Geometric variability of organs at risk in head and neck radiotherapy
Brouwer, Charlotte Louise
Summarizing discussion and future perspectives
The overall purpose of this thesis was to quantify the effect of two of the major factors of uncertainty in current radiotherapy treatment with regard to head and neck organs at risk (OARs): delineation variability and anatomical variability during the course of treatment. Reduction of these uncertainties is of major importance since advanced treatment techniques, such as proton therapy, can be delivered with high conformity and therefore require an adequate definition of structures before and during the course of therapy.

**Variability in delineation**

Variation in delineation of anatomical structures in the head and neck region involved in swallowing was demonstrated by major differences between delineation guidelines for OARs in the head and neck region (Chapter 2). The 3-dimensional dose distributions to some of these swallowing structures can be used as input parameters in Normal Tissue Complication Probability (NTCP) models in order to predict the probability of radiation-induced swallowing dysfunction in individual patients. In this chapter, the variability in NTCP values due to the variation between guidelines and subsequent variation in dose parameters was studied. In some cases, the variation in dose parameters resulted in relatively large absolute differences in anticipated NTCP-values (i.e. >10%). These findings emphasize the importance of using the same delineation guidelines as those that were used to develop and validate the NTCP-models. For a broad introduction of NTCP-models into routine clinical practice, generally accepted consensus guidelines and corresponding NTCP-models are required.

In Chapter 3, interobserver variability within specific delineation guidelines was quantified for a set of head and neck OARs, including the spinal cord, parotid and submandibular glands, thyroid cartilage and glottic larynx by different measures of variability and 3D analysis over sub-regions. The results from this analysis revealed the regions with the largest variability for each OAR. Overall, the largest variability was observed for the cranial region of the OARs. A number of reasons for this variability could be identified, including indistinctness of the delineation guidelines, the lower image resolution in the cranial-caudal direction (determined by the slice thickness), and poor difference in contrast with adjacent tissues. Apart from improving the delineation guidelines, the importance of delineation meetings and training was stressed to optimize compliance to the guidelines.

The main difficulty in the interpretation of results in our but also in other studies that investigated interobserver variability, is the lack of ‘gold standard’ contours. Ideally, surgical specimens could be used to verify the accuracy of delineations as has been done for prostate cancer delineation [1], but obviously this is not an option for normal tissues in the head and neck.

Therefore, the accuracy of OAR delineation is generally derived from interobserver variability relative to other observers, measuring the variation amongst the different contours [2–4]. Commonly used measures include the analysis of volumes [5, 6], the conformity / conformality / (dis)concordance index or Jaccard coefficient [2, 7–10], the Dice Similarity
Coefficient (DSC) [5], and the Hausdorff distance [6]. In addition, the distance of each contour to a reference contour is commonly used, giving a common frame of reference to determine the statistical variation of each of the contouring metrics. The choice of this ‘gold standard’ reference contour varies in the literature from a mathematical average contour [11–13], a STAPLE (simultaneous truth and performance level estimation) consensus contour [14], a radiologist- or experienced oncologist-defined contour [15] or a consensus contour that is decided upon by a panel of experts [16, 17].

There is no consensus on a general set of tools to use to describe interobserver variability in contouring. A combination of measures of volume, position and 3D distance, independent of the number of observers, would provide a comprehensive overview of interobserver variability. Objective measures on interobserver variability are useful for guideline development, and could serve as a benchmark against automated contouring procedures. Valentini et al. [18] presented a set of recommendations with regard to the ontology definition, performance evaluation tools and benchmark evaluation methods. These authors also advice to describe the existing contouring variability from different points of view, using a combination of multiple different tools.

Based on the results described in Chapter 2 and 3, and in collaboration with a number of international experts in the field of head and neck radiation oncology, we proposed CT-based consensus delineation guidelines for head and neck OARs in Chapter 4. These consensus guidelines have been endorsed by all major cooperative head and neck cancer research groups worldwide, including DAHANCA, EORTC Head and Neck Cancer Group, EORTC Radiation Oncology Group, GORTEC, HKNPCSG, NCIC CTG, NCRI, NRG Oncology and TROG. Furthermore, the guidelines have been integrated in ESTRO teaching courses. In addition, the Dutch Platform for Head and Neck Radiotherapy (LPRHHT) decided to use these guidelines in the Netherlands which was further supported by the organization of a workshop for head and neck radiation oncologists in the Netherlands.

