-related toxicities.

However, the time-consuming and resource-intensive nature of ART is a major challenge for its routine clinical implementation. ART strategies rely heavily on the acquisition of on-treatment images for assessing the dosimetric impact of anatomic changes, calculating dose on the new anatomy, automatically segmenting structures, and deciding whether a new treatment plan is required. Challenges such as establishing appropriate ART workflows, ensuring that deformable image registration and contour propagation is accurate, and understanding the potential clinical benefits of ART have so far limited the current application of ART.

Optimising ART requires a balance between the potential clinical gains that could be achieved by replanning to spare a particular OAR and the workload involved with the procedure. The projects within this thesis have been undertaken within this context, and the aims identified in the introduction are reiterated below:

1. To determine the relative plan robustness to weight loss during radiotherapy of VMAT and IMRT.

2. To determine which metrics should be used for assessing automatic segmentation accuracy in the head and neck.
3. To develop a method of performing patient-specific quality control of contour propagation accuracy.

4. To develop a statistically robust image-based data mining technique for identifying structures exhibiting a dose-response relationship with a continuously variable toxicity outcome.

5. To determine which structures in the head and neck have a dose-response relationship with trismus.

These aims have been successfully addressed in a series of publications, which are discussed in this chapter. Suggestions for future work to develop these aims are also identified, as well as additional areas of work for further optimising the application of ART for head and neck cancer patients.

6.1 Relative robustness of IMRT and VMAT to weight loss during head and neck radiotherapy

In this project, the dosimetric effect of weight loss during head and neck radiotherapy was investigated for IMRT and VMAT. It was shown that IMRT and VMAT are equally robust to weight loss during head and neck radiotherapy, addressing aim 1.

The optimal head and neck ART strategy in terms of the frequency and timing of replans is an important consideration, and there have been several studies seeking to determine the most appropriate strategy. Differing numbers of replans have been suggested, with one (Ahn et al., 2011; Zhao et al., 2011; Schwartz et al., 2012), two (Wu et al., 2009) and three (Zhang et al., 2016) replans reported in the literature. The vast majority of this literature concerns ART for IMRT; however, VMAT has become more common due to its more efficient delivery (Teoh et al., 2011). This study was the first to investigate the relative robustness of IMRT
and VMAT during head and neck radiotherapy, and showed that the dosimetric changes that occur during head and neck IMRT are equivalent to those that occur during head and neck VMAT. This means that the literature concerning the dosimetric changes during head and neck IMRT is applicable to VMAT, and that a plan adaptation strategy for IMRT and VMAT would be equivalent in terms of frequency and number of replans. This is reassuring for centres considering replacing IMRT with VMAT, as there should be no change in the amount of replanning due to weight loss for head and neck cancer.

Although ART strategies for photons can be considered equivalent, this is not the case for proton beam therapy. The way in which protons deposit dose is fundamentally different to that of photons, with the finite range of the Bragg peak depending on the incident energy of the protons and the amount and density of tissue upstream. This results in superior OAR sparing in the head and neck for proton beam therapy relative to photons (van de Water et al., 2011). These fundamental differences mean that proton radiotherapy is likely to be more sensitive to anatomic changes (Kraan et al., 2013; Góra et al., 2015), and the optimal ART strategy has not been established. There is much opportunity for further research to establish the requirement for plan adaptation for proton beam therapy and to determine the optimal ART strategy.

There is some disagreement in the literature concerning the optimum replanning strategy for head and neck IMRT, which could to some extent be explained by heterogeneities in patient populations and the small numbers of patients generally included in studies. However, this highlights the potential role of patient selection for ART: some patients will likely benefit more than others, and predicting these patients from pre- or early treatment patient parameters has been the subject of some recent studies (Guidi et al., 2015; Brouwer et al., 2016; Castelli et al., 2016). Although these studies were applied to IMRT, Chapter 2 has shown that they are also applicable to VMAT. Such methods of patient selection show
promise for improving the efficiency of ART, and present exciting opportunities for further work, linking well with ongoing predictive modelling activities in our group.

6.2 The suitability of common metrics for assessing automatic segmentation accuracy

This study compared the suitability of different metrics for assessing automatic segmentation accuracy. Based on the assertion that geometric accuracy should correlate with dosimetric discrepancies for a given dose distribution, it was found that the widely used Dice similarity coefficient (DSC) was one of the least suitable metrics for assessing automatic segmentation accuracy of the parotids and larynx for head and neck VMAT. Automatic segmentation is an important component of ART, and validation before clinical implementation is essential. There is no agreement in the literature on which metrics should be used to assess the geometric accuracy of an algorithm, and this study was the first to address this lack of consensus in the head and neck, achieving aim 2.

