Comparison of Investigator-Delineated Gross Tumor Volumes and Quality Assurance in Pancreatic Cancer: Analysis of the Pretrial Benchmark Case for the SCALOP Trial

Emmanouil Fokas[1*], Charlotte Clifford[2*], Emiliano Spezi[3], George Joseph[4], Jennifer Branagan[5], Chris Hurt[6], Lisette Nixon[6], Ross Abrams[7], John Staffurth[8#], Somnath Mukherjee[1],

[1]Department of Oncology, CRUK/MRC Oxford Institute for Radiation Oncology; NIHR Biomedical Research, Oxford, UK

[2]Radiotherapy Physics Department, University Hospitals Birmingham NHS Foundation Trust, Queen Elizabeth Hospital, Queen Elizabeth Medical Centre, Birmingham,UK

[3]School of Engineering, Cardiff University, Cardiff, UK

[4]Velindre Cancer Centre, Velindre Hospital, Cardiff, UK

[5]Oncology Department, Northampton General Hospital, Northampton, UK

[6]Wales Cancer Trials Unit, School of Medicine, Cardiff University, Cardiff, UK

[7]Department of Radiation Oncology, Rush University Medical Center, Chicago, IL, USA

[8]Institute of Cancer and Genetics, Cardiff University, Cardiff, UK, Cardiff NCRI RTTQA Centre, Velindre NHS Trust, Cardiff, UK

Short title: Volume Delineation in the SCALOP Trial

Address for Reprint Requests
Emmanouil Fokas, MD DPhil
Department of Oncology
CRUK/MRC Oxford Institute for Radiation Oncology
University of Oxford
Old Road Campus Research Building
Roosevelt Drive
Headington, Oxford
OX3 7DQ, UK
Phone: +44-1865-225832
Fax: +44-1865-857127
E-mail: emmanouil.fokas@oncology.ox.ac.uk

[*] These two authors are joint first authors; [#] These two authors are joint last authors.

Total number of pages: 13; **Total number of main tables and figures:** 6;

ABSTRACT

Background and Purpose: To evaluate the variation in investigator-delineated volumes and assess plans from the radiotherapy trial quality assurance (RTTQA) program of SCALOP, a phase II trial in locally advanced pancreatic cancer.

Materials and Methods: Participating investigators (n=25) outlined a pre-trial benchmark case as per RT protocol, and the accuracy of investigators' GTV (iGTV) and PTV (iPTV) was evaluated, against the trials team-defined gold standard GTV (gsGTV) and PTV (gsPTV), using both qualitative and geometric analysis. The median Jaccard Conformity Index (JCI) and Geographical Miss Index (GMI) were calculated. Participating RT centres also submitted a radiotherapy plan for this benchmark case, which was centrally reviewed against protocol-defined constraints.

Results: Twenty-five investigator-defined contours were evaluated. The median JCI and GMI of iGTVs were 0.57 (IQR: 0.51-0.65) and 0.26 (IQR: 0.15-0.40). For iPTVs, these were 0.75 (IQR: 0.71 to 0.79) and 0.14 (IQR: 0.11-0.22) respectively. Qualitatative analysis showed largest variation at the tumor edges and failure to recognize a peri-pancreatic lymph node. There were no major protocol deviations in RT planning, but three minor PTV coverage deviations were identified. .

Conclusions: SCALOP demonstrated considerable variation in iGTV delineation. RTTQA workshops and real-time central review of delineations are needed in future trials.

INTRODUCTION

Radiotherapy (RT) Trial Quality Assurance (RTTQA) is an important component of radiation trials to ensure delivery of high quality treatment and allow consistency in outcome among multiple centres [1, 2]. In that context, RT protocols, outlining atlases and benchmark cases allow investigators to familiarize themselves with the radiation protocol and demonstrate their ability to execute protocol treatment. At least two studies, including one in pancreatic cancer, have shown that RT protocol violations may adversely affect overall survival [1, 3]. Sub-analysis of the RTOG 9704, a study of adjuvant chemoradiotherapy (CRT) in pancreatic cancer showed that the clinical outcome of patients treated as per radiotherapy protocol was significantly superior to those treated outside the protocol (overall survival: 1.74 vs. 1.46 years, p=0.0077; local failure: p=0.016).

Previous RTTQA studies have principally focused on the quality of radiotherapy plans, with little emphasis on testing the quality of tumour outlines [4]. The quality of tumour delineation is likely to have major impact on toxicity and survival outcomes [4], especially in a region like pancreas where the tumour is often difficult to delineate and is related to critical organs at risk (OARs). We therefore wanted to ascertain the degree of variability amongst investigator contours within a multi-centre pancreatic radiation trial despite the availability of a detailed RT protocol including atlas. We hypothesized that demonstrating a large variation in contours would have major implication for conduct of any future pancreatic radiation trials, as this would highlight the need for clinician education through regular outlining workshops and consideration of real-time central review of tumour contours before trial participants begin radiotherapy.