At present, over 900 head and neck cancer patients that have been included in the Head and Neck Standardized Follow-up Program (SFP) of the department of radiation oncology of the UMCG from 2007 on, are re-delineated according to the new consensus guidelines to develop a complete set of new NTCP models for head and neck cancer patients (Dutch Cancer Society Project: CITOR).

After the introduction of the consensus guidelines, we demonstrated that the interobserver variability was reduced for most OARs investigated [19]. The average 3D variation in distance for 6 observers before and after the introduction of the guidelines was 3.7 mm and 2.4 mm (1 SD), respectively. Out of 20 OARs, 16 OARs showed reduced 3D variation (reduction range 0.1-10.8 mm) using the consensus guidelines. The largest reduction of 4.3 mm was seen for the oral cavity, from 6.4 mm to 2.1 mm (Figure 1).
In the consensus guidelines, the use of MRI in addition to CT is advised in case of limited discrimination between the OAR and its surrounding tissue, in particular for the brainstem, optic chiasm and optic nerve. The addition of MRI to CT has demonstrated to improve delineation accuracy for GTVs in the head and neck [20, 21]. Also for the parotid glands, MRI would facilitate the delineation of the deep lobe, and allow for an easier distinction between the vascular and the glandular structures of the parotid glands [8].

With the increasing use of MR (and PET) imaging in radiotherapy treatment planning, it is of importance to follow established, consistent image registration procedures. Furthermore, improved geometric accuracy of MRI is required. Phantoms and analysis software have been developed to characterize distortions in MRI scans [22, 23]. Moreover, commercial solutions have become available to measure and correct for geometric distortion.

FIGURE 1

Contours of the oral cavity delineated on CT (left) without (above) and with (below) consensus guidelines. Right side shows 3D variation (1SD).
As stated in Chapter 4, we anticipate that the guidelines will be updated in the near future, incorporating new relevant anatomical and functional information, e.g. using (f)MRI and incorporating (subregions of) OARs that are most relevant for side effects. This is illustrated by a recent publication on the parotid gland, demonstrating that the radiation dose to the region containing the stem and progenitor cells correlated better with post-treatment saliva production than the radiation dose to the whole gland [24].

We expect that the updated guidelines will contribute to a further improvement of NTCP modelling (Figure 2).

Delineation guidelines (I) and accounting for geometric changes (II) will lead to most accurate and consistent DVH parameters (III). This provides the possibility to develop improved NTCP models (IV), which will in turn define the updated set of delineation guidelines (I).
Anatomical variation during the treatment

As stated in the introduction, the number of publications on anatomic and dosimetric changes increased rapidly in the last decade. Therefore, we systematically reviewed the literature to obtain better insight in anatomic and dosimetric changes of head and neck OARs during radiotherapy, including the implications of these changes for the rate and severity of complications and quality of life (Chapter 5).

The results of this review revealed that the parotid gland was the most studied OAR. The parotid glands normally shrink to 60-80% of their initial volume at the end of treatment. Various results regarding dosimetric impact of this shrinkage were reported in the different studies, with a median value of the mean parotid gland dose difference over 25 studies of 1.7 [-1.9-10.4] Gy.

The reason that the parotid glands have been subject to so many studies is twofold. Firstly, the dose to the parotid glands is related to the most frequently reported late radiation induced side effect [24–26], i.e. salivary dysfunction and subsequent xerostomia. Secondly, the parotid glands show large anatomic changes over time during the course of treatment (Chapter 5).

The reason of the large variety in the results of the dosimetric effects can be explained by differences in patient population (e.g., nasopharynx vs. other locations) and differences in treatment techniques (e.g., different margins, OAR-based treatment plan optimization versus conventional treatment planning, etcetera). Another limitation is that most studies only consisted of limited patient populations.