DSC is the most widely used metric; it is easily calculated and easily interpreted. However, DSC is sensitive to the volume of the structures being assessed (Sykes, 2014), and Rohlfing (2012) has previously showed that it is not reliable for assessing the accuracy of image registration algorithms. The results from Chapter 3 provide further evidence of the pitfalls of using DSC for assessing automatic segmentation accuracy. DTA, a metric based on the euclidean distance between two surfaces, was found to be more clearly related to dosimetric discrepancies for both structures investigated, and is more suitable for assessing automatic segmentation accuracy.

One of the challenges implicit in any validation of an automatic segmentation
algorithm is the uncertainty in the underlying ground truth against which accuracy is measured. The ground truth is typically manually drawn by a clinician, but is subject to inter- and intra-observer variation (Loo et al., 2012). The use of contouring guidelines can reduce these uncertainties (Lobefalo et al., 2013), but cannot remove them completely. This uncertainty in the ground truth propagates into the measured geometric accuracy of an automatically generated contour, and so it is difficult to determine an acceptable value for geometric accuracy. Metrics should be benchmarked against the inter-observer variation for that structure, but further work is required to determine acceptable values for the different metrics used for automatic segmentation.

One of the interesting findings of this study was that the choice of the most suitable metrics was dependent on the specific structure in question. For example, the euclidean distance between the centres of mass was found to be suitable for the parotid, but not for the larynx. The position of the structure relative to typical dose gradients has a strong impact on the choice of metric to be used and highlights the importance of choosing a suitable metric when assessing automatic segmentation accuracy.

The methodology developed in this study could be easily extended to other treatment sites to determine which metrics are most appropriate for other structures. By establishing a consensus on metrics used to assess automatic segmentation accuracy, comparing and benchmarking different algorithms will become easier. This will enable the performance of different algorithms to be compared in an objective manner, ultimately ensuring accurate automatic segmentation for ART.
6.3 An automated workflow for quality control of contour propagation

In this chapter, a novel methodology for performing patient-specific quality control of contour propagation was introduced and validated for the parotids. No tools for assessing the accuracy of contour propagation currently exist, and the method presented in Chapter 4 is the first example of such a tool. This has addressed aim 3.

Automatic segmentation of on-treatment images is an essential requirement for ART, enabling an evaluation of the dose distribution at each fraction and the need for plan adaptation. It is therefore essential that the contours used to perform this assessment reflect the true anatomy. Segmentation of on-treatment images for ART has previously been manually performed *de novo* (Acharya et al., 2016) or by manually reviewing propagated contours (Schwartz et al., 2013). This process is labour-intensive, requiring significant clinician time, and precludes routine implementation of ART. The novel workflow for patient-specific quality control of contour propagation introduced in Chapter 4 could reduce this burden, automatically flagging contours that are likely to be incorrect. Only these contours would require review, increasing the efficiency of ART and potentially reducing some of the barriers that have so far prevented its routine adoption.

The methodology introduced here has been developed within the context of head and neck ART, for which offline plan adaptation is generally considered sufficient. Whilst this method would improve the efficiency of the replanning process, the potential benefits are even greater for online ART, such as for the bladder (Vestergaard et al., 2014) or cervix (Heijkoop et al., 2014), in which replanning is subject to acute time pressures. For this, propagated contours must be used directly for plan assessment and optimisation whilst the patient is immobilised in the treatment position. Automated means of quickly assessing
the accuracy of propagated contours is essential and so the workflow developed here could help facilitate the introduction of online ART.

Contour propagation derives from the deformation vector field (DVF) resulting from deformable image registration (DIR) between a planning CT (pCT) and an on-treatment image. The DVF can also be used for dose accumulation, in which doses to individual voxels are tracked and accumulated onto a single reference image (typically the pCT). Whilst beyond the scope of this thesis, dose accumulation will play an important role in ART, and further discussion is warranted. Dose accumulation requires accurate DIR not just at boundaries between anatomic structures, but also in regions of low contrast within anatomic structures. Furthermore, structures that exhibit sliding motion cause discontinuities in the DVF that can be challenging to model. This means that techniques for assessing contour propagation accuracy, for which only the accuracy of anatomic borders is important, cannot be used for assessing DIR with respect to dose accumulation. Different methods of assessing dose accumulation accuracy have been suggested (Brock, 2010; Yeo et al., 2012; Saleh et al., 2014; Roussakis et al., 2015), and tools for integrating it into the clinical workflow are being developed (Park et al., 2016).