The Selective CRT in Advanced Localized pancreatic cancer study (SCALOP I, EudraCT No: 2008-001394-1) was a multi-centre randomized phase II trial where patients received four months of gemcitabine/capecitabine chemotherapy followed by randomization to either gemcitabine- or capecitabine-based CRT [5]. One hundred fourteen patients from 28 centres were registered, and 74 received CRT. Radiotherapy consisted of 50.4Gy in 28 fractions delivered conformally; dose of concomitant gemcitabine was 300mg/m^2 weekly and dose of capecitabine was 830mg/m^2 administered twice daily on days of radiotherapy. Pre-trial component of the SCALOP RTTQA included a pretrial benchmark case. The principal radiation oncologists from centres participating in the trial were required to outline the tumour volumes, and corresponding RT plans were also submitted. Herein we report the variation in investigator outlines and quality of RT plans from the bench-mark case of the

SCALOP trial. We also evaluated the applicability of a novel concept, the planning assessment form (PAF), a paper-based proforma which allowed rapid, real-time review of adherence to protocol constraints.

MATERIALS AND METHODS

SCALOP trial RTTQA program

The details of the SCALOP RTTQA Program are described in Supplementary Methods, and the full protocol can be accessed at http://www.wctu.org.uk/trial.php?trial=scalop. The protocol had a detailed radiotherapy section including a detailed slice-by-slice atlas of delineated 'head' and 'body/tail' tumours and organs at risk. The study also required the participating centres to outline and plan a 'benchmark case, from a patient with a pancreatic head tumour scanned as per the SCALOP RT protocol, reviewed and agreed by the SCALOP RTTQA group. The planning CT scan dataset in Digital Imaging and Communications in Medicine (DICOM)-RT format was sent to all participating centres for tumour volume delineation by the principal radiation oncologist investigator, with input from a radiologist with expertise in gastrointestinal cancer. All investigators (n=25) were asked to follow the trial protocol and outline the gross tumor volume (iGTV) and create a planning target volume (iPTV) directly from the iGTV by the addition of treatment planning system (TPS) generated margins as per the SCALOP protocol: 2.0cm in superior-inferior direction and 1.5cm in all other directions. The centers also generated a plan following the protocol stipulated dose/volume constraints and filled in a planning assessment form (PAF - described in Supplementary Methods; shown in Supplementary Figure 1). In summary, the PAF was a tabulated form where protocol-specific dose constraints were pre-populated with corresponding spaces to allow the investigator to record the constraint dose achieved in that particular case. The pre-populated values served as a quick check-list of the protocol constraints to ensure these limits were not breached. Once completed, the DICOM-RT dataset including all outlined target volumes and organs at risk (OARs), the 3D dose cube and the PAF were submitted for central review.

Review of tumour outlines of the test case

In order to facilitate review of tumour outlines, a 'gold standard' set of reference structures (gsGTV, gsPTV) were created by the Chief Investigator (SM) and the Chief trial radiologist (GJ). The iGTVs and iPTVs of the individual investigators (n=25) were compared against gsGTV and gsPTV using qualitative and geometric analysis.

Qualitative Analysis

For qualitative analysis, the n=25 DICOM-RT datasets from the 25 centers that participated in the pre-trial benchmark contouring case (one investigator per center) were individually imported into the ImSimQA^{contour} software (Oncology Systems Limited,UK), allowing superposition of investigator contours against the gsGTV and gsPTV. Three reviewers, including a gastrointestinal radiologist, visually evaluated each iGTV against the gsGTV on a slice-by-slice basis. Each slice of each iGTV was scored as significant over-contour (outlining of unequivocal OAR, usually normal pancreas, stomach or duodenum), significant under-contour (omission of unequivocal tumor) or as acceptable variation (outlining variation within acceptable limits on visual inspection, where the deviation from the gold standard was felt to be unlikely to result in compromised tumour control, or excessive toxicity).

Geometric Analysis

For quantitative analysis, the investigators' DICOM-RT datasets were individually imported into Computational Environment for Radiotherapy Research (CERR) [6]. CERR was used to assess the whole volume iGTVs and iPTVs for over- and under-outlining compared to the respective gold standard volumes. We used the Jaccard Conformity Index (JCI) and Geographical Miss Index (GMI) [7], defined in Supplementary Table 1. The iGTV was then re-analyzed in CERR on a slice-by-slice basis generating the JCI, GMI and Mean Distance to Conformity (MDC) for each slice as we described previously [8, 9].