Selection of a subgroup of patients for an adaptive radiotherapy approach would enable adaptive radiotherapy for more radiotherapy departments, since the procedure is still highly labour intensive and thus relatively expensive. Therefore, selection of patients that will benefit most from this approach may be more cost effective and may improve a more widely clinical introduction.

Potential selection criteria for adaptive radiotherapy can be divided in pre-treatment and per-treatment selection factors. We identified a number of potential pre-treatment factors (Chapter 5) including primary tumour site, age, body mass index, planned dose to the parotid glands, the initial parotid gland volume, and the amount of overlap volume of the parotid glands with the clinical target volumes around the lymph node metastases. Potential per-treatment selection factors were weight loss, reduction of lateral neck diameter, parotid gland volume decrease, parotid gland density decrease, distances between parotid gland centres of mass and GTV volume decrease.
An important limitation of the studies performed so far is that many of these consisted of relatively small patient populations and were mainly based on univariable analyses. To identify prognostic factors for volume changes that eventually result in relevant dose distortions, prospective studies are desperately needed including assessment of anatomic and dosimetric changes as well as multivariable analysis. Since only limited data on the effect of dosimetric changes on the risk of radiation-induced side effects and/or on quality of life were found (Chapter 5), these aspects should be taken into account as well in future studies to determine the ultimate clinical impact of adaptive radiotherapy.

**FIGURE 3**

NTCP model Xerostomia 6 months after radiotherapy according to Beetz et al. [25]. A mean dose difference of 10 Gy to the contralateral parotid gland results in a different NTCP difference, depending on the initial dose level.
We conducted a retrospective analysis on prospectively derived data to develop and validate a method to select head and neck cancer patients for adaptive radiotherapy in Chapter 6.

Two different patient cohorts were available for the development and validation of a multivariable prediction model to identify patients with dose deviations to the parotid gland of more than 3 Gy (Δdose3Gy). The only parameter in the multivariable analysis with a significant association with Δdose3Gy was the mean dose to the parotid gland. Using this single parameter, a sensitivity of 80% could be obtained when a threshold of parotid gland Dmean of 22.2 Gy was used. This would spare 24% of the patients from the labour-intensive adaptive radiotherapy procedure.

The question arises whether using a threshold for Δdose to the parotid gland of 3 Gy for patient selection results in a clinically relevant reduction of side effects in terms of the risk of xerostomia. For some patients, a Δdose of more than 3 Gy translates to a clinically relevant ΔNTCP of more than 10%, while in other patients, the same Δdose is not expected to translate in a clinically relevant NTCP reduction (Figure 3). The effect of Δdose on the NTCP depends on the baseline dose level and the NTCP model applied [25, 26]. When proper externally validated NTCP models would be available, the expected ΔNTCP could become part of the selection process for ART, analogous to the selection procedure for proton therapy [27].

**Future perspectives**

The work presented in this thesis focussed on delineation variability for manual contouring and anatomical variability during the course of treatment. Four different other aspects may further reduce uncertainty with regard to head and neck OARs definition and/or administered dose.

First of all, to facilitate uniform contouring procedures, dedicated delineation software may increase consistency among radiation oncologist world-wide. Fortunately, (semi) automatic delineation software for OARs is increasingly applied in routine clinical practice. This software is designed to integrate different imaging modalities and offers optimized contouring tools, such as the use of threshold contouring and the ability to delineate OARs in every desired image orientation. Software dedicated to contouring often offers the ability for (semi-) automatic and/or atlas-based delineation. Application of (semi-) automatic delineation has proven its efficacy for certain OARs [28–35], but clinicians should still carefully review and if necessary adjust every automatically generated contour. At present, efforts are being made to automate the quality assurance of automatic delineation, by using a database of validated contours and review every new contour by comparison with the database contours [36, 37]. Such a system would also be helpful for training purposes. For consistency in retrospective analysis and multicentre studies, it is of importance to use similar validated atlases and deformable image registration algorithms for every patient.
Second, by the increasing use of improved fixation and online image guidance, setup variations have been reduced significantly. Online verification using conebeam CT increased the set-up accuracy, allowing a margin reduction between CTV and PTV of 2-3 mm [38]. It should be emphasized that position verification as such is not able to compensate for anatomical changes. An adaptive procedure is necessary in which the treatment plan is re-optimized for the altered anatomy.