Online ART requires both accurate dose accumulation and accurate contour propagation. Accumulating the dose delivered over all treatment fractions enables calculation of the cumulative delivered dose and, therefore, deviations from the planned dose. These deviations, which could originate from anatomic changes relative to the original pCT, could then be used along with propagated contours as part of online plan optimisation to restore the intended dose. An optimisation system that could perform online plan optimisation accounting for new anatomy has recently been developed by Kontaxis et al. (2015). Illustrated in Figure 6.1, they developed an optimisation pipeline with an adaptive sequencer (ASEQ) that accounts for the changing patient anatomy. There is urgent need for further work
to develop and optimise online ART to ensure the safety of this approach, including means of performing effective quality control of the imaging and treatment plan deliverability.

Methods of assessing DIR and contour propagation will be particularly important for advanced adaptive radiotherapy with onboard magnetic resonance imaging (MRI) guidance. Integration of MRI into the radiotherapy workflow will enable daily imaging with excellent soft tissue contrast, providing the opportunity to perform daily plan adaptation. Several systems are being developed (Lagendijk et al., 2008; Tadic and Fallone, 2010; Mutic and Dempsey, 2014), but in order to fully exploit the benefits afforded by ART, fast and accurate contour propagation is required.

Manual review of propagated contours is currently the gold standard for ensuring accurate patient-specific contour propagation, but this is labour-intensive and is not feasible for routine implementation. Tools for ensuring that propagated
contours, such as that presented in Chapter 4, will be essential for the implementation of such advanced radiotherapy techniques. This tool would flag contours that are subject to a large degree of uncertainty, and by only reviewing these flagged contours, this would dramatically increase the efficiency of MRI-guided ART, facilitating its clinical implementation.

Tools that make ART more efficient will make it easier to adapt plans in response to anatomic and dosimetric changes. Currently the evidence suggesting a benefit for ART is based on retrospective studies that have shown dosimetric benefits for replanning compared to no replanning. Some studies have also related the dosimetric benefits to risk of treatment-related toxicities for xerostomia. However, these retrospective studies are subject to a selection bias, as there could be underlying biological or clinical reasons for anatomic changes that could be reflected in the decision to replan. There is a need for prospective studies, such as the one performed by Schwartz et al. (2012), incorporating ART into the clinical workflow. This will become easier as ART becomes more efficient, and tools for performing accurate automatic segmentation and DIR will make studies investigating the clinical benefit more feasible.

6.4 Permutation testing for image-based data mining

In this chapter, a method of correcting for multiple comparisons in image-based data mining in radiotherapy was introduced. Based on the technique of permutation testing introduced by Chen et al. (2013) for binary radiotherapy outcome data, this work is the first to extend the concept to continuously variable outcome data, addressing aim 4. Upon application of the technique to a cohort of head and neck cancer patients with respect to trismus, a region, overlapping with the ipsilateral masseter, was identified as having a statistically significant dose-response relationship. This addressed aim 5.
With improving image processing technology and increased availability of DIR software, image-based data mining is becoming more feasible. It has been applied to prostate radiotherapy with respect to treatment failure (Witte et al., 2010), urinary obstructions (Heemsbergen et al., 2010), rectal bleeding (Acosta et al., 2013), proctitis (Wortel et al., 2015) and international prostate symptom score (Improta et al., 2016; Palorini et al., 2016), as well as Hodgkin’s lymphoma with respect to radiation-induced lung damage (Palma et al., 2016). A similar technique has also been used to identify sub-regions of the parotid that exhibit a stronger dose-response relationship with xerostomia than the whole gland (van Luijk et al., 2015). However, the large number of voxels in medical images introduces a multiple comparisons problem that must be addressed to ensure confidence in any findings.

