Inter-observer variability in target delineation increases during adaptive treatment of head-and-neck and lung cancer

Apolle, R.; Appold, S.; Bijl, H. P.; Blanchard, P.; Bussink, J.; Faivre-Finn, C.; Khalifa, J.; Laprie, A.; Lievens, Y.; Madani, I.; Ruffier, A.; de Ruysscher, D.; van Elmpt, W.; Troost, E. G. C.;

Originally published:

July 2019

Acta Oncologica 58(2019)10, 1378-1385

DOI: https://doi.org/10.1080/0284186X.2019.1629017

Inter-observer variability in target delineation increases during adaptive treatment of head-and-neck and lung cancer

Rudi Apolle[1,2], Steffen Appold[1,3], Henk P. Bijl[4], Pierre Blanchard[5], Johan Bussink[6], Corinne Faivre-Finn[7], Jonathan Khalifa[8], Anne Laprie[8], Yolande Lievens[9], Indira Madani[10], Amandine Ruffier[5], Dirk de Ruysscher[11], Wouter van Elmpt[11], Esther G. C. Troost[1,2,3,12,13].

[1]OncoRay – National Center for Radiation Research in Oncology, Faculty of Medicine and University Hospital Carl Gustav Carus- Technische Universität Dresden- Helmholtz-Zentrum Dresden - Rossendorf, Dresden, Germany.

[2]Helmholtz-Zentrum Dresden - Rossendorf, Institute of Radiooncology – OncoRay, Dresden, Germany.

[3]Faculty of Medicine and University Hospital Carl Gustav Carus - Technische Universität Dresden, Department of Radiotherapy and Radiation Oncology, Dresden, Germany.

[4]University Medical Center Groningen, Department of Radiation Oncology, Groningen, The Netherlands.

[5]Gustave Roussy Cancer Campus, Department of Radiation Oncology, Villejuif, France.

[6]Radboud University Medical Center, Department of Radiation Oncology, Nijmegen, The Netherlands.

[7]The University of Manchester and The Christie NHS Foundation Trust, Division of Cancer Science, Manchester, United Kingdom.

[8]Institut Claudius Regaud / Institut Universitaire du Cancer de Toulouse - Oncopole, Department of Radiotherapy, Toulouse, France.

[9]Ghent University Hospital and Ghent University, Radiation Oncology Department, Ghent, Belgium.

[10]University Hospital Zürich, Department of Radiation Oncology, Zürich, Switzerland.

[11]Maastricht University Medical Centre, GROW – School for Oncology and Developmental Biology- Department of Radiation Oncology (MAASTRO), Maastricht, The Netherlands.

[12]German Cancer Consortium DKTK, Partner Site Dresden- and German Cancer Research Center DKFZ- Heidelberg, Dresden, Germany.

[13]National Center for Tumor Diseases NCT, Partner Site Dresden- German Cancer Research Center DKFZ- Heidelberg- Faculty of Medicine and University Hospital Carl Gustav Carus- Technische Universität Dresden- Dresden- and- Helmholtz Association / Helmholtz-Zentrum Dresden - Rossendorf HZDR, Dresden, Germany.

Corresponding author

Rudi Apolle, MSc

OncoRay, National Center for Radiation Research in Oncology

Händelallee 26

01309 Dresden, Germany

Telephone: +49-(0)351-458-6532

E-mail: rudi.apolle1@oncoray.de

Manuscript information

3400 words

2 (+1) figures

1 (+1) tables

30 references

(Numbers in brackets refer to supplementary material)

Key words

Target volume delineation

Inter-observer variability

Adaptive radiotherapy

Head-and-neck squamous cell carcinoma

(Non-)small cell lung cancer

Abstract

Introduction

Inter-observer variability (IOV) in target volume delineation is a well-documented source of geometric uncertainty in radiotherapy. Such variability has not yet been explored in the context of adaptive re-delineation based on imaging data acquired during treatment. We compared IOV in the pre- and mid-treatment setting using expert primary gross tumour volume (GTV) and clinical target volume (CTV) delineations in locoregionally advanced head-and-neck squamous cell carcinoma (HNSCC) and (non-)small cell lung cancer [(N)SCLC].

Materials and Methods

Five and six observers participated in the HNSCC and (N)SCLC arm, respectively, and provided delineations for five cases each. Imaging data consisted of CT studies partly complemented by FDG-PET and was provided in two separate phases for pre- and mid-treatment. Global delineation compatibility was assessed with a volume overlap metric (the Generalised Conformity Index), while local extremes of IOV were identified through the standard deviation of surface distances from observer delineations to a median consensus delineation. Details of delineation procedures, in particular GTV to CTV expansion and adaptation strategies, were collected through a questionnaire.

Results

Volume overlap analysis revealed a worsening of IOV in all but one case per disease site, which failed to reach significance in this small sample (p-value range 0.063-0.125). Changes in agreement were propagated from GTV to CTV delineations, but correlation could not be formally demonstrated. Surface distance based analysis identified longitudinal target extent as a pervasive source of disagreement for HNSCC. High variability in (N)SCLC was often associated with tumours abutting consolidated lung tissue or potentially invading the mediastinum. Adaptation practices were variable between observers with fewer than half stating that they consistently adapted pre-treatment delineations during treatment.

Conclusion

IOV in target volume delineation increases during treatment, where a disparity in institutional adaptation practices adds to the conventional causes of IOV. Consensus guidelines are urgently needed.

Introduction

Target volume delineation is a crucial step in the radiotherapy (RT) workflow and has also long been acknowledged as a major source of geometric uncertainty. It is increasingly based on multi-modal imaging aiming to provide sufficient contrast between tumour and unaffected regions. However, this process relies on individual physicians' experience in cases where this differentiation cannot easily be made. This leads to a large degree of inter-observer variability (IOV), which has been well-documented for many solid tumour sites [1].