Review of RT planning of the test case

The DICOM-RT planning datasets for each submitted contour were centrally reviewed using the Visualization and Organization of Data for Cancer Analysis program (VODCA 4, Medical Software Solutions, Hagendorn, Switzerland), a research software that can read DICOM-RT datasets from multiple planning systems. Plan quality and adherence to dose volume constraints were assessed. We also evaluated the accuracy of PAF completion by comparing figures enumerated on PAF against the actual dose values generated from VODCA. Investigating centers were provided individualized written feedback on the performance of the benchmark case. Real-time central review of each patient's iGTV was not conducted but on-trial review of radiotherapy planning was performed through assessment of the PAF before each patient received radiotherapy. There was high compliance to all aspects of dose prescription to PTV with only three minor deviations reported using the PAF (Table

1); all of these were in the volume of PTV receiving 95% of the prescribed dose falling below the required 99% (95.1%, 96.7% and 97.6%, respectively).

Statistical Analysis

The Stata 13 package (StatCorp LP, College Station, Texas, USA) was used to conduct statistical analyses. Variables were assessed for their normality using the Shapiro-Wilk test before summary statistics and a number of correlations were assessed. For variables that were normally distributed the mean and standard deviation (SD) were cited and the Pearson correlation coefficient (R) was used. For the variables that were not normally distributed the median and interquartile range (IQR) were cited and the Spearman correlation coefficient (Rho) used.

RESULTS

Qualitative review of tumour outlines of the test case

The volumes of the n=25 investigators were assessed. The gsGTV extended over twelve 3mm slices. Qualitative review of contours identified several areas showing poor agreement between iGTV and gsGTV. Under-contouring errors were most common with five centers significantly under-contouring on ≥10 slices. In the central section of the tumor, 23 investigators missed a peritumoral lymph node on two slices (Figure 1A). The peritumoral lymph node was included in the PTV in all cases. The other main area of significant under-contouring was at the inferior end of the tumor where 19 centres failed to contour a small area around the superior mesenteric artery (Figure 1B). Eight investigators systematically over-contoured the duodenum on multiple slices, with five centres significantly over-contouring on >10 slices. Over-contouring of the duodenum was the most common error (Figure 1C). In two cases, investigators had to resubmit as the contours were felt to be unacceptable – one investigator had outlined the bile

duct above the tumor as iGTV and the other had outlined almost the whole of the second part of the duodenum within the iGTV.

Geometric Analysis of the whole volume

The iGTV volumes, with the corresponding GMI and JCI data, for all investigators are shown in Figure 2a. The volume of gsGTV and gsPTV were 26.3 cm^{3}and 223.3 cm^3, respectively. In comparison, the median iGTVs and iPTVs were 23.2cm^3 (IQR: 17.2-32.5) and 209.5 cm^3 (IQR: 183.1-234), respectively. The ratio of the largest to the smallest iGTV and iPTV was

6.8 and 2.8, respectively. The median whole volume JCI of the iGTVs compared to gsGTV was 0.6 (IQR: 0.5-0.7) and the median GMI was 0.3 (IQR: 0.15-0.4). For iPTVs, the median JCI was 0.75 (IQR: 0.7-0.8) and the median GMI was 0.1 (IQR: 0.1-0.2). Additionally, the GTV was highly linearly correlated with GMI (Rho = -0.9), as less GTV is missed as volume increases. JCI and GTV volume (Figure 2a) have a more complex, bell-shaped relationship with lower JCI scores with higher or lower iGTV volumes, as JCI falls with both over and under-outlining errors. A similar relationship is shown for PTV volume and GMI (Rho = -0.9) and JCI (Figure 2b). The slice by slice analysis of conformity between each iGTV and the gsGTV performed in CERR is described in Supplementary Results and shown in Supplementary Figure 2.

Review of **RT planning of the test case**

The protocol advised complete sparing of one kidney, where possible, which was achieved by 13 centres that restricted the contralateral kidney Dmax to less than 2%. Additionally, we imported the planning values into VODCA and compared them to the PAF (Figure 3A-F). In total, values from 22 investigators were available. Two centres recorded the maximum dose to spinal cord as the percentage of the prescription dose as opposed to the requested absolute dose i.e. 18.6Gy in VODCA and 37.1% on PAF. Unexpectedly, four centres misidentified the ipsilateral kidney, erroneously recording the dose to the left kidney on the PAF in patients with pancreatic head tumours. Finally, with regard to the outlining of the OAR, the volumes of outlined structures from the approved cases are shown in Supplementary Table 3. Despite all iGTVs being approved, considerable variation remained.

DISCUSSION

Quality assurance is necessary for safe and effective delivery of radiation. A systemic review identified only nine phase II-III RT studies published between January 1994 and December 2011 that reported clinical outcomes of a RTTQA program of which only two trials had a prospective interventional QA process [4]. Studies reporting impact of RTTQA on clinical outcome showed worse survival in patients who had major protocol deviations [10]. The overall rate of protocol compliance was 65% (2309 evaluable patients) and 13% of the trials had a 'deviation unacceptable' rate >10%. Previous studies incorporated a pre-treatment review of radiotherapy planning – all of these studies showed enhanced protocol adherence [10]. The importance of RTTQA has led to international collaboration of RTTQA processes [11, 12].

