Received:
6 August 2012

Revised:
31 December 2012

Accepted:
7 January 2013

doi: 10.1259/bjr.20120398

Cite this article as:

Gwynne S, Spezi E, Sebag-Montefiore D, Mukherjee S, Miles E, Conibear J, et al. Improving radiotherapy quality assurance in clinical trials: assessment of target volume delineation of the pre-accrual benchmark case. *Br J Radiol* 2013;86:20120398.

REVIEW ARTICLE

Improving radiotherapy quality assurance in clinical trials: assessment of target volume delineation of the pre-accrual benchmark case

^{1,2}S GWYNNE, MD, FRCR, ³E SPEZI, BSc, PhD, ⁴D SEBAG-MONTEFIORE, FRCP, FRCR, ⁵S MUKHERJEE, MD, FRCR, ⁶E MILES, DCR(T), MPhil, ⁶J CONIBEAR, MSc, FRCR, and ⁷J STAFFURTH, MD, FRCR, on behalf of the NCRI RTTQA OUTLINING AND IMAGING SUBGROUP

¹Department of Clinical Oncology, Singleton Hospital, Swansea, UK

²Department of Clinical Oncology, Velindre Cancer Centre, Cardiff, UK

³Department of Medical Physics, Velindre Cancer Centre, Cardiff, UK

⁴University of Leeds, St James' Institute of Oncology, Leeds, UK

⁵Gray Institute for Radiation Oncology and Biology, Oxford, UK

⁶Mount Vernon Cancer Centre, London, UK

⁷Division of Cancer, Cardiff University, Cardiff, UK

Address correspondence to: S Gwynne

E-mail: sarah.gwynne@wales.nhs.uk

ABSTRACT

As the complexity of radiotherapy (RT) trials increases, issues surrounding target volume delineation will become more important. Some form of outlining assessment prior to trial entry is increasingly being mandated in UK RT trials. This document produced by the Outlining and Imaging Subgroup (OISG) of the National Cancer Research Institute will address methods to reduce interobserver variation in clinical trials and how to conduct an assessment of outlining through a pre-accrual benchmark case. We review currently available methods of describing the variation and identify areas where further work is needed. The OISG would encourage ongoing discussion with chief investigators in order to provide advice on individual aspects of benchmark case assessment for current and future trials.

The importance of quality assurance (QA) in radiotherapy (RT) trials has been highlighted by recent publications showing that protocol deviations

can negatively impact on survival in head and neck and pancreatic cancer trials [1, 2]. These studies have focused predominantly on planning rather

than target volume delineation (TVD), as they were mostly conducted in the two-dimensional (2D) era. In the three-dimensional (3D) era, issues surrounding TVD will become increasingly important, as this is an inherently observer-biased procedure and the extent of tumour can be interpreted differently by different oncologists [3–5].

The impact of interobserver variation in TVD on trial outcome is yet to be fully explored, but one of the major concerns is that it can represent a systematic error—both for an individual patient and for all the patients treated by the investigator/centre [5, 6]. The typical reported magnitude of interclinician variation commonly exceeds that of geometric systematic error, potentially having tumour control implications for some patients [4] and toxicity implications in others [7]. Variations in the outlining of critical normal structures and outlining errors of patients recruited to the trial are additional important issues, but they are beyond the focus of this paper.

REDUCING INTEROBSERVER VARIATION

There have been a variety of interventions to try and reduce interobserver variation within and outside of clinical trials. These include writing of a clear protocol [8, 9], access to an outlining atlas [10], outlining of pre-trial benchmark cases, review of clinical cases from each centre (“dummy runs”) [3, 11–14] and workshop attendance [14, 15]. Access to a protocol and an outlining atlas has been shown to improve consistency in outlining prostate [16] and rectal cancers, respectively [10, 17]. Educational sessions or workshops have been shown to reduce TVD variation in a range of settings, such as prostate [18] and lung cancer [19].

THE OUTLINING AND IMAGING SUBGROUP

The National Cancer Research Institute Radiotherapy Trials QA (NCRI RTTQA) group has scientific subgroups responsible for the development and implementation of QA in specific technical areas. The Outlining and Imaging Subgroup (OISG) has responsibility for outlining and imaging for target delineation. Although we recognise that there are several sources of error during the TVD stage of multicentre trials, our work to date suggests that the errors of greatest magnitude have been a result of misinterpretation of radiological anatomy [20]. Variations will occur for other reasons, such as different window settings or different planning system algorithms for calculating volumes. In the on-trial setting, further

variations will occur owing to the use of different CT slice number and thickness and differences between image fusion systems. The latter will be important as multi-centre trials requiring image fusion are developed in the UK, and the OISG is developing an appropriate QA programme to address this issue.

