Observer variation in the delineation of organs at risk for head and neck radiation therapy treatment planning: a systematic review protocol

Alisha Moore¹,²

¹Faculty of Health and Medicine, School of Health Sciences, University of Newcastle, Newcastle, Australia, ²Trans-Tasman Radiation Oncology Group (TROG) Cancer Research, Newcastle, Australia

Review question: The objective of this review is to examine inter- and intra-observer agreement and reliability in the delineation of head and neck organs at risk (OAR) as part of the radiation therapy treatment planning process.

More specifically, the objectives are to identify:

i. The level of agreement and reliability among multiple observers in defining normal tissue in the head and neck region in both computed tomography (CT) and magnetic resonance imaging (MRI) scans.

ii. The level of agreement and reliability for single observers at different time points in defining normal tissue in the head and neck region on both CT and MRI scans.

iii. The contouring methods used to delineate head and neck organs at risk (OAR) in studies of intra and inter-observer reliability and agreement. Such contouring methods may include the use of software packages, education, technology and multi-modality imaging such as CT and MRI.

Keywords: radiation therapy; observer variation; organs at risk; observer variation; head and neck neoplasms

Introduction

Head and neck cancer definition and impact

Head and neck cancer (HNC) is defined as primary malignant neoplasms of the pharynx (hypopharynx, oropharynx and nasopharynx), larynx, salivary glands, lip, oral cavity, paranasal sinus and nasal cavity, as well as lymph node metastases from unknown primaries, head and neck sarcomas and lymphomas.¹ In Australia, in 2016, HNC is estimated to represent approximately 3.5% (4631 cases) of all newly diagnosed malignancies, with 550,000 cases expected worldwide.² Although the incidence of HNC is relatively small compared to other disease sites, the physical, mental and economic burden of HNC is significant.³,⁴ Adverse patient outcomes occur as a result of both uncontrolled tumor cell growth invading nearby tissue as well as from the modalities employed to treat the disease itself, including surgery, chemotherapy and radiation therapy.¹ Malignancies in this region of the body can have a profound impact on quality of life, resulting in functional and physical morbidities.³ Morbidities such as difficulty swallowing, breathing, impaired speech, changes in taste, sight and physical appearance as well as muscle weakness and neuralgia can occur.³,⁵-⁷ With overt and life changing morbidities, subsequent financial, social and psychological implications frequently manifest.¹,³,⁴

Role of radiation therapy in head and neck cancer

Radiation therapy (RT) is a key modality in the treatment of HNC. Both the stage and pathology of the disease will determine whether RT is used as the primary treatment modality or as an adjuvant therapy.⁸,⁹ Radiation therapy is typically delivered via multiple external beams of megavoltage ionising radiation, generated by a linear accelerator and focused towards a cancer target.¹⁰,¹¹ Essential to the successful delivery of radiation therapy is the accurate prediction of target coverage and the avoidance of healthy tissue.¹⁰
The head and neck region bears inherent complexity due to the proximity of multiple radiation sensitive structures which can dramatically impact basic human function and quality of life. A compounding factor to this complexity is that head and neck cancers are typically treated with high radiation doses to achieve a tumorcidal effect. The high doses required for tumor control, in combination with the physical and functional anatomy of the head and neck, place normal tissue in this region at risk of impaired function and toxicity.

**Radiation therapy treatment planning**

Treatment planning is undertaken prior to RT delivery and is a critical step in ensuring that the balance between high tumor dose and low normal tissue dose is maintained. A treatment planning system (TPS) with purpose built software is used to construct the RT treatment plan for each patient. Typically, a digital CT scan is imported into the TPS and used as the primary dataset, representing a surrogate for the patient. The digital CT scan is comprised of voxels, that is, three-dimensional units of pictorial information that form the matrix of the scan. The TPS utilizes electron density data from the CT scan to model interactions of ionizing radiation within the body. The user must define physical radiation treatment parameters to simulate treatment delivery, including but not limited to, beam geometry, prescribed dose, fractionation, beam energy, immobilization/treatment position, target volumes and normal tissue contours. The defined parameters are used to create a visual and statistical estimation of the radiation dose distribution within the patient.