Whereas adherence to radiotherapy protocol undeniably improves patient outcome, RTTQA programs have primarily focused on dosimetry and other components of treatment delivery. Over and under-contouring errors are highly likely to impact toxicity and efficacy outcomes, but accuracy of GTV outlining has not been prospectively evaluated in many pancreatic trials [4, 11]. Consistency of GTV outlining is especially important in multi-centre trials of high dose radiotherapy (including stereotactic radiotherapy), where small margins are applied and therefore inappropriate delineation may result in geographic miss or serious gastro-intestinal toxicity.

The SCALOP trial results suggested that outcome with the capecitabine-based regimen was preferable to the gemcitabine-based regimen as it was associated with superior survival and lower toxicity [5]. We believe that the high quality of RTTQA within SCALOP contributed to these results and have laid the foundation for future trials in this disease. To our best of knowledge, SCALOP is the first multicenter pancreatic trial that has comprehensively assessed the variation in investigator-defined contours using qualitative as well as conformity indices in a prospective manner. In the present study, we found poor agreement in iGTV delineation amongst 25 investigators and the corresponding largest: smallest ratio was 6.8. Importantly, this variation occurred despite the availability of a detailed RT outlining protocol, trial-specific clinical atlas and input from radiologists. For the tumour itself, there was greatest concordance in the central sections and the uncertainty lay in defining the superior and inferior borders. A peritumoral lymph node was, however, missed by 23 investigators.

An older study reported variation in investigator-delineated iGTV in locally-advanced pancreatic cancer (LAPC) following assessment of the outlines of 11 radiation oncologists from eight institutions [13]. The ratio of the largest: smallest iGTV contour was 9, with a median iGTV of $31.8cm^3$ for head tumor; a ratio of 3.3 with a median iGTV volume of 67.9 cm^3 for body tumor. However, conformity indices were not used and, in contrast to our trial, a detailed RT planning protocol was lacking [13]. In a recent work, Caravatta et al examined the inter-observer variability in clinical target volume contouring by using the coefficient of variation and the Dice Similarity Index in two dummy runs and revealed acceptable agreement among participating physicians [14].

This work supports the findings of Gwynne *et al*. [9] who studied investigator variation in target delineation in the SCOPE 1 esophageal trial that proposed the concept of slice conformity indices as a tool to identify areas of maximum discrepancy. Our work confirms the utility of slice MDC as a sensitive index that could be used to direct more detailed review.

The completion of the PAF for the pre-trial benchmark identified two areas of misunderstanding in the protocol. Four of 23 centres mis-identified the kidneys as contralateral/ipsilateral. The PAF asked centres to report both the volume of the PTV receiving 95% of the prescription dose (47.9Gy), which a constraint of 99% and the PTV minimum dose, for which there was no clear constraint: 'The aim should be to encompass the PTV with the 95% isodose. Although in routine clinical practice this is generally taken to mean that the absolute minimum PTV dose should always be >90% and ideally >93%, we should have made this a constraint. There was also excellent compliance to the other aspects of the protocol, with no cases exceeding OAR dose constraints. This highlights the preference for PTV underdose rather than OAR overdose. Altogether, the pre-trial educational package and RTTQA program ensured a comprehensive understanding of the radiation protocol, as suggested by the benchmark case and highlight the importance of RTTQA to facilitate high-quality clinical trials.

Given the variability seen with CT-based planning, the role of integrating additional imaging modalities, such as FDG-PET, in pancreatic radiotherapy planning warrants investigation. In one study where CT-based GTVs were compared with co-registered FDG-PET-based GTVs, five of 14 patients required changes based on PET-CT information [15].The average GTVs increased by 29.7%, mainly due to addition of lymph node metastases and extension of primary tumor beyond that defined by CT. We have recently conducted a pilot study comparing GTV delineation using CT versus FDG-PET [16]. This showed a reduction in the standard deviation and range of volumes in favor of PET-CT [16].

Our study had several limitations. First, the use of a single oncologist-radiologist team to outline gsGTV was a potential limitation as it was subject to bias. In hindsight, the use of a consensus outline by a group of national experts, as proposed by Myerson *et al.* [17] may have been more appropriate. Secondly, despite the well-defined planning and RTTQA criteria and the assistance by a radiologist, variability in iGTVs was observed that may reflect a lack of extensive clinical experience in outlining for pancreatic RT. Third, we did not assess whether the variation in outline decreased with time as centres gained more experience with pancreatic radiotherapy contouring. Irrespective, the study does highlight that measures need to be taken to account for inter-observer variation, possibly through the use of interactive on-treatment real-time contour reviews, and intensive contouring workshops involving participating investigators. It should be mentioned that our study is based on only one patient. In summary, our data demonstrated inter-observer variability in GTV contouring despite the availability of a detailed RT protocol, educational atlas and the radiologist's input.