A third aspect that will reduce uncertainty in the definition and administered dose to head and neck OARs is the future application of offline or online adaptive (proton) therapy. Different adaptive radiotherapy strategies including one or two plan adaptations during the treatment have been proposed to be effective to reduce dose to OARs [39–42]. Potentially, adaptive radiotherapy would also enable margin reduction and therewith further reduce the dose to OARs.

We believe that adaptive radiotherapy is an essential tool for a successful introduction of proton therapy, where anatomic changes may have a significantly larger impact on the delivered dose distribution than in photon therapy [43].

Important requirements for adaptive therapy are in-room and/or out-room (CB)CT and/or MR imaging in combination with advanced patient transport systems from CT to gantry, QA controlled deformable image registration for contour propagation and dose warping, fast treatment planning and plan QA, and integration of hard- and software of all of the aforementioned components. We are currently putting efforts into realizing these different aspects for our proton therapy facility (the UMC Groningen Particle Therapy Center (GPTC)). Fortunately, more and more vendors provide software for workflow automation of adaptive radiotherapy. On line 3D image guidance is relatively new in proton therapy as the first cone beam-CT-scan systems are only recently installed.

From a clinical perspective, the exact role of adaptive radiotherapy remains to be defined. Recent studies [44] showed encouraging post-treatment functional recovery with regard to late toxicity (12 months) for 22 patients treated with adaptive radiotherapy. Other studies [45, 46] reported significantly lower incidences of side effects in the patient group treated with adaptive radiotherapy with respect to those observed in the group treated without adaptive radiotherapy. Patients were, however, not randomly selected for the adaptive radiotherapy scheme, and therefore differences between patient groups might be explained by other factors such as differences in socioeconomic status between the two populations. The ESTRO annual meeting of 2016 entitled a proffered paper session ‘Adaptive radiotherapy: Hype or Hope?’ to update the community with recent findings on this subject (including our findings as presented in Chapter 6 [47]). This underlines that the clinical benefit of adaptive radiotherapy is still under discussion in the radiation oncology community. We need to wait for more clinical results on adaptive radiotherapy to get more
knowledge about the effect of adaptive radiotherapy on tumour control and the occurrence of complications.

A fourth and last aspect that will aid to reduce (uncertainty of) dose to OARs is robust treatment planning. The main objective of robust treatment planning is to produce treatment plans that are robust with regard to 3D dose distributions to the different sources of uncertainty. Robust treatment planning originated in proton therapy to account for different sources of range uncertainty [48], but also finds its application in photon therapy [49]. Including robustness into a multicriteria optimisation framework would provide the possibility to balance robustness and conformity [48]. Current implementations of robust treatment planning simulate geometric changes by applying setup errors [49] using scenario-based treatment plan optimization [50]. The latter accounts for different scenarios of geometric uncertainty, and re-optimizes the treatment plan to cover the plan criteria in all of the proposed scenarios. Apparently, uncertainties will be more complex and not only depend on the setup error, but also on the anatomical location, IGRT workflow, type of geometric changes and use of adaptive planning. Further research including analysis of repeated imaging over the course of treatment will be necessary for each specific indication.

The choice of position verification procedures, adaptive radiotherapy schedules and levels of robustness of the treatment plan involves a hierarchy of decisions, each of which balance complexity versus flexibility and conformity versus robustness.

Apart from reducing factors of uncertainty in current radiotherapy treatment, it is also important to improve the accuracy of endpoints to develop reliable NTCP models. In case of patient or physician rated endpoints, a consistent way of scoring and grading is required, which requires adequate instructions. Objectification of endpoints would also contribute in this perspective. For quantification of xerostomia and sticky saliva by image biomarkers, the first studies showed promising results [51]. Changes of biomarkers over time, such as reduction of the parotid gland volume, have also shown to be related to the severity of late xerostomia [52]. So information concerning changes of OARs during RT are not only important because of their consequential dosimetric changes, it could also help us to identify patients that are at excess risk to develop complications.
References