**Target volume and organ at risk delineation**

Defining normal tissue and target volumes on digital scans (such as CT) is critical to the dosimetric and volumetric accuracy of RT treatment planning and delivery. Clinical target volumes (CTVs) are delineated for treatment planning and encompass gross tumor regions as well as regions of potential subclinical extension, determined by the stage and pathology of the disease. Organs at risk (OAR) are normal tissues that limit the feasible treatment plan parameters, based on their radiation sensitivity and proximity to the target lesion. In the head and neck region, OAR may include the brachial plexus, parotid glands, mandible, temporomandibular joint, cranial nerves, optic chiasm, orbits, lens, submandibular glands, oral cavity, glottic larynx, pharyngeal constrictor muscles, inner ear, middle ear, brainstem and spinal cord. For each patient, computer software tools are used to physically define the anatomical boundaries of cancer targets and OAR in digital imaging scans, via both manual and automatic processes. Manual delineation requires the user to trace anatomic boundaries by hand, using a computer mouse or digitizer. Automated methods may include algorithms that search for areas of common electron density, interpolation algorithms and computer modelling of anatomy based on a stored “bank” of atlases.

**Multi-modality imaging**

Multiple imaging modalities, including but not limited to, MRI and positron emission tomography (PET) can provide complementary information for both target volume and normal tissue delineation. Imaging datasets may be used in isolation or as complementary (secondary) scans. For radiation therapy treatment planning, a primary CT planning scan is acquired. Visualisation is enhanced by registering different imaging modalities (such as PET and MRI) to the primary CT planning scan. A mutual coordinate system is established and images are subsequently overlaid with the visualization advantages of each modality utilized in unison to inform the RT planning process. Positron emission tomography highlights areas characteristic of malignant growth and is used in radiation therapy treatment planning to delineate tumor volumes. Magnetic resonance imaging is a modality of choice for soft tissue visualisation including normal tissue and tumor volumes. Magnetic resonance imaging provides significant advantages in the head and neck due to the complexity and magnitude of soft tissue structures located in this region.

**Organs at risk contouring impact**

Head and neck RT treatment has escalated in complexity in line with advances in technology. Inversely planned, intensity modulated radiation therapy (IMRT) and volumetric modulated arc therapy (VMAT) are now well-established radiotherapy treatment planning and delivery techniques for HNC. Inverse planning requires each voxel within the patient to be prioritized and assigned dose volume criteria. Accurately defining the voxels...
that belong to each structure is therefore critical in ensuring correct priorities and dosimetric criteria are applied. Inaccurate delineation will ultimately lead to inaccurate modelling, visualization and estimation of radiation dose within the patient, which has the potential to impact tumor control and normal tissue toxicity. In addition, a lack of uniformity in delineation will inevitably reduce confidence in observed results, particularly for clinicians and collaborative groups who seek to correlate patient outcomes with definitive interventions.

**Inter-observer variation**
Numerous studies have been conducted which evaluate observer variation (across a range of anatomical sites) for both target volume delineation and normal tissue contouring in radiation therapy treatment planning. There is considerable evidence to indicate that contouring accuracy and consistency vary among clinicians. Interventions that seek to improve contouring accuracy have also been studied. Vinod et al. reviewed the factors that may reduce inter-observer variation, such as image quality and multimodality imaging, individual clinician training and expertise as well as the presence of ongoing education, standard protocols and collaborative efforts. Radiation oncology clinical trial groups have recognized the need to minimize inter-clinician variation with evidential links between protocol compliance, data integrity and patient outcome. Peters et al. linked protocol non-compliance in the TROG 02.02 head and neck trial with poorer patient outcomes. Rigorous RT quality assurance (QA) programs are a feature of many radiation oncology clinical trials, incorporating measures to promote structure delineation consistency. Furthermore, systematic reviews addressing inter-observer contouring variation have been published and highlight an ongoing pursuit to inform clinical practice regarding this issue.