Altogether, our RTTQA, which included the use of PAF, ensured adherence to the radiotherapy planning protocol in the majority of cases, with all identified deviations being minor. Our data provide valuable insight on the importance of well-defined RTTQA and contouring guidelines in the radiotherapy of LAPC. More intense RTTQA, including radiotherapy workshops and real-time central review of delineated GTVs should be considered in future trials.

REFERENCES

[1] Abrams RA, Winter KA, Regine WF, Safran H, Hoffman JP, Lustig R, et al. Failure to adhere to protocol specified radiation therapy guidelines was associated with decreased survival in RTOG 9704--a phase III trial of adjuvant chemotherapy and chemoradiotherapy for patients with resected adenocarcinoma of the pancreas. International journal of radiation oncology, biology, physics. 2012;82:809-16.

[2] Roques TW. Patient selection and radiotherapy volume definition - can we improve the weakest links in the treatment chain? Clinical oncology. 2014;26:353-5.

[3] Peters LJ, O'Sullivan B, Giralt J, Fitzgerald TJ, Trotti A, Bernier J, et al. Critical impact of radiotherapy protocol compliance and quality in the treatment of advanced head and neck cancer: results from TROG 02.02. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2010;28:2996-3001.

[4] Weber DC, Tomsej M, Melidis C, Hurkmans CW. QA makes a clinical trial stronger: evidence-based medicine in radiation therapy. Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology. 2012;105:4-8.

[5] Mukherjee S, Hurt CN, Bridgewater J, Falk S, Cummins S, Wasan H, et al. Gemcitabine-based or capecitabine-based chemoradiotherapy for locally advanced pancreatic cancer (SCALOP): a multicentre, randomised, phase 2 trial. Lancet Oncol. 2013;14:317-26.

[6] Deasy JO, Blanco AI, Clark VH. CERR: a computational environment for radiotherapy research. Medical physics. 2003;30:979-85.

[7] Hanna GG, Hounsell AR, O'Sullivan JM. Geometrical analysis of radiotherapy target volume delineation: a systematic review of reported comparison methods. Clinical oncology. 2010;22:515-25.

[8] Jena R, Kirkby NF, Burton KE, Hoole AC, Tan LT, Burnet NG. A novel algorithm for the morphometric assessment of radiotherapy treatment planning volumes. The British journal of radiology. 2010;83:44-51.

[9] Gwynne S, Spezi E, Wills L, Nixon L, Hurt C, Joseph G, et al. Toward semi-automated assessment of target volume delineation in radiotherapy trials: the SCOPE 1 pretrial test case. International journal of radiation oncology, biology, physics. 2012;84:1037-42.

[10] Willett CG, Moughan J, O'Meara E, Galvin JM, Crane CH, Winter K, et al. Compliance with therapeutic guidelines in Radiation Therapy Oncology Group prospective gastrointestinal clinical trials. Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology. 2012;105:9-13.

[11] Bekelman JE, Deye JA, Vikram B, Bentzen SM, Bruner D, Curran WJ, Jr., et al. Redesigning radiotherapy quality assurance: opportunities to develop an efficient, evidence-based system to support clinical trials--report of the National Cancer Institute Work Group on Radiotherapy Quality Assurance. International journal of radiation oncology, biology, physics. 2012;83:782-90.

[12] Ibbott GS, Followill DS, Molineu HA, Lowenstein JR, Alvarez PE, Roll JE. Challenges in credentialing institutions and participants in advanced technology multi-institutional clinical trials. International journal of radiation oncology, biology, physics. 2008;71:S71-5.

[13] Yamazaki H, Nishiyama K, Tanaka E, Koiwai K, Shikama N, Ito Y, et al. Dummy run for a phase II multi-institute trial of chemoradiotherapy for unresectable pancreatic cancer: inter-observer variance in contour delineation. Anticancer research. 2007;27:2965-71.

[14] Caravatta L, Macchia G, Mattiucci GC, Sainato A, Cernusco NL, Mantello G, et al. Inter-observer variability of clinical target volume delineation in radiotherapy treatment of pancreatic cancer: a multi-institutional contouring experience. Radiation oncology. 2014;9:198.

[15] Topkan E, Yavuz AA, Aydin M, Onal C, Yapar F, Yavuz MN. Comparison of CT and PET-CT based planning of radiation therapy in locally advanced pancreatic carcinoma. Journal of experimental & clinical cancer research : CR. 2008;27:41.