Permutation testing corrects for multiple comparisons, but has previously only been described in radiotherapy for binary outcome data (Chen et al., 2013; Wortel et al., 2015; Palma et al., 2016). The work presented in Chapter 5 is the first to describe a method of performing permutation testing for continuously variable outcome data, and enables more detailed clinical endpoints to be studied. Both categorical and continuous clinical endpoints are used in radiotherapy, and being able to perform image-based data mining using both types of outcome data will enable more treatment-related toxicities to be studied. This could ultimately lead to a more comprehensive understanding of toxicities in radiotherapy and the potential for reducing them by carefully managing the dose delivered to key structures during initial treatment planning and ART.

As part of the study performed in Chapter 5, it was found that permutation testing using cluster-based methods is more powerful than that based on extreme statistics. This is particularly useful for radiotherapy, in which the presence of confounding variables often weakens correlation. Despite the advantages of cluster-based methods, the simulated data showed a clear dependence of the
results on the choice of threshold used to create the clusters. Threshold-free cluster-based techniques have been suggested (Smith and Nichols, 2009), and could mitigate the dependence of the result on the specific choice of threshold. This is currently being investigated in our group.

The effect of changing the number of randomisations used for the permutation test could also be investigated in future work. Although it has been shown that 1000 randomisations generally provide an effective estimation of the true permutation distribution (Edgington, 1969), the simulations performed as part of the study could be used to investigate this further. By modifying the number of randomisations in the permutation test, its effect could be investigated for the different summary statistics, sample sizes and levels of noise in the outcome data.

The results from the application of the image-based data mining technique to a cohort of head and neck cancer patients identified a cluster of voxels, overlapping with the ipsilateral masseter, exhibiting a dose-response relationship with trismus. This supports recent evidence that the dose to this structure is more predictive of the incidence of trismus than the doses to the other muscles of mastication (Lindblom et al., 2014; Pauli et al., 2016; Rao et al., 2016). Furthermore, recent analysis using delineated structures on the same cohort of patients used in Chapter 5 has shown that the ipsilateral masseter exhibits a stronger dose-response relationship with the severity of trismus at 6 months post-radiotherapy than any of the other muscles of mastication (Thor et al., 2016). For this, the correlation of the mean dose with the MID at 6 months was measured for each delineated structure, which is complementary to the image-based data mining approach. These recent studies support the findings of the present study.

Image-based data mining is a retrospective technique that enables exploration of dose-response relationships from previously treated patients. It provides a robust statistical correction for multiple comparisons, but validation in an independent cohort is essential to confirm any findings. This is the subject of ongoing
and future work. It is planned to analyse a cohort of 40 patients treated at Memorial Sloan Kettering Cancer Centre (MSKCC) in the near future. Propagation of the identified cluster onto these patients will enable a region-based analysis that includes the other muscles of mastication. This will test the generalisability of the findings across different patient cohorts. In addition, it is planned to perform an additional validation on a cohort of patients treated at MSKCC for whom trismus was graded according to the Common Terminology Criteria for Adverse Events (CTCAE), which is an alternative method of assessing the impact of trismus. Up to 400 patients will be available, and will test the generalisability of the findings across different methods of assessing trismus. This work is ongoing, and will be combined with the results presented in Chapter 5 before final publication.

Having identified the ipsilateral masseter as a structure with a dose-response relationship with radiation-induced trismus, the potential for sparing this structure during radiotherapy should be investigated. The majority of studies have so far characterised the dose to ipsilateral masseter by the mean dose, but further work is required to determine which dosimetric parameter best correlates with trismus. This will enable treatment planning studies to determine the extent to which the dose can be reduced and allow an estimate of the clinical benefit of sparing this structure. Furthermore, proton beam therapy could offer further sparing, and this should be explored.

Image-based data mining, with appropriate correction for multiple comparisons, is a tool for generating dose-response hypotheses. The novel methodology developed in Chapter 5 now allows this tool to be applied to any radiotherapy outcome, and there is much potential for applying this technique to other toxicities. Once validated, this could inform radiotherapy treatment planning and ART.

ART is extremely labour-intensive, and the benefit of ART is likely modest when considered on a population level. Nevertheless, some patients can undergo
extreme anatomic changes throughout treatment that can result in significant deviations in the delivered dose relative to the intended treatment plan. Understanding the impact of these dosimetric changes on a specific toxicity is essential in order to identify the patients for whom plan adaptation will be of most benefit, as well as the optimal timing for replanning. This will become even more pertinent with the introduction of more advanced radiotherapy techniques such as proton beam therapy, which will likely make the treatment plan more sensitive to anatomic changes, making ART more important. Dose-response hypotheses generated through image-based data mining approaches can therefore help to optimise ART by establishing which patients will benefit from ART, ultimately making implementation of ART easier.
7 Conclusion

In this thesis, a number of projects were performed with the overall aim of optimising adaptive radiotherapy. Anatomic changes occur during head and neck radiotherapy, which means that a treatment plan based solely on a CT obtained prior to treatment can be sub-optimal. Adaptive radiotherapy is therefore often required to mitigate the resulting dosimetric consequences. Establishing ART workflows by identifying the requirement for replanning is an important aspect of optimising ART, and in Chapter 2 it was shown that head and neck ART workflows for IMRT and VMAT are equivalent with respect to replanning frequency.