**Inclusion criteria**

**Participants**
The systematic review will be limited to a patient population diagnosed with malignancy arising in the head and neck region, where radiation therapy is clinically indicated as a primary or adjuvant treatment. The population of interest must have undergone a cross sectional imaging media scan (including CT with or without MRI) for the purpose of radiation therapy treatment planning. No limit will be placed on disease stage or histology. Imaging scans will be from participants of at least 18 years of age.

**Instruments**
Studies that measure inter- and intra-observer agreement and variation of organ at risk delineation in the head and neck region will be included.

Organ at risk delineation will include both manual and automatic contouring methods used in radiation oncology treatment planning such as physically tracing anatomic boundaries by hand as well as using computer algorithms respectively. No limit will be placed on organ at risk delineation method.

Cross sectional imaging scans are defined as CT with a z-axis resolution consistent with reasonable requirements for head and neck radiation therapy treatment planning (that is, a minimum 3 mm z-axis resolution [or slice thickness]) as the primary dataset with or without MRI of equivalent z-axis resolution as the secondary or complementary media. Studies that include PET scans will not be specifically included as an imaging media of interest. Studies that include PET scans will be considered, provided the study includes at a minimum a CT (+or- a MRI) scan where normal tissue contouring has been conducted for the purpose of radiation therapy treatment planning.
Head and neck organs at risk: for the purpose of this review these will include the brachial plexus, parotid glands, submandibular glands, oral cavity, glottic larynx, pharyngeal constrictor muscles and spinal cord.

Observers: these will include medical professionals involved in physically defining anatomical structures on cross sectional imaging media for radiation therapy treatment planning, including radiation oncologists, radiation therapists, medical physicists as well as trainees of the aforementioned professions. The level of variation and agreement between multiple observers as well as single observers at different time points will be included.

Instrument properties
This review will consider studies that measure inter and intra rater reliability and agreement. The review will not be limited to any singular statistical measure, rather any statistic used to measure inter and intra rater reliability and agreement will be considered. Measures including but not limited to percent agreement, Cohen’s kappa (two raters), Fleiss’ kappa (three or more raters), the contingency coefficient, Pearson correlation coefficient, Spearman’s rank coefficient, intra-class correlation coefficient, weighted kappa, and the concordance correlation coefficient will be considered.

Types of studies
A range of study designs will be considered for inclusion in the review. Experimental study designs including quasi-experimental, before and after studies, prospective and retrospective cohort studies, case control studies and analytical cross sectional studies will be considered. This review will also take into account descriptive study designs such as case series, individual case reports and descriptive cross sectional studies.

Methods
Search strategy
The search strategy aims to find both published studies and gray literature. A three-step search strategy will be utilized in this review. An initial limited search of MEDLINE and CINAHL will be undertaken followed by analysis of the text words contained in the title and abstract, and of the index terms used to describe the article. A second search using all identified keywords and index terms will then be undertaken across all included databases. Thirdly, the reference list of all identified reports and articles will be searched for additional studies. Studies published in English language will be considered for inclusion in this review. Studies published between 1990 and 2017 will be considered for inclusion in this review. Studies prior to 1990 will be excluded due to the limited use of CT and MRI for head and neck radiation therapy treatment planning then.

The databases to be searched include: MEDLINE, Embase, Scopus, The Cochrane Database, CINAHL, Health Technology Assessment Database (HTA), the Database of Abstracts of Reviews of Effects (DARE) and Turning Research into Practice (TRIP).

The search for gray literature will include: conference papers, government agency reports and proceedings, clinical trial protocols and resources, Trove, the Agency for Healthcare Research and Quality, MedNar, The Grey Literature Report, Open Grey, Australia and New Zealand Clinical Trials Registry, ClinicalTrials.gov, EU Clinical Trials Registry, World Health Organization, International Clinical Trials Registry Platform and thesis repositories.