[16] Wilson JM, Mukherjee S, Chu KY, Brunner TB, Partridge M, Hawkins M. Challenges in using (1)(8)F-fluorodeoxyglucose-PET-CT to define a biological radiotherapy boost volume in locally advanced pancreatic cancer. Radiation oncology. 2014;9:146.

[17] Myerson RJ, Garofalo MC, El Naqa I, Abrams RA, Apte A, Bosch WR, et al. Elective clinical target volumes for conformal therapy in anorectal cancer: a radiation therapy oncology group consensus panel contouring atlas. International journal of radiation oncology, biology, physics. 2009;74:824-30.

FIGURE LEGENDS

Figure 1. Examples of the investigating centers gross tumor volumes (iGTVs), pink, and the gold standard gross tumor volumes (gsGTV), yellow, on slices that caused the most common under- and over-contouring errors. A geographical miss of the peritumoral lymph node (red arrow) was observed for 23 of 25 investigators (Figure 1A).). At the inferior end of the tumor, 19 centers failed to contour a small area around the superior mesenteric artery (Figure 1B). Over-contouring of the duodenum was the most common error (Figure 1C).

Figure 2. Scatter plot showing the correlation of the Geographical Miss Index (GMI) and the Jaccard Conformity Index (JCI) with A) the investigator gross tumor volumes (iGTVs) and B) the investigator planning target volumes (iPTVs).

Figure 3. Illustration of the relationship between planning assessment form (PAF) value and the Visualization and Organization of Data for Cancer Analysis program (VODCA) values for benchmark cases, as indicated (A-F). Two minor deviations of dose constraints were identified from the PAF for PTV D95 (<99%) (black data points) (A). VODCA values for maximum dose to spinal cord was significantly lower than the PAF value in 2 cases (black data points).e.g. 18.6 in VODCA and 37.1 on PAF. However, 18.6Gy is 36.9% of the prescription dose and this value would agree with that written on the PAF (B). Incorrect identification of the ipsilateral kidney occurred in four cases (black data points) (C).
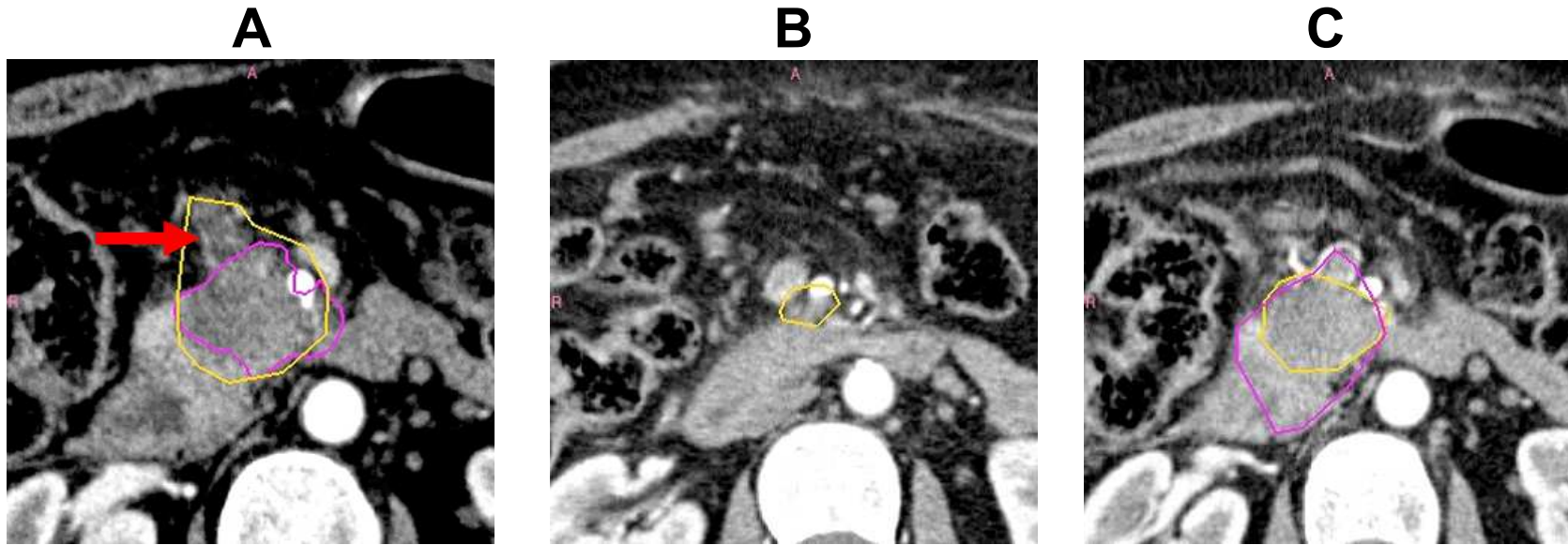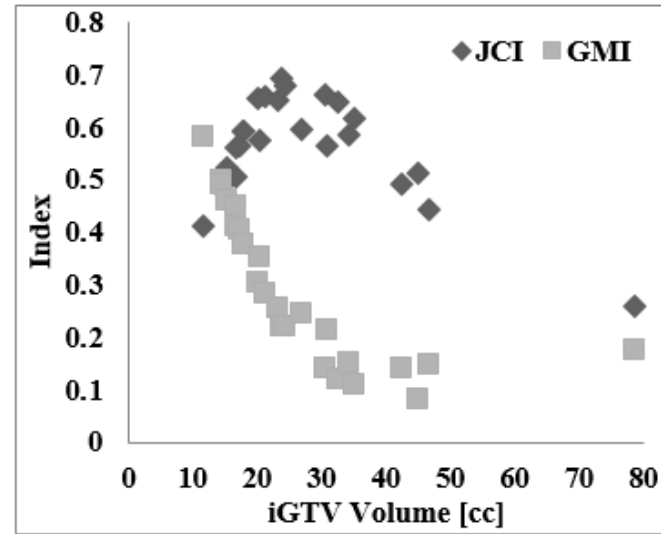
Figure 1