ART requires accurate deformable image registration to enable dose accumulation and contour propagation. Delineation is one of the most time consuming parts of the radiotherapy workflow, and automated contour propagation is required to make ART more feasible. Pre-clinical assessment of automatic segmentation algorithms is essential, but it remains necessary to review propagated structures for each patient. In Chapter 3, the suitability of different metrics for assessing automatic segmentation accuracy for the parotids and larynx was measured. It was found that the commonly used Dice similarity coefficient was one of the least suitable metrics, and that metrics based on distance-to-agreement should be used for these structures. Identification of appropriate metrics for assessing automatic segmentation accuracy will enable objective comparisons between different algorithms, ultimately ensuring accurate automatic segmentation for ART.

Assessing contour propagation accuracy during clinical use is labour-intensive, and no tools currently exist for this purpose. A novel method of measuring patient-specific contour propagation accuracy was introduced in Chapter 4. This could be used to highlight instances in which contour propagation has failed, increasing the efficiency of ART and removing some of the barriers that have so
far limited its routine implementation.

**ART** minimises the dose to organs at risk, and accurate knowledge of dose-response relationships will enable the role of ART for reducing different toxicities to be explored. Radiotherapy-related toxicities are often complex, and image-based data mining techniques will enable dose-response hypotheses to be generated and, after validation, will inform radiotherapy treatment planning and ART. Current methods can only be used for categorical data, and in Chapter 5 a novel method of correcting image-based data mining techniques for multiple comparisons was developed for continuously variable clinical endpoints. This will allow image-based data mining with any radiotherapy outcome data. This technique was applied to a cohort of head and neck cancer patients treated with radiotherapy, and the ipsilateral masseter muscle was identified as being the most important muscle of mastication with regards to radiation-induced trismus. After appropriate validation, this could inform future treatment planning and ART studies with respect to reducing the severity of radiation-induced trismus.

**ART** mitigates the dosimetric effects of anatomic changes, and this thesis has addressed some of the technical and clinical challenges that have so far limited its clinical implementation. Detailed knowledge of dose-response relationships will enable selection of patients requiring plan adaptation based on potential clinical benefit, and fast and accurate DIR and contour propagation will make ART more efficient, facilitating its routine implementation.
References


Hoang Duc, A. K., Eminowicz, G., Mendes, R., Wong, S.-L., McClelland, J., Modat, M., Cardoso, M. J., Mendelson, A. F., Veiga, C., Kadir, T., D’souza,


Palma, G., Monti, S., D’Avino, V., Conson, M., Liuzzi, R., Pressello, M. C.,
voxel-based approach to explore local dose differences associated with
radiation-induced lung damage. *International Journal of Radiation Oncology

Palorini, F., Cozzarini, C., Gianolini, S., Botti, A., Carillo, V., Iotti, C.,
pixel-wise analysis on bladder dose-surface maps in prostate cancer

Park, S., McNutt, T., Plishker, W., Quon, H., Wong, J., Shekhar, R., and Lee,

Pauli, N., Johnson, J., Finizia, C., and Andrél, P. (2013). The incidence of
trismus and long-term impact on health-related quality of life in patients with

Pauli, N., Olsson, C., Pettersson, N., Johansson, M., Haugen, H., Wilderång,

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O.,
Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos,

registration of medical images: a survey. *IEEE Transactions on Medical

Auto-segmentation of normal and target structures in head and neck CT
38(11):6160–70.

dynamic programming approach to adaptive fractionation. *Physics in

Rancati, T., Fiorino, C., Fellin, G., Vavassori, V., Cagna, E., Casanova Borca,
V., Girelli, G., Menegotti, L., Monti, A. F., Tortoreto, F., Delle Canne, S.,
modelling of late rectal toxicity after high dose radiotherapy for prostate


172