Initial keywords to be used will be: organs at risk; observer variation; radiation therapy; head and neck neoplasms; parotid; submandibular; brachial plexus; mandible; glottic larynx; pharynx; pharyngeal constrictor; spinal cord; radiotherapy planning, computer-assisted; radiotherapy, intensity-modulated; tomography, x-ray computed; radiation oncology; magnetic resonance imaging; atlases as topic; anatomic landmarks/ra [radiography]; reproducibility of results.

Assessment of methodological quality
Quantitative papers selected for retrieval will be assessed by two independent reviewers for methodological validity of measurement properties using the COSMIN (COnsensus-based Standards for the selection of health Measurement INstruments) checklist. Any disagreements that arise between the reviewers will be resolved through discussion, or with a third reviewer.

Data extraction
Quantitative data will be extracted using the Guidelines for Reporting Reliability and Agreement Studies (GRRAS) (Appendix I). Data will be extracted
into tabular form and will include specific details about the instruments, populations, measurement properties and outcomes of significance to the review question and specific objectives. Additional information will be captured including a summary of both the reviewer and author conclusions as well as documentation of the normal tissue delineation method used.

**Data synthesis**
Quantitative data will, where possible be pooled in statistical meta-analysis using Joanna Briggs Institute System for the Unified Management, Assessment and Review of Information (JBI SUMARI). All results will be subject to double data entry. Effect sizes expressed as weighted mean differences (for continuous data) and their 95% confidence intervals will be calculated for analysis. For example, the 95% confidence interval for kappa statistics and the weighted mean will be calculated for analysis. Where statistical pooling is not possible the findings will be presented in narrative form including tables and figures to aid in data presentation where appropriate.

**Acknowledgements**
We acknowledge the assistance of Yolanda Surjan and Helen Warren-Forward, Faculty of Health and Medicine, School of Health Sciences, University of Newcastle, Newcastle, Australia.

**References**
Appendix I: Guidelines for reporting reliability and agreement studies (GRRAS)\textsuperscript{43}

<table>
<thead>
<tr>
<th>TITLE AND ABSTRACT</th>
<th>1. Identify in title or abstract that interrater/intrarater reliability or agreement was investigated.</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTRODUCTION</td>
<td>2. Name and describe the diagnostic or measurement device of interest explicitly.</td>
</tr>
<tr>
<td></td>
<td>3. Specify the subject population of interest.</td>
</tr>
<tr>
<td></td>
<td>4. Specify the rater population of interest (if applicable).</td>
</tr>
<tr>
<td></td>
<td>5. Describe what is already known about reliability and agreement and provide a rationale for the study (if applicable).</td>
</tr>
<tr>
<td>METHODS</td>
<td>6. Explain how the sample size was chosen. State the determined number of raters, subjects/objects, and replicate observations.</td>
</tr>
<tr>
<td></td>
<td>7. Describe the sampling method.</td>
</tr>
<tr>
<td></td>
<td>8. Describe the measurement/rating process (e.g. time interval between repeated measurements, availability of clinical information, blinding).</td>
</tr>
<tr>
<td></td>
<td>9. State whether measurements/ratings were conducted independently.</td>
</tr>
<tr>
<td></td>
<td>10. Describe the statistical analysis.</td>
</tr>
<tr>
<td>RESULTS</td>
<td>11. State the actual number of raters and subjects/objects which were included and the number of replicate observations which were conducted.</td>
</tr>
<tr>
<td></td>
<td>12. Describe the sample characteristics of raters and subjects (e.g. training, experience).</td>
</tr>
<tr>
<td></td>
<td>13. Report estimates of reliability and agreement including measures of statistical uncertainty.</td>
</tr>
<tr>
<td>DISCUSSION</td>
<td>14. Discuss the practical relevance of results.</td>
</tr>
<tr>
<td>AUXILIARY MATERIAL</td>
<td>15. Provide detailed results if possible (e.g. online).</td>
</tr>
</tbody>
</table>