A

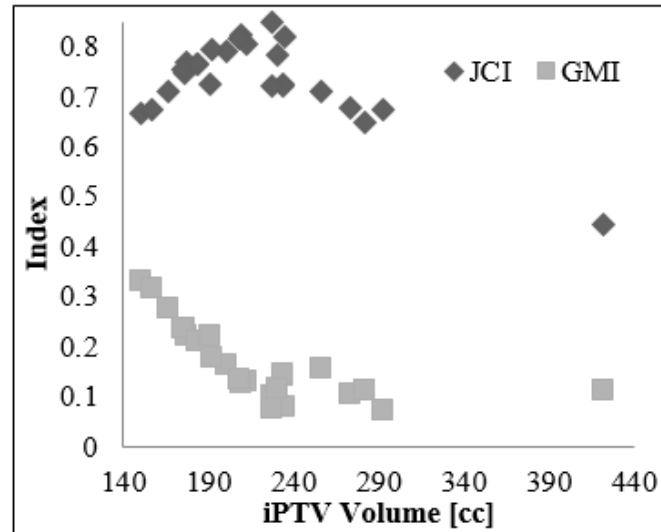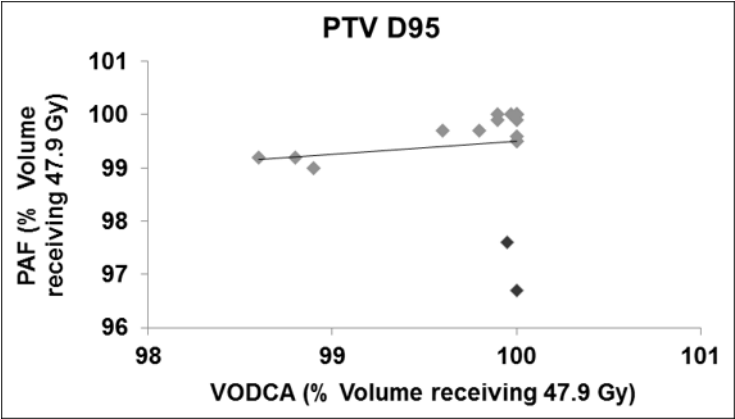**B**
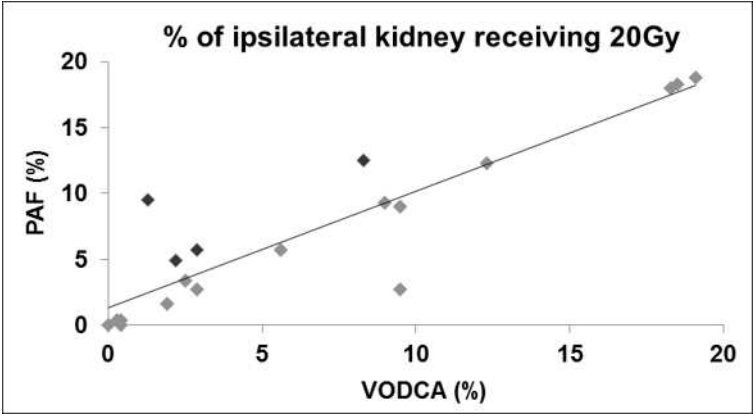
C