AIM

The aim of this document is to inform those involved in RT clinical trials of the issues surrounding the assessment of outlining in a pre-accrual benchmark case. We have focused on relatively complex outlining protocols and emerging metrics that can be used for the analysis. Other essential components of an RTTQA programme, such as pre-accrual assessment of planning and on-trial evaluation of outlining and planning, are beyond the scope of this paper. The trials referred to in this paper are detailed in [Box 1](#).

Box 1. UK trials with estimated start and end dates for recruitment

ARISTOTLE—Advanced Rectal study with Standard Therapy Or a novel agent, Total mesorectal excision (TME) and Long term Evaluation. A Phase III trial comparing standard vs novel chemotherapy as pre-operative treatment for MRI defined locally advanced rectal cancer. ISRCTN09351447 (September 2011 to August 2016)

CHHiP—Conventional or Hypofractionated High-dose Intensity modulated radiotherapy for Prostate cancer. ISRCTN97182923. NCT00392535 (closed on June 2011)

CONVERT—Concurrent ONce-daily VErSus twice-daily RadioTherapy (small cell lung cancer). ISRCTN91927162. NCT00433563 (April 2008 to 2012)

NeoSCOPE—Neo-adjuvant Study of Chemoradiotherapy in Oesophageal Cancer (in set-up). (January 2013 to January 2015)

PIVOTAL—Prostate and pelvis Versus proState Alone treatment for Locally advanced prostate cancer. ISRCTN48709247 (closed on October 2012)

SCOPE 1—Study of Chemoradiotherapy in Oesophageal Cancer Plus or Minus Erbitux. ISRCTN47718479. NCT00509561 (closed on February 2012)

SCALOP—Selective Chemoradiation in Advanced Localised Pancreatic cancer. A multicentre randomised Phase II study of induction chemotherapy followed by gemcitabine- or capecitabine-based chemoradiotherapy for locally advanced non-metastatic pancreatic cancer. ISRCTN96169987 (closed on October 2011)

ROCS—Radiotherapy after Oesophageal Cancer Stenting. ISRCTN12376468 (April 2013 to October 2017)

Four steps in the process of outlining assessment are identified:

- Step 1: RT protocol, atlases and workshops
- Step 2: assessment of protocol adherence—role of the pre-accrual benchmark case
- Step 3: definition of the reference volumes for the benchmark case
- Step 4: assessment of investigator outlines of the benchmark case.

Step 1: RT protocol, atlases and workshops

The RT protocol is the first step towards achieving high-quality TVD within a clinical trial, having been shown to reduce interobserver variation in several cancer sites. It aims to give clear instructions to outliners on how to define the target volumes, including identification of key clinical “subsites” that might require separate outlining assessment—*e.g.* low and intermediate risk of seminal vesicle cases in the CHHiP (Conventional or Hypofractionated High-dose Intensity Modulated Radiotherapy for Prostate Cancer) trial and middle and lower third/junctional oesophageal tumours in the upcoming NeoSCOPE (Neo-adjuvant Study of Chemoradiotherapy in Oesophageal Cancer) trial.

The writing of an RT protocol will be improved by an understanding of the potential errors in TVD for each anatomical subsite, allowing them to be addressed. This is likely to be an ongoing process as it is not possible to anticipate all the errors and account for them within the protocol, especially if this is the first trial in an anatomical subsite. Ambiguities in the protocol can lead to inconsistencies/errors in outlining, and this can be minimised at an early stage of protocol development by a small number of expert individuals, *e.g.* the trial management group (TMG), outlining one or two cases according to the protocol, with any necessary modifications being made before circulation to potential investigators.

An evidence-based approach to improving consistency in TVD is through attendance at an educational workshop. The protocol is presented with opportunity for discussion and outlining of a case. Errors seen in TVD within any previous trials of the same subsite could also be presented. It is possible that further ambiguities in the protocol would be detected at this stage and could be addressed prior to the outlining assessment. This approach has been taken by the ARISTOTLE TMG [21].

An atlas, where images with worked examples of target volumes are defined according to the protocol (Figure 1), or showing areas of uncertainty in both the target volume and the organ at risk outlining may also be helpful. Such an approach was taken in SCOPE (Study of Chemoradiotherapy in Oesophageal Cancer Plus or Minus Erbitux) 1 and CONVERT (Concurrent ONce-daily VErSus twice-daily RadioTherapy) for heart and oesophagus, respectively. This atlas development, like the protocol, can be informed by the ongoing experience with TVD in the anatomical subsite.

Step 2: assessment of protocol adherence—role of the pre-accrual benchmark case

The pre-accrual benchmark case, as a part of a wider pre-accrual QA programme, aims to assess and enhance understanding of and adherence to the RT protocol, with the aim of reducing interobserver and intra-

Figure 1. Images from the Advanced Rectal study with Standard Therapy Or a novel agent, Total mesorectal excision (TME) and Long term Evaluation (ARISTOTLE) trial radiotherapy protocol showing delineation of clinical target volume B (CTVB). CTVB comprises the sites of elective nodal irradiation, including the mesorectum, presacral space and internal iliac lymph nodes. A 7-mm margin (dotted yellow line) is applied to the internal iliac arteries (yellow line) to define the anterior border of the CTVB (blue line) above the mesorectum and a 1-cm margin in front of the mesorectum defines the anterior border of the CTVB at the level of the mesorectum.

observer variation in outlining within the actual trial. Increasingly, an outlining assessment of some kind for potential investigators is being mandated prior to trial QA approval and patient recruitment. Exceptions would be trials with protocols very similar to a previous study where previous QA is accepted or where simple palliative fields are used, *e.g.* Radiotherapy after Oesophageal Cancer Stenting. The assessment of TVD in these situations would be adequately addressed within the “on-trial” setting.

The intensity of assessment will vary according to a number of criteria: treatment intent (radical *vs* palliative), the complexity of treatment [2D *vs* 3D conformal *vs* IMRT (intensity-modulated radiation therapy)], the design of the trial (limited centre Phase II *vs* multicentre Phase III) and the experience of potential investigators with previous similar outlining protocols. A more intensive approach is required if the trial introduces a novel outlining protocol into a multicentre setting for the first time. It will make adequate outlining (reflecting understanding of the protocol and knowledge of the anatomy), a prerequisite for an investigator to be able to enter patients into the study. In addition, variations detected within the initial benchmark cases may identify lack of clarity (or even errors) within the RT protocol, which can be fed back to the TMG and future investigators. This is seen as a vital step in protocol development.

There are various approaches for the selection of the pre-accrual benchmark case. In SCOPE 1 a patient recently treated by the chief investigator was chosen, while in ARISTOTLE two challenging cases were outlined by the protocol development group to inform the protocol and a case of moderate difficulty chosen for the actual pre-accrual benchmark case.

Box 2. Aims of pre-accrual outlining assessment

1. Identify discordance between the investigator outline and the reference volume.
2. Determine the significance of the discordance? Minor enough to “pass” investigator and open centre.
3. Determine if resubmission of outlining assessment needed.
4. Determine cause of discordance.
 - A. misunderstanding of a well-written protocol
 - B. ambiguities in protocol, *i.e.* consistent errors across investigators from different centres
 - C. misinterpretation of radiological anatomy.

The aims of this pre-accrual benchmark case assessment are detailed in [Box 2](#).

Step 3: definition of the reference volumes for the benchmark case/creation of a reference volume

In order to assess the outlining in a pre-accrual benchmark case, investigator outlines are usually compared with a reference volume. Different ways to define a reference volume can be found in the literature. It has sometimes been defined by an individual clinician [10], but it is becoming more common for a group of senior radiation oncologists to agree a consensus manually [14, 22–26]. The Radiation Therapy Oncology Group (RTOG) have created consensus volumes for delineation atlases using the STAPLE (simultaneous truth and performance level estimation) algorithm, which utilises statistical methods to create a single contour from multiple expert contours [27–29], which can be subsequently edited by the constituent outliners (For further details on the STAPLE algorithm, the reader is referred to Warfield et al [30]). Each of these approaches has led to the creation of a single contour. Another approach, taken by European Organisation for Research and Treatment of Cancer trials, is the creation of a reference volume with a maximum and minimum extent, based on the outlines of a group of experts [31].

A consensus approach to defining a reference volume has clear advantages in reflecting the uncertainties and variations that occur even among expert outliners [27, 28]. In the UK, a group made up of experienced clinical oncologists, such as the TMG members, would be the obvious starting point, but the radiologist input is encouraged inline with Royal College of Radiologist guidance [32]. A single consensus contour or a volume with a minimum and maximum extent can be used, but with recognition of the limitations of the reference volume. Data from the SCOPE 1 pre-accrual test case showed a statistically significant effect on the number of investigators classified as achieving excellent conformity when the reference volume was changed from a single clinician/radiologist-defined volume to a TMG consensus volume using the STAPLE algorithm [33] ([Figure 2](#)). It is recommended that at least four individuals’ contours are used to create the STAPLE volume [30].

Figure 2. Four slices of the SCOPE 1 (Study of Chemoradiotherapy in Oesophageal Cancer Plus or Minus Erbitux) pre-trial test case. (A–D) were identified as having <20% of investigator gross tumour volumes meeting the Jaccard conformity index value of 0.7 threshold, indicating that slice-by-slice assessment can detect areas of maximal variation in the volume. Reasons for the poor concordance in these slices included inclusion of bronchus (A) and inclusion of pericardium (D). Figure reproduced with permission from Gwynne et al [20].

Although not the final volume to be treated, gross tumour volume (GTV) definition is the most important step in TVD, as all subsequent volumes are based on this. The methods required to create clinical target volume (CTV) and planning target volume (PTV) from a consensus GTV will depend on the TVD protocol—whether the CTVs are delineated according to anatomy or grown by the treatment planning system (TPS). There may be instances where a CTV grown automatically by the TPS is edited to exclude bone and muscle, such as in PIVOTAL (Prostate and pelvis Versus prOsTate Alone treatment for Locally advanced prostate cancer) and PTV is edited away from the skin surface, such as in head and neck IMRT planning. Clear instructions need to be provided in trial protocols to increase compliance in these aspects of TVD. When assessing PTV conformity, the possibility of variation in TPS margining algorithms [34] needs to be considered. Site-specific solutions will need to be developed for each tumour site but could involve a combination of STAPLE and single investigator-derived volumes.

Step 4: assessment of investigator outlines of the benchmark case/current methods of assessing pre-accrual benchmark case outlining

Visual inspection of outlining, with or without comparison to a reference (gold standard) volume, along with simple measurements, such as length and volume, form the basis of most current pre-trial outlining assessments, while some trials have specified protocol deviations, such as percentage volume over- or under-outlined. Visual inspection of pre-accrual benchmark case outlining can be subjective, labour intensive, error prone and can risk creating a conflict of interest if the chief investigator is performing the assessment. Length, width and volume are very simple to obtain and are entirely objective, but they do not give any information on the spatial relationship between the outlines, with the possibility of two very different outlines having an identical length or a very similar volume. Protocol deviations are often subjective values lacking a robust evidence base, yet applied objectively to outlining assessment. Group analysis and descriptive

Figure 3. Example from the SCOPE 1 (Study of Chemo-radiotherapy in Oesophageal Cancer Plus or Minus Erbitux) trial. (A) shows the original reference gross tumour volume (GTV) (yellow), defined by the chief investigator and an upper GI radiologist and the GTVs delineated by the other 3 clinical oncology members of the SCOPE 1 trial management group. (B) shows the STAPLE derived consensus GTV (red) with the original reference GTV in yellow.

(a)

(b)

statistics of all investigators (mean, median, standard deviation etc.) can be used to identify outliers, but analyses beyond this require software that is able to visualise and

analyse multiple outlines performed on the same CT data set. Before these tools were available, important analyses were performed [12] but were limited in their scope.

Various commercial and non-commercial tools have been developed to process the large amounts of data collected as a part of RTTQA and to view and analyse the outlining data and are currently being used within the OISG. This has allowed complex methods of outlining assessment, such as the use of conformity indices (CIs), to be undertaken. These can represent the variation in volume and spatial relationship in a single metric, which may have distinct advantages. These include the Jaccard conformity index (JCI) [35], geographical miss index (GMI) [36], discordance index (DI) [11], mean distance to conformity (MDC) [37] and the κ statistic [38]. For further details, the reader is referred to a review by Hanna et al [35] and a recent publication by Fotina et al [39].

In the UK, to date, the application of CIs for assessment of pre-accrual benchmark case outlining has been limited to the research setting. Work has been conducted on the SCOPE 1 [20, 40], SCALOP [41] and ARISTOTLE [42] trials. From this work, the advantages and disadvantages of the different CIs can be determined, and these are detailed in Table 1. We have found that a site-specific approach is needed, with different CIs being more suited to different tumour sites.

As discussed in Table 1, these whole-volume metrics are summary statistics, describing the conformity of the outlined volume as a whole. It is possible that variations on a single slice within the volume could be “blurred out” by these. A method of measuring CIs on a slice-by-slice basis may be more informative. At the time of writing, there are tools available to calculate both JCI/GMI/DI (local conformity) [20] and the MDC on a slice-by-slice basis (axial conformity), and these tools have been applied to the SCOPE 1 [40], SCALOP (SM, unpublished work, Northampton General Hospital, 2011–12) and ARISTOTLE accrual benchmark cases [42].

The development of the local and axial conformity tools has made progress towards semi-automated assessment possible. These tools are able to detect areas of discordance within the volume that are potentially missed by the whole-volume statistics, and they can direct visual inspection to these areas, eliminating the need for review of all slices, which will in turn increase

Table 1. Advantages and disadvantages of currently available metrics to describe outlining variation, with values for the 50 SCOPE 1 (Study of Chemoradiotherapy in Oesophageal Cancer Plus or Minus Erbitux) pre-trial test cases and the value for that metric in the context of perfect concordance

Metric	Description of metric	SCOPE 1 GTV values	Value if perfect concordance	Advantage	Disadvantage
JCI/Van't Riet/ DICE	Ratio of the volume of overlap of two structures over union volume of the two structures	0.69 (JCI)	1	Widely used in the literature for multiple tumour sites and with different imaging modalities	Whole-volume metric may miss areas of variation within the volume
		0.65 (Van't Riet)		Benchmark level defined for poor concordance (breast cancer)	Concordance will increase with larger volumes
		0.80 (DICE)			Correlates poorly with length
					Failure to detect small but potentially clinical significant anatomical errors such as the bronchus in the SCOPE 1 pre-trial test case
					No information on the direction of the error
GMI	Calculates the amount of underoutlining	0.09	0	Well correlated with volume	No benchmark for comparison, tumour site- and case-dependent
				Calculates the amount of geographical miss, <i>i.e.</i> underoutlining	
DI	Calculates the amount of overoutlining	Mean 0.25	0	Calculates the amount of overoutlining	No benchmark for comparison, tumour site- and case-dependent
				Well correlated with volume	
Kouwenhoven index	Ratio of the volume of overlap of two structures over union volume of the two or more structures	0.66	1	No reference volume required for calculation	Value dependent on conformity to other investigators and not cf. with gold standard
κ statistic (Fleiss)	Measurement of magnitude of agreement between multiple outlines	0.61	1	No reference volume required for calculation	Value dependent on investigators and not cf. with gold standard
				Objective benchmark values to assess agreement	Only valid for multiple investigator outlines
					Decision required about what level of agreement is acceptable
					No information on the direction of the error

(Continued)

Table 1. (Continued)

Metric	Description of metric	SCOPE 1 GTV values	Value if perfect concordance	Advantage	Disadvantage
κ statistic (Cohen)	Measurement of magnitude of agreement between two outlines	Not analysed	1	Can be used to compare two outlines, e.g. the investigator volume and the reference volume	Not been previously used to assess outlining variation
				Objective benchmark values to assess agreement	Decision required about what level of agreement is acceptable
					No information on the direction of the error
MDC	Shape-based statistic that measures the mean displacement needed to transpose every voxel in the investigator volume onto the reference volume	2.29 mm	0 mm	Gives measurements of variation (in mm)	Overoutlining and underoutlining MDC values that are high in one direction could cancel each other out
				Has an over- and an underoutlining component	Use of the underoutlining and overoutlining MDC results in two metrics, offsetting the advantages of a single metric to describe outlining
				Independent of size of volumes under comparison	No information on the direction of the error
					Correlates poorly with length and volume

Values are median unless otherwise specified.

DI, discordance index; GMI, geographical miss index; GTV, gross tumour volume; JCI, Jaccard conformity index; MDC, mean distance to conformity.

objectivity of the assessment (Figure 3). It will also allow the QA group to identify systematic error issues during the lifetime of a trial that can be used to clarify the protocol and to inform future trial groups.

Semi-automated assessment relies on pre-defined criteria to determine acceptable outlining. Petersen et al, in breast cancer, determined that a JCI value of <0.5 represented poor concordance [24], while Jena et al [37] suggested that a JCI value of 0.8 or above would be an acceptable level of conformity [37], based on their work on glioblastoma TVD with MR/CT fusion [43]. For SCOPE 1, no investigator GTVs met this 0.8 threshold, despite being considered clinically acceptable by the QA team, suggesting that this threshold is too

high for this tumour subsite. Conversely, this threshold was met by all ARISTOTLE investigator CTVAs (GTV+1 cm) Data for whole-volume analysis of pre-trial test case assessment in SCOPE 1, SCALOP and ARISTOTLE is shown in Table 2. Threshold values for JCI, MDC and other CIs are critical and are likely to be cancer subsite dependent, if not also case dependent, and will only reliably increase the objectivity of the assessment if appropriate thresholds are set. They are also critically dependent on the reference volume by which the investigator GTV is compared. In SCOPE 1, the number of investigator GTVs achieving a JCI value of ≥ 0.7 increased from 28% against the original reference GTV to 81% against the new consensus STAPLE-derived GTV ($p \leq 0.0001$) (Figure 2).

Table 2. Results of the evaluation metrics for the 3 trials

Trial	SCOPE 1 GTV ^[20]	SCALOP GTV ^[41]	ARISTOTLE CTVA ^{a, [42]}
No of investigator test cases	50	25	10 (analysis of further cases ongoing)
JCI	0.69	0.56 (mean)	0.86
GMI	0.09	0.29 (mean)	0.10
DI	0.26	NA	0.06

Values are median unless otherwise specified.

DI, discordance index; GMI, geographical miss index; GTV, gross tumour volume; JCI, Jaccard conformity index.

^aCTVA = GTV (primary and involved lymph nodes plus 1 cm).

In addition to pre-defined levels of acceptability, a successful semi-automated approach requires data submitted from the centres to be in a usable format. Issues such as different scan unique identifiers, nomenclature and erroneous outlines will lead to erroneous results and will increase work for the RTTQA group. Clear instructions for investigators will be required also in order to ensure that the only variability between outlines is owing to interpretation of the anatomy and not variation in the window setting and contouring line thickness. QA of planning systems and any image fusion will also need to be addressed.

Data from oesophageal cancer have shown that there are differences in investigator performance, as measured by JCI, between three different mid-oesophageal cases, suggesting that there may be a role for more than one test case to assess outlining in the pre-trial QA programme. 12 investigators who achieved a JCI of ≥ 0.7 against a pre-defined reference volume in the SCOPE 1 pre-accrual benchmark case (case one) re-outlined a GTV on this and two further mid-oesophageal cancer cases on a total of two occasions, with a reference volume for each defined in the same way as the benchmark case. Median JCI for the first and second attempts at case one were 0.73 (interquartile range 0.71–0.74) and 0.72 (0.64–0.74), case two 0.67 (0.64–0.73) and 0.70 (0.64–0.72) and case three 0.65 (0.58–0.68) and 0.61 (0.57–0.67), respectively. Only one investigator achieved a JCI value of ≥ 0.7 on all outlining attempts. No investigator achieved a JCI value of < 0.5 on

their first attempt, but two investigators achieved a JCI value of < 0.5 on their second attempt for cases two and three, respectively.

The analysis was conducted on investigators who performed well ($JCI \geq 0.7$), and this level of conformity identifies investigators that are unlikely to perform badly ($JCI < 0.5$) in future cases [44]. As a minimum, it is recommended that an outlining assessment is made for each anatomical subsite within a clinical trial.

With regard to the approval of centres with multiple clinical oncologists, different approaches can be taken. One solution is to approve a single outliner per centre (PIVOTAL) or extending to this delegated outliner overseeing all subinvestigators at the same clinical site (CHHiP and ARISTOTLE). An alternative approach is central approval of all outliners (SCOPE 1)

Practicalities of pre-accrual benchmark case assessment

The requirement for satisfactory completion of the pre-accrual benchmark case can cause delays in setting up the trial in a specific centre. However, we believe that, for validity of trial outcomes, it is critically important to show QA in the outlining.

The assessment of the pre-accrual benchmark case is only the first step in the QA of outlining within a clinical trial. Defining optimal methods to review the first case(s) from each centre prior to treatment and the ongoing “on trial” QA poses a further challenge, and is yet to be fully explored.

CONCLUSION

The RTTQA group is a valuable resource for the delivery of RT within high-quality clinical trials, by reducing variation in all aspects of the RT process. Multiple issues need to be addressed when conducting an outlining assessment within a clinical trial. Although significant advances in the tools available have been made, they remain at a research stage and need to be tested prospectively in each cancer subsite. This will also allow the linking of observer variation with outcome. The OISG would encourage ongoing discussion with chief investigators in order to provide advice on individual aspects of benchmark case assessment for current and future trials.

FUNDING

SM is supported by the Oxford National Institute of Health Research Biomedical Research Centre

ACKNOWLEDGMENT

The authors would like to acknowledge Charlotte Halle, Naomi Cole and Richard Adams for contributing data for this manuscript.

REFERENCES

- Abrams RA, Winter KA, Regine WF, Safran H, Hoffman JP, Lustiq 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. *Int J Radiat Oncol Biol Phys* 2012;82: 809–16. doi: 10.1016/j.ijrobp.2010.11.039 2011.
- 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. *J Clin Oncol* 2010;28: 2996–3001.
- Begnozzi L, Benassi M, Bertanelli M, Bonini A, Cionini L, Cote L, et al. Quality assurances of 3D-CRT: indications and difficulties in their applications. *Crit Rev Oncol Haematol* 2009;70:24–38.
- Hamilton CS, Ebert MA. Volumetric uncertainty in radiotherapy. *Clin Oncol (R Coll Radiol)* 2005;17: 456–64.
- McLaughlin PW, Evans C, Feng M, Narayana V. Radiographic and anatomic basis for prostate contouring errors and methods to improve prostate contouring accuracy. *Int J Radiat Oncol Biol Phys* 2010;76: 369–78.
- van Herk M. Errors and margins in radiotherapy. *Semin Radiat Oncol* 2004;14:52–64.
- Van de Steene J, Linthout N, de May J, Vinh-Hung V, Claassens C, Noppen M, et al. Definition of gross tumour volume in lung cancer: interobserver variability. *Radiother Oncol* 2002;62:37–49.
- Weiss E, Hess CF. The impact of gross tumour volume (GTV) and clinical target volume (CTV) definition on the total accuracy in radiotherapy theoretical aspects and practical experiences. *Strahlenther Onkol* 2002;179:21–30.
- Jansen EP, Nijkamp J, Gubanski M, Lind PA, Verheij M. Interobserver variation of clinical target volume delineation in gastric cancer. *Int J Radiat Oncol Biol Phys* 2010;77: 1166–70.
- Fuller CD, Nijkamp J, Duppen JC, Rasch CR, Thomas CR Jr, Wang SJ, et al. Prospective randomized double-blind pilot study of site-specific consensus atlas implementation for rectal cancer target volume delineation in the cooperative group setting. *Int J Radiat Oncol Biol Phys* 2011;79:481–9.
- Valley JF, Bernier J, Tercier PA, Fogliata-Cozzi A, Rosset A, Garavaglia G, et al. Quality assurance of the EORTC radiotherapy trial 22931 for head and neck carcinomas: the dummy run. *Radiother Oncol* 1998;47:37–44.
- Seddon B, Bidmead M, Wilson J, Khoo V, Dearnaley D. Target volume definition in conformal radiotherapy for prostate cancer: quality assurance in the MRC RT-01 trial. *Radiother Oncol* 2000;56:73–83.
- Poortmans PM, Veneslaar JL, Struikmans H, Hurkmans CW, Davis JB, Huyskens D, et al. The potential impact of treatment variations on the results of radiotherapy of the internal mammary lymph node chain: a quality-assurance report on the dummy run of EORTC Phase III randomised trial 22922/10925 in Stage I–III breast cancer(1). *Int J Radiat Oncol Biol Phys* 2001;49: 1399–408.
- van Sörnsen de Koste JR, Senan S, Underberg RW, Oei SS, Elshove D, Slotman BJ, et al. Use of CD-rom based tool for analysing contouring variations in involved-field radiotherapy for Stage III NSCLC. *Int J Radiat Oncol Biol Phys* 2005;63:334–9.
- Coles CE, Hoole AC, Harden SV