Advances in medical imaging technology, standardisation of image acquisition protocols, and consensus guidelines for image interpretation have brought about some reduction of IOV in delineation of the Gross Tumour Volume (GTV), which comprises the demonstrable extent of gross disease [2,3]. Such improvements do not easily alleviate uncertainty in delineating the Clinical Target Volume (CTV), which aims to encompass potential sub-clinical disease surrounding the GTV. Since this cannot be detected by medical imaging, particular regions must be delineated based on their likelihood of involvement, which is rarely known with great certainty [4,5]. The CTV is generally constructed as an expansion of the GTV either by adding a geometrical margin or by including an anatomical area at risk of microscopic involvement.

With the advent of highly conformal dose delivery techniques (e.g. intensity-modulated and light-ion RT) and technologies to manage other sources of geometrical uncertainty (e.g. image guidance), the relative impact of delineation uncertainty on treatment success is increasing. The growing availability and frequency of repeat imaging during treatment allows for changes of both the target and normal tissue anatomy to be detected and subsequently for target volumes and RT plans to be adapted [6,7]. This has the potential to preserve the initially optimised quality of treatments throughout their course, but introduces further uncertainties in interpreting evolving imaging information. There is a lack of experience and data in the field of adaptation of target volumes during RT, which is most likely to be associated with an increase in IOV. An investigation of IOV in adaptive scenarios is therefore justified in order to compare the levels of agreement before and during treatment and to determine potential sources of disagreement arising during treatment.

The present work reports on the RETRACE study (radiooncological evaluation of target and risk structures adaptively contoured by international experts) which was launched in the Spring of 2017. Experts in the field of head and neck and lung cancer were asked to delineate target volumes on pre- and mid-treatment imaging datasets. Their compatibility between observers was evaluated and contrasted between the time points, and local sources of disagreement identified.

Materials and Methods

Patient Cases

Five patients from routine clinical practice were retrospectively selected per disease site, with HNSCC cases retrieved from University Hospital Carl Gustav Carus (Dresden, Germany) and (N)SCLC cases from MAASTRO clinic (Maastricht, The Netherlands). All patients had consented to the sharing of their data in a pseudonymised form for research purposes and ethical approval was granted by the relevant bodies of both institutions [references: EK341082016 (Dresden), P0152 (Maastricht)]. Cases of locoregionally advanced disease of different anatomical subsites were chosen to demonstrate a variety of changes on mid-treatment imaging, which are likely to prompt adaptation (table 1).

Imaging Data and Data Distribution

Imaging data consisted of computed tomography (CT) and [^{18}F]fluorodeoxyglucose positron emission tomography (FDG-PET) studies. While CT imaging was available for all cases and time points, FDG-PET was not always provided (see table 1). CT imaging had an in-plane resolution of approx. 1 × 1 mm^2 and a slice thickness of either 2 or 3 mm and was often acquired without intravenous iodine contrast in order not to jeopardize renal function, particularly in patients receiving concomitant chemotherapy. FDG-PET resolutions were approx. 4 × 4 mm^2 laterally and 3 or 5 mm longitudinally. Combined FDG-PET-CTs were acquired on Biograph

16 or Biograph 40 systems (Siemens Healthineers, Erlangen, Germany) and CT only imaging performed on Somatom Sensation Open or Somatom Definition AS scanners (Siemens Healthineers). Imaging data was stored and distributed to observers via RadPlanBio [8], a research platform hosted by the German Cancer Consortium, which provides authenticated and encrypted data exchange. It was also used to confidentially collect observer delineations in the form of DICOM-RT structure sets.

Observers and Target Volume Delineation

Five and six international radiation oncologists with respective expertise in HNSCC or (N)SCLC were recruited and asked to delineate the primary tumour GTV and CTV. They were instructed to adhere to their routine procedures and utilise their in-house treatment planning system, such that a representative sample of current clinical practice can be obtained. Once pre-treatment delineations were received, mid-treatment imaging data was made available after a period of at least one week. Alongside imaging data, observers were provided with a summary of clinical findings for each case and questionnaires soliciting comments on their delineation process, in particular on GTV to CTV expansion technique and on the influence of pre-treatment delineation on mid-treatment delineations.

Data Preparation

All processing of delineations was carried out in custom software built using facilities provided by the SciPy library [9]. Observer delineations were transformed into binary volumes discretised at the corresponding CT resolution by labelling all voxels whose centre lies inside of or on an observer's delineation. These volumes were subsequently resampled longitudinally to yield a voxel size of approx. 1 × 1 × 1 mm^3. Median consensus delineations were then generated as the collection of all voxels included by at least half the observers.

Delineation Compatibility Assessment

A variety of methods for comparing volume segmentations exists [10,11]. We utilised a metric based on volume overlap for global compatibility assessment and one based on surface distances to identify local extremes of IOV.

Volume Overlap

Given a set of delineations $\{D_i\}$, volume-based metrics are generally computed from their intersection (common delineation) $D_i \cap D_j$, and their union (encompassing delineation) $D_i \cup D_j$. We employed an extension of the Dice Similarity Coefficient for omnibus comparisons, the Generalised Conformity Index (CI_{gen}) defined as:

$$\mathrm{CI}_{\mathrm{gen}} = \frac{\sum_{j>i}\left|D_i \cap D_j\right|}{\sum_{j>i}\left|D_i \cup D_j\right|}$$

, where the sums run over all unique inter-pairings of delineations [12]. Its central advantage lies in the independence of the number of observers, thus enabling a meaningful comparison of cases with incomplete sets of delineations.

Surface Distance

We implemented the Bi-directional Local Distance Measure (BLDM) described by Kim *et al*. [13] as the primary tool to evaluate distances between delineations. It works on surface voxels and is an extension of minimum distance type metrics, designed to handle asymmetric comparisons, wherein forward associations between a voxel on the reference surface p_{ref} and its closest counterpart on the test surface p_{test} are not reproduced when the search is run in reverse. After finding the so-called forward minimum distance at p_{ref}, BLDM identifies all points on the test surface whose closest point on the reference surface is p_{ref}. The largest of this set of reverse minimum distances and the forward minimum distance is then assigned as the surface distance at p_{ref}. The authors did not explicitly state how to treat cases where no test-voxel is reversely associated to p_{ref}, but this is likely to only occur when the forward minimum distance is already the largest distance which can be assigned, and so we did.

BLDM was extended to differentiate between inward and outward associations, by assigning a negative distance to p_{ref} if the associated p_{test} lies on the interior of the reference surface. The set of surface distances between a point on the median consensus delineation and all observer delineations should then be roughly

distributed symmetrically around zero and their surface distance standard deviation (SDSD) be a meaningful measure of local IOV at that point. In order to summarise surface-based IOV for a whole volume, the root-mean-square error (RMSE) was computed from all voxel SDSDs.

Statistical Analysis

IOV was individually quantified for each disease site, case, target volume, and time point, in terms of CI_{gen} and RMSE. Differences in IOV between the two time points were probed with two-sided Wilcoxon signed-rank tests and correlations between GTV and CTV IOV evolution with Spearman rank correlation. A significance criterion of $p<0.05$ applies throughout and all hypothesis tests and correlations were carried out in R (version 3.5.1, [14]).

In order to identify local extremes of IOV, a χ^2 statistic was computed for each surface voxel i of a given consensus delineation [15]:

$$\chi_i^2 = (N_{\text{obs}} - 1)\left(\frac{SDSD_i}{RMSE}\right)^2$$

, where N_{obs} is the number of observers whose delineations are included in the comparison and $RMSE^2$ is taken as an estimate of the surface distance variance. Voxels were then classified as significantly variable if their χ_i^2 exceeded the one-sided $\alpha = 0.001$ critical value of the χ^2 distribution with $N_{\text{obs}} - 1$ degrees of freedom.

Results

A total of 206 delineations were received out of an anticipated 220, with omissions due to a perceived lack of imaging quality (n=10) and observer drop-out (n=4). Mean pre-treatment GTVs aggregated over all cases and observers were 44 ml (range: 1-122) and 271 ml (range: 35-723) for HNSCC and (N)SCLC, respectively. Mid-treatment GTVs were reduced to 34 ml (range: 2-108) and 236 ml (range: 24-1113). Mean pre-treatment CTVs respectively measured 110 ml (range: 11-265) and 447 ml (rage: 90-1073) for HNSCC and (N)SCLC, reducing to 91 ml (range: 13-216) and 412 ml (range: 77-1530) mid-treatment. None of the changes in mean volumes were significant. Supplementary table 1 lists detailed figures per case.

In terms of volume overlap (CI_{gen}), IOV worsened for nearly all target volumes from pre- to mid-treatment, except for HNSCC case 2 whose GTV suffered a reduction of agreement, but whose CTV was delineated marginally more consistently, and (N)SCLC case 1, which saw moderate improvement for both GTV and CTV. This general reduction of agreement failed to reach significance in this small selection of cases. It was most notable for HNSCC GTVs (p=0.063) while the remaining comparisons (HNSCC CTV and both (N)SCLC volumes) all yielded p=0.125. Correlations between CI_{gen} changes of corresponding GTVs and CTVs were hinted at, but not significant, and slightly less pronounced in the HNSCC cohort (ρ=0.8, p=0.133) than for (N)SCLC (ρ=0.9, p=0.083). Figure 1 provides a summary and

illustration of CI_{gen} changes, while detailed results can be found in supplementary table 1.

Surface distance based global IOV (RMSE) was characterised by less noticeable changes from pre- to mid-treatment. All cases but one per anatomical site (HNSCC case 2, (N)SCLC case 3) showed reduced agreement for both GTV and CTV, with changes mostly minute and non-significant (supplementary figure 1, supplementary table 1). In HNSCC locally significant deviations frequently clustered around caudal and cranial borders, both of the target volumes themselves and of anatomical features which observers excluded (e.g. borders of thyroid cartilage). Significantly variable voxels on the GTV surface often had no analogue on the CTV surface. In (N)SCLC large discrepancies were generally associated with consolidated lung tissue or potential invasion of the mediastinum, and significant deviations often propagated from GTV to CTV surfaces. An example of such disecrepancies is shown in Figure 2, (N)SCLC case 5, due to severe lung atelectasis.

Contouring questionnaires revealed differing GTV to CTV expansion practices for the two disease sites. In (N)SCLC almost all observers adopted an isotropic expansion of 5 mm with rare editing for anatomical boundaries at both time points. In the HNSCC cohort, on the other hand, the expansion from GTV to CTV employed larger isotropic margins (mean: 7.6 mm) and CTV delineations were more extensively edited to exclude anatomy not at risk of involvement. Mid-treatment

GTV delineations were more often adapted from their pre-treatment counterparts after image registration in the HNSCC cohort (14/23 vs 9/28 observers and cases), the remainder having been delineated *de novo* on the mid-treatment imaging data. If pre-treatment HNSCC delineations were used after registration, they were consistently adaptated to anatomical changes by two observers, and the mid-treatment CTV extended to cover at least the pre-treatment GTV by one observer. Similar adaption was performed by one and two observers, respectively, for (N)SCLC cases. Observers universally indicated that all provided imaging had been used for delineation, and frequently stated that additional imaging data would have been desirable, in particular magnetic resonance imaging (MRI) for HNSCC and FDG-PET in (N)SCLC cases for which it was not provided. The lack of contrast enhanced CT was cited as a limitation by both groups of observers .

Discussion

We investigated differences in IOV between primary target volume delineations performed on pre- and mid-treatment imaging of HNSCC and (N)SCLC patients. There was a general trend of reduction in agreement for both disease sites and volumes, albeit one which failed to reach significance. Uncertainty in craniocaudal target extent and interpretation of consolidated lung tissue or invasion of the mediastinum were identified as the most common causes of IOV in HNSCC and (N)SCLC, respectively, at either stage.

Delineation accuracy is dependent on the quality and tumour-specific contrast of the imaging data. Mid-treatment imaging studies will likely remain at a disadvantage in this regard, due to the lower availability and greater cost of high-contrast imaging modalities (i.e. PET and MRI). They are also sensitive to general treatment induced changes (e.g. increased perfusion), further complicating their interpretation.

The complete lack of FDG-PET imaging for mid-treatment HNSCC delineations is a possible explanation of the IOV increase seen for GTVs there, which was the most drastic change found in this study. In clinical practice, however, FDG-PET acquisition prior to treatment adaptation is rarely performed, for it is scarcely available at short notice and holds significant drawbacks caused by peritumoral inflammation [16]. Investigations of the impact of FDG-PET on IOV in pre-treatment HNSCC target

delineation are rare, but the few that have been performed report no significant change based on GTV size comparisons alone while acknowledging that more experience and training is needed to fully utilise the additional information offered by this then relatively novel modality [17,18]. More faith is placed in the inclusion of MRI for certain HNSCC subsites (e.g. oropharynx, oral cavity) with community guidelines recommending it more highly [19]. Its effect on pre-treatment IOV is debateable, however, with no significant changes found in primary GTVs for HNSCC of the pharynx and larynx by Geets *et al*. [20] based on volume overlap analyses. More recently Rasch *et al*. [21] did find improved agreement in nasopharyngeal CTVs when providing co-registered MRI alongside CT imaging, as well as giving more precise instructions, to which the authors attribute much of the improvement prior to treatment. They originally found surface distance based IOV to be highest at the inferior target edge, which was successfully reduced by instructing observers to consider non-axial image reconstructions. MRI is uniquely capable of providing primary non-axial acquisitions with excellent longitudinal resolution and might thus be well placed to alleviate the largest local source of IOV found also in this study.

There is a stronger evidence for the benefit of FDG-PET in NSCLC, in particular in the pre-treatment setting. Steenbakkers *et al*. [3] found a general reduction in IOV when adding co-registered pre-treatment FDG-PET to CT imaging. They were able to show that local IOV was initially largest at tumour-atelectasis boundaries, by performing surface distance based analyses on sub-regions of the median surface

classified by interface type. This source of IOV was most strikingly reduced by the additional information and specific instructions to exclude atelectatic regions showing no FDG uptake, but still remained considerable. A similar analysis by Karki *et al*. [22] confirmed these findings and also investigated the use of MRI for GTV delineation. They found FDG-PET-based delineations to be most consistent among observers for all interface types except where the target is adjacent to normal lung parenchyma where CT imaging had a slight advantage. MRI did not offer IOV improvement over FDG-PET-CT with the MR sequences trialled, but was conjectured to be of benefit in scenarios where FDG-PET specificity is limited (e.g. inflammation). Perplexingly, the instances affected by atelectasis in this study were among the most highly variable despite the availability of FDG-PET. This suggests that additional imaging information can only be utilised to reduce IOV if complemented with specific guidelines for interpretation. Atelectasis is not only a major source of disagreement, it is also common and dynamic. Kwint *et al*. [23] found nearly 20% of a series of 1800 mid-treatment cone beam CTs to demonstrate development or resolution of atelectasis. Tumour regression and progression were observed in 35% and 10%, respectively, signalling a great potential for adaptive target re-delineation. Such changes were generally well reacted to in our study, but no comment can be made on the specific benefit of FDG-PET in this regard.

It must be acknowledged that IOV reduction does not necessarily improve delineation accuracy, which can only be directly assessed by comparisons of

imaging-derived target volumes against macro- and microscopic disease extent found in resection specimen. Some such efforts have been undertaken to determine the accuracy of various imaging modalities and segmentation methods for GTV definition [e.g. 24-26]. Studies have also been conducted to find pathology-guided CTV margins, either based on the distances of macroscopic tumour extension beyond GTVs defined on various imaging modalities [27], or considering microscopic extension around macroscopic tumour borders [28,29]. These and the very few similar studies have demonstrated that validation against a true reference is possible, but plagued by many difficulties, including small sample sizes, the laborious correction of tissue deformations, and differences in the spectrum of disease suitable for resection and that typically treated with RT. The resulting scarcity of high-quality measurements hinders their acceptance and utilisation in clinical practice and more detailed studies are urgently needed. This is especially true for resected tumours which have undergone neo-adjuvant RT, since very little is known about the correspondence between macroscopic imaging changes and the evolving spatial distribution of microscopic disease under therapy.

Differing adaptation practices are likely to have contributed to the increase in IOV observed during therapy, with a few observers conservatively maintaining coverage of pre-treatment targets, while others redefined targets more freely based on new imaging information. This is where consensus building is most acutely required, and it should start with a comprehensive categorization of relevant intra-therapeutic

changes for different disease sites and imaging modalities. Options for treatment adaptation could then be recommended based on the capability of the available imaging modalities to detect those changes, which will ensure that guidance is available even if the most state-of-the-art imaging is not. Moreover, a framework of adaptation options should be developed in order to provide a systematic nomenclature for the continuing implementation and subsequent evaluation of adaptive RT procedures. Current developments are expected to lead to increased utilisation of treatment adaptation either to counteract delivery uncertainties (light-ion RT) or due to an increased availability of high-quality mid-treatment imaging (integrated MR-guidance). This must be supported by community-driven consensus guidelines if we are to take meaningful advantage of it.

One limitation of this study was its small sample size, resulting in severely underpowered hypothesis tests. Our main instrument for detecting significant changes in IOV (Wilcoxon signed-rank test on CI_{gen}) can only yield $p \geq 0.063$ with five cases, but more powerful parametric alternatives cannot credibly be employed. When testing for CI_{gen} changes pooled across disease sites, one obtains p-values of 0.006 and 0.014 for GTVs and CTVs, respectively. While not entirely meaningful, these results demonstrate that IOV increase during therapy is a significant effect overall. It must also be recognised that most comparison metrics require a reference against which to compare observer delineations, yet no true reference exists. While the volume overlap metric is reference-free, surface distance based

assessments depended on the median consensus delineation as a reference. Since only the spread of measurements relative to it was ultimately analysed, its influence on the final result was reduced, but remains a limitation. Another limitation lies in the quality of imaging provided. It often fell short of current recommendations [19,30], especially the lack of iodine contrast in CT studies and the range of other modalities offered.

In summary, this study has demonstrated that IOV in target volume delineation increases during treatment. While the prominent causes of disagreement are similar at both phases, additional IOV is introduced at mid-treatment due to differing adaptation practices. This disparity and its underlying uncertainties regarding the development of macro- and microscopic target volumes under treatment require more detailed investigations, while high-contrast and comprehensive imaging should be acquired whenever possible to provide clarity in scenarios known to lead to large IOV without it.

Acknowledgments

The authors thank Dr. Tomas Skripcak for the efficient operation of the RadPlanBio data exchange platform and his prompt assistance with all related issues.

Disclosure of interest

The authors report no conflicts of interest.

References

[1] Vinod S, Jameson M, Min M, et al. Uncertainties in volume delineation in radiation oncology: A systematic review and recommendations for future studies. Radiother Oncol. 2016;121:169-179.

[2] Berson AM, Stein NF, Riegel AC, et al. Variability of gross tumor volume delineation in head-and-neck cancer using PET/CT fusion, Part II: the impact of a contouring protocol. Med Dosim. 2009;34: 30–35.

[3] Steenbakkers R, Duppen J, Fitton I, et al. Reduction of observer variation using matched CT-PET for lung cancer delineation: a three-dimensional analysis. Int J Radiat Oncol Biol Phys. 2006;64:435–448.

[4] Moghaddasi F, Bezak E, Marcu L. Current challenges in clinical target volume definition: tumour margins and microscopic extensions. Acta Oncol. 2012;51: 984–995.

[5] Apolle R, Rehm M, Bortfeld T, et al. The clinical target volume in lung, head-and-neck, and esophageal cancer: Lessons from pathological measurement and recurrence analysis. Clinic Transl Radiat Oncol. 2017;3:1–8.

[6] Surucu M, Shah KK, Roeske JC, et al. Adaptive Radiotherapy for Head and Neck Cancer. Technol Cancer Res Treat. 2017;16:218–223.

[7] Møller D, Holt M, Alber M, et al. Adaptive radiotherapy for advanced lung cancer ensures target coverage and decreases lung dose. Radiother Oncol. 2016;121:32-38.

[8] Skripcak T, Just U, Simon M, et al. Toward Distributed Conduction of Large-Scale Studies in Radiation Therapy and Oncology: Open-Source System Integration Approach. IEEE J Biomedic Health Inform. 2016;20:1397–1403.

[9] Jones E, Oliphant E, Peterson P, et al. SciPy: Open Source Scientific Tools for Python. 2001; https://www.scipy.org [online; accessed 04.03.2019]

[10] Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med Imaging. 2015;15:29.

[11] Jameson M, Holloway L, Vial P, et al. A review of methods of analysis in contouring studies for radiation oncology. J Med Imaging Radiat Oncol. 2010;54:401–410.

[12] Kouwenhoven E, Giezen M, Struikmans H. Measuring the similarity of target volume delineations independent of the number of observers. Phys Med Biol. 2009;54:2863–2873.

[13] Kim HS, Park SB, Lo SS, et al. Bidirectional local distance measure for comparing segmentations. Med Phys. 2012;39:6779–6790.

[14] R Core Team. R: A Language and Environment for Statistical Computing. 2018; https://www.r-project.org [online; accessed 04.03.2019]

[15] Wu J, Murphy MJ, Weiss E, et al. Development of a population-based model of surface segmentation uncertainties for uncertainty-weighted deformable image registrations. Med Phys. 2010;37:607–614.

[16] Troost EGC, Bussink J, Slootweg PJ, et al. Histopathologic validation of 3'-deoxy-3'-18F-fluorothymidine PET in squamous cell carcinoma of the oral cavity. J Nucl Med. 2010;51:713–719.

[17] Riegel AC, Berson AM, Destian S, et al. Variability of gross tumor volume delineation in head-and-neck cancer using CT and PET/CT fusion. Int J Radiat Oncol Biol Phys. 2006;65:726–732.

[18] Breen SL, Publicover J, de Silva S, et al. Intraobserver and interobserver variability in GTV delineation on FDG-PET-CT images of head and neck cancers. Int J Radiat Oncol Biol Phys. 2007;68:763–770.

[19] Grégoire V, Evans M, Le Q, et al. Delineation of the primary tumour Clinical Target Volumes (CTV-P) in laryngeal, hypopharyngeal, oropharyngeal and oral cavity squamous cell carcinoma: AIRO, CACA, DAHANCA, EORTC, GEORCC, GORTEC, HKNPCSG, HNCIG, IAG-KHT, LPRHHT, NCIC CTG, NCRI, NRG Oncology, PHNS, SBRT, SOMERA, SRO, SSHNO, TROG consensus guidelines. Radiother Oncol. 2018;126:3–24.

[20] Geets X, Daisne J-F, Arcangeli S, et al. Inter-observer variability in the delineation of pharyngo-laryngeal tumor, parotid glands and cervical spinal cord: comparison between CT-scan and MRI. Radiother Oncol. 2005;77:25–31.

[21] Rasch CRN, Steenbakkers RJHM, Fitton I, et al. Decreased 3D observer variation with matched CT-MRI, for target delineation in Nasopharynx cancer. Radiat Oncol. 2010;5:21.

[22] Karki K, Saraiya S, Hugo GD, et al. Variabilities of Magnetic Resonance Imaging-, Computed Tomography-, and Positron Emission Tomography-Computed Tomography-Based Tumor and Lymph Node Delineations for Lung Cancer Radiation Therapy Planning. Int J Radiat Oncol Biol Phys. 2017;99:80–89.

[23] Kwint M, Conijn S, Schaake E, et al. Intra thoracic anatomical changes in lung cancer patients during the course of radiotherapy. Radiother Oncol. 2014;113:392–397.

[24] Daisne J, Duprez T, Weynand B, et al. Tumor volume in pharyngolaryngeal squamous cell carcinoma: comparison at CT, MR imaging, and FDG PET and validation with surgical specimen. Radiology 2004;233:93–100.

[25] van Baardwijk A, Bosmans G, Boersma L, et al. PET-CT-based auto-contouring in non-small-cell lung cancer correlates with pathology and reduces interobserver variability in the delineation of the primary tumor and involved nodal volumes. Int J Radiat Oncol Biol Phys. 2007;68:771–778.

[26] Wanet M, Lee J, Weynand B, et al. Gradient-based delineation of the primary GTV on FDG-PET in non-small cell lung cancer: a comparison with threshold-based approaches, CT and surgical specimens. Radiother Oncol. 2011;98:117–125.

[27] Ligtenberg H, Jager EA, Caldas-Magalhaes J, et al. Modality-specific target definition for laryngeal and hypopharyngeal cancer on FDG-PET, CT and MRI. Radiother Oncol. 2017;123:63–70.

[28] Giraud P, Antoine M, Larrouy A, et al. Evaluation of microscopic tumor extension in non-small-cell lung cancer for three-dimensional conformal radiotherapy planning. Int J Radiat Oncol Biol Phys. 2000;48:1015–1024.

[29] van Loon J, Siedschlag C, Stroom J, et al. Microscopic disease extension in three dimensions for non-small-cell lung cancer: development of a prediction model using pathology-validated positron emission tomography and computed tomography features. Int J Radiat Oncol Biol Phys. 2012;82:448–456.

[30] Nestle U, de Ruysscher D, Ricardi U, et al. ESTRO ACROP guidelines for target volume definition in the treatment of locally advanced non-small cell lung cancer. Radiother Oncol. 2018;127:1–5.

Tables

Table 1: Overview of cases and provided imaging data.

Disease site	Case	Tumour characteristics		PET availability		MT imaging time
		Location	Classification*	PT	MT	[treatment week]
HNSCC	1	Oropharynx	T4N2M0	yes	no	4
	2	Hypopharynx	T4N3M0	yes	no	4
	3	Palatine tonsils	T1N2bM0	no	no	5
	4	Larynx	T3N2cM0	yes	no	4
	5	Oropharynx	T4N2bM0	yes	no	4
(N)SCLC	1	Left pulm. hilum	T4N2M0	yes	yes	2
	2	Right upper lobe	T3N0M0	no	no	3
	3†	Right pulm. hilum	T4N2M0	no	no	2
	4	Left upper lobe	T4N0M0	yes	yes	1
	5	Right lower lobe	T3N0M0	no	yes	1

Abbreviations: PET, positron emission tomography; (P/M)T, (pre/mid)-treatment; pulm., pulmonary.
Notes: *clinical assignments cT cN cM; †small cell lung cancer.

Supplementary table 1: Delineated volumes and their agreement per case and time point.

Disease site	Case	Delineation summaries													
		N_{obs}		Volume [ml] (mean±SD)				CI_{gen}				RMSE [mm]			
				GTV		CTV		GTV		CTV		GTV		CTV	
		PT	MT	PT	MT	PT	MT	PT	MT	PT	MT	PT	MT	PT	MT
HNSCC	1	5	5	88 ± 23	69 ± 28	186 ± 23	156 ± 31	0.59	0.46	0.69	0.63	2.9	3.1	2.6	2.8
	2	3	4	39 ± 14	36 ± 17	109 ± 40	104 ± 32	0.49	0.40	0.50	0.50	7.0	3.8	8.6	4.5
	3	5	5	5 ± 3	5 ± 3	25 ± 8	24 ± 8	0.34	0.29	0.44	0.41	3.4	3.6	3.3	3.4
	4	5	5	15 ± 3	13 ± 5	54 ± 16	49 ± 18	0.61	0.40	0.51	0.47	1.4	3.2	2.8	3.0
	5	5	4	71 ± 19	51 ± 27	178 ± 48	133 ± 50	0.55	0.33	0.61	0.47	2.5	4.7	3.3	4.4
(N)SCLC	1	6	6	97 ± 70	66 ± 32	207 ± 155	179 ± 105	0.33	0.41	0.35	0.38	5.5	10.4	7.0	10.5
	2	6	6	296 ± 27	190 ± 29	469 ± 50	354 ± 76	0.78	0.66	0.77	0.64	2.1	3.0	2.7	4.2
	3	4	5	161 ± 39	90 ± 47	336 ± 119	226 ± 98	0.55	0.36	0.50	0.44	5.9	4.3	6.6	6.4
	4	6	6	643 ± 59	680 ± 209	942 ± 92	1035 ± 256	0.77	0.55	0.79	0.58	3.0	5.6	3.2	6.4
	5	6	6	120 ± 15	130 ± 79	243 ± 52	235 ± 122	0.73	0.30	0.69	0.35	2.0	14.0	2.9	14.0

Abbreviations/Symbols: (P/M)T, (pre/mid)-treatment; N_{obs}, Number of observers providing delineations; SD, standard deviation; CI_{gen}, generalised conformity index; RMSE, root-mean-square error.

Figures

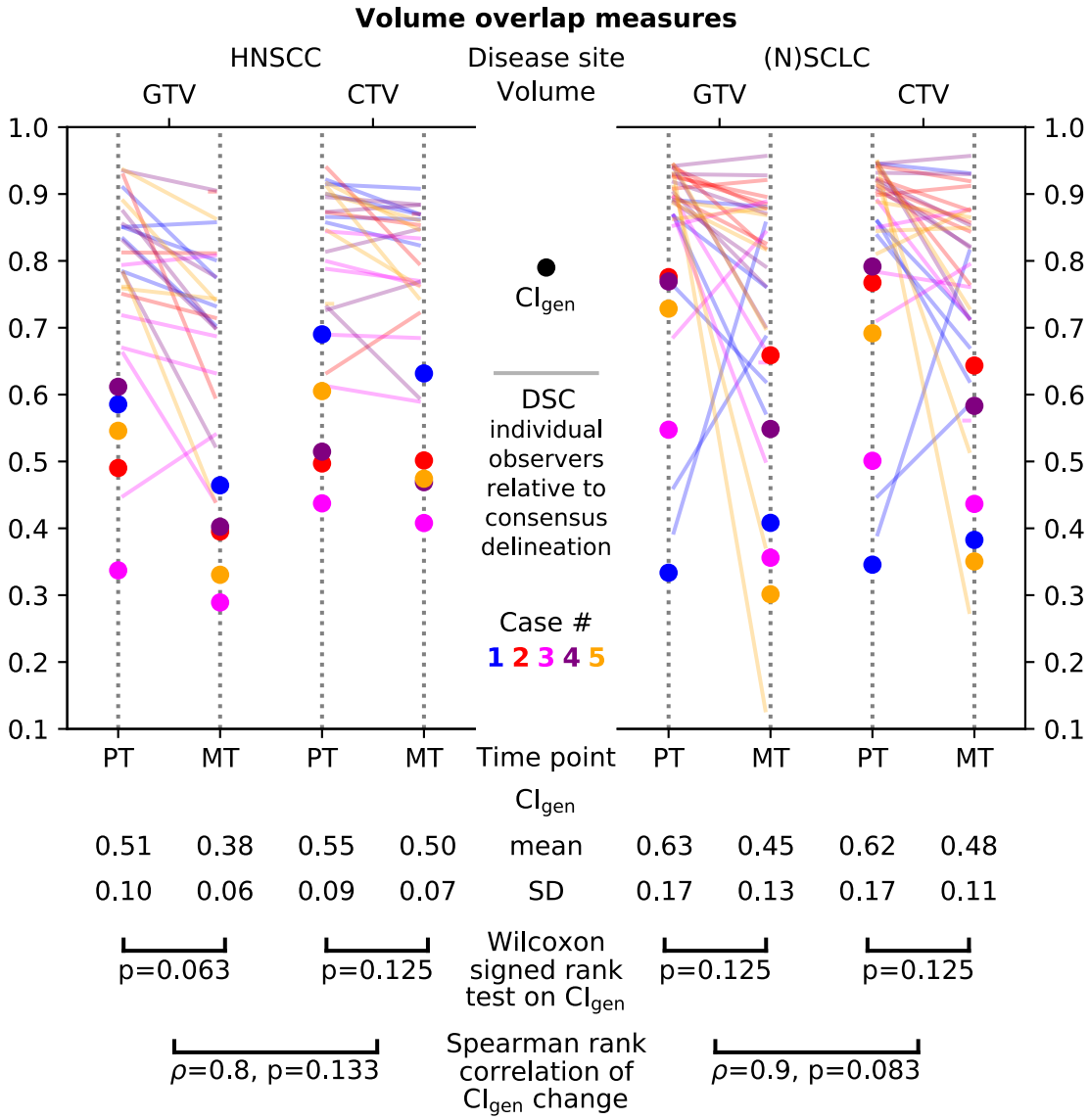


Figure 1: Statistical overview of delineation overlap metrics. The Generalised Conformity Index (CI_{gen}) is shown for each case and time point for Gross Tumour Volume (GTV) and Clinical Target Volume (CTV) delineations. Its summary statistics are printed underneath alongside the results of 1) hypothesis tests for a non-zero change in CI_{gen} when transitioning from pre- to mid-treatment time points (PT to MT), and 2) the correlations of those CI_{gen} changes between GTV and CTV

delineations. The changes in Dice Similarity Coefficients (DSC) for each observer relative to a median consensus delineation (i.e. the collection of voxels delineated by at least half the observers) are depicted for comparison.

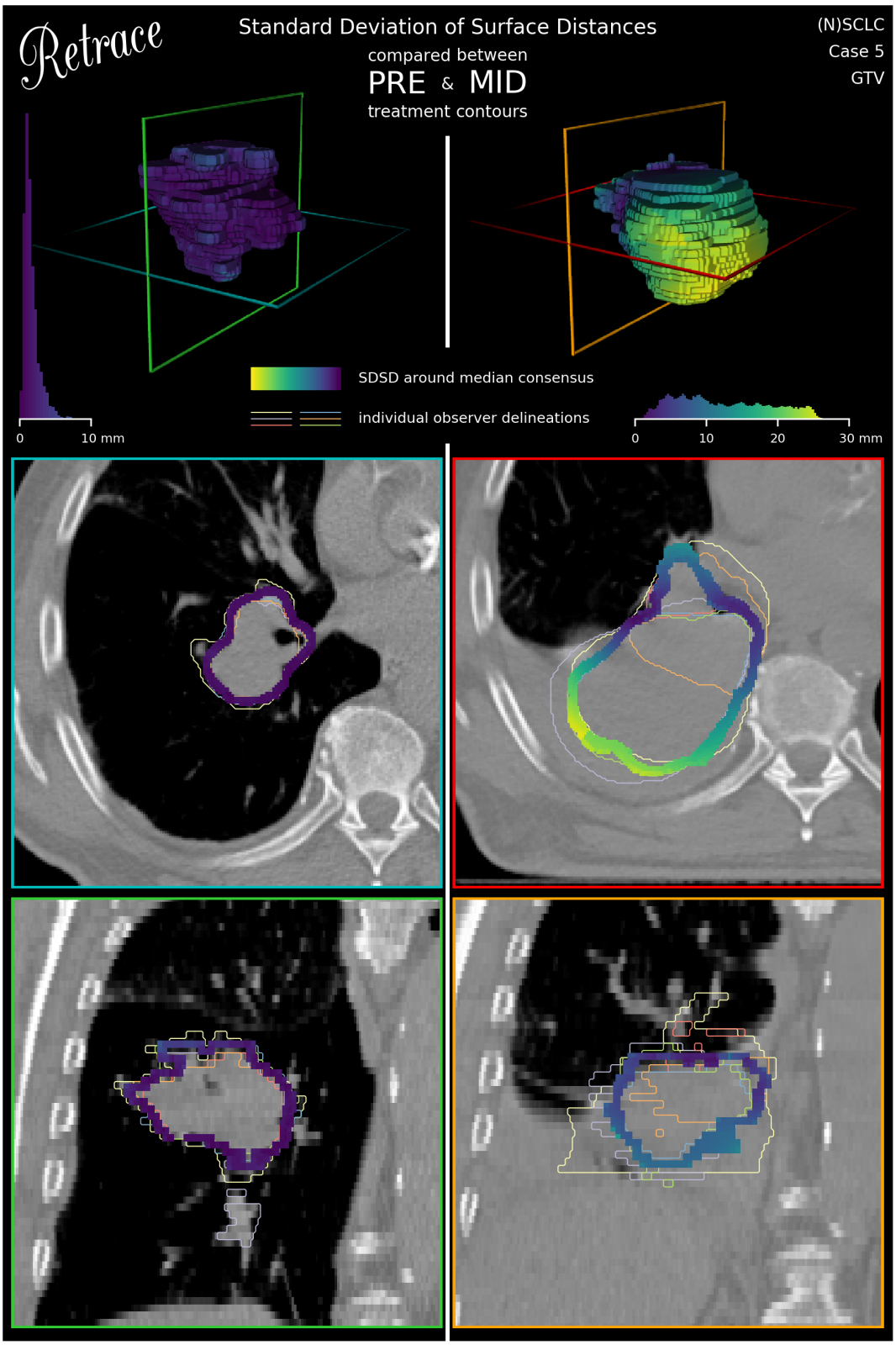
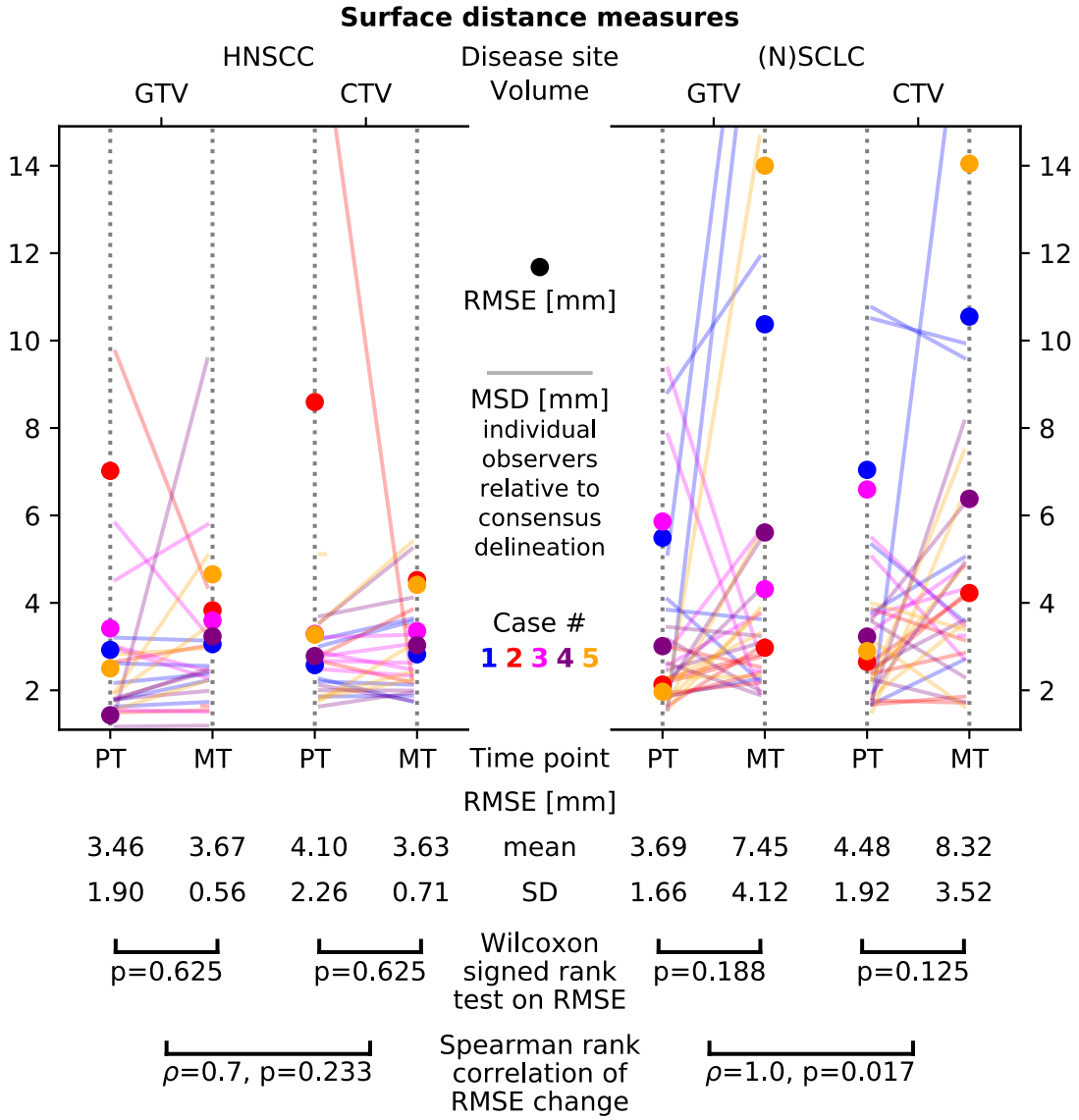


Figure 2: Illustration of surface distance based IOV in GTV delineation for a cT3N0M0 NSCLC of the right lower lobe before (left column) and during (right

column) treatment. The local standard deviation of surface distances from observers' delineations to the median consensus delineation (SDSD) is displayed on the median consensus surface (top row) in a saturated colour scale. Axial (middle row) and coronal (bottom row) CT slices are superimposed with the median consensus delineation alongside individual observer delineations drawn in pastels. Distributions of SDSD values and slice positions are shown in the top row.



Supplementary figure 1: Statistical overview of surface distance based compatibility metrics. The standard deviations of distances from observer delineations to a median consensus delineation (i.e. the collection of voxels delineated by at least half the observers) are computed per voxel and aggregated over the whole volume as a root-mean-square error (RMSE). RMSE is shown for each case and time point for Gross Tumour Volume (GTV) and Clinical Target Volume (CTV) delineations. Its summary statistics are printed underneath alongside

the results of 1) hypothesis tests for a non-zero change in RMSE when transitioning from pre- to mid-treatment time points (PT to MT), and 2) the correlations of those RMSE changes between GTV and CTV delineations. The evolution of observers' Mean Surface Distance (MSD) over the whole surface relative to the consensus delineation are shown for comparison.