Figure 2

A

B

Figure 3
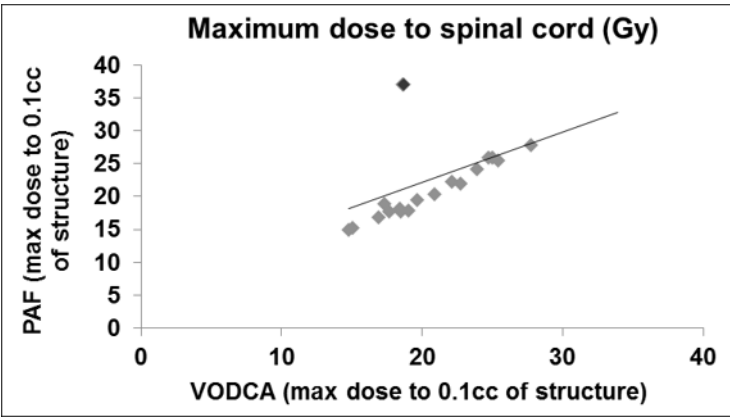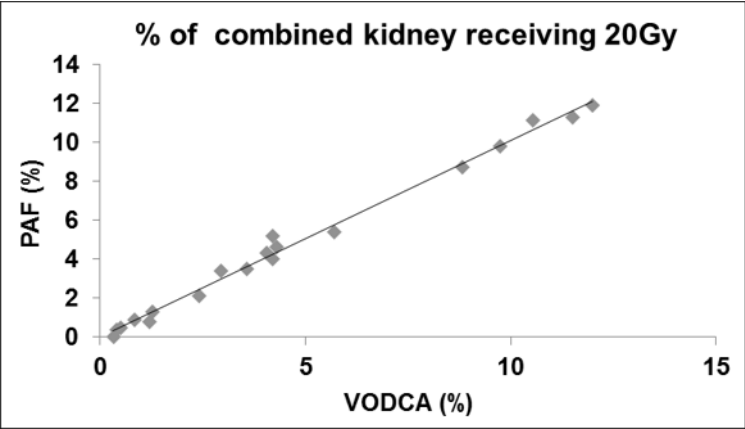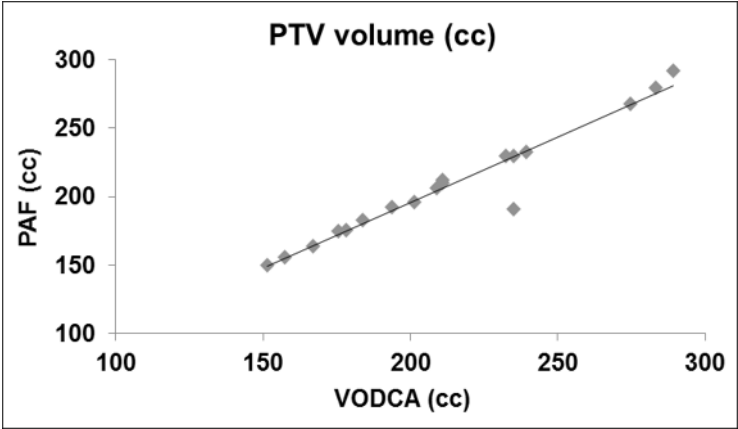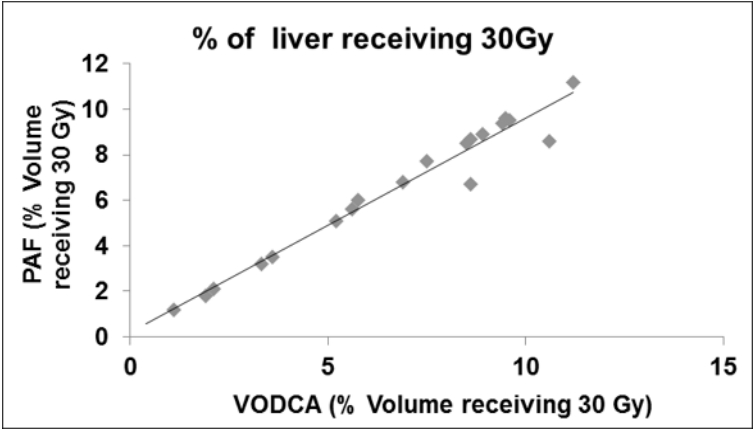
Table 1. Dose to PTV and OARs from benchmark case

Structure/ Dose	Constraint	Gold Standard	PAF values (N=22 cases)		
			Mean (SD)	Minor deviations	Major deviations
PTV D95 (%)	>99.0%	99.9%	99.2% (1.3)	3	none
PTV Dmin (%)	Not defined	94.1%	90.3 % (13)	N/A	N/A
ICRU defined maximum dose (%)	107%	103.4%	103% (1.3)	none	none
Liver V30 (%)	<40%	5.4%	7.3% (4)	none	none
Ipsilateral Kidney V20 (%)	<40%	1.0%	7.7% (7.2)	none	none
Combined Kidney V20 (%)	<30%	1.4%	5.5% (6)	none	none
Spinal cord PRV Dmax (Gy)	<40Gy	0	21.7Gy (8.2)	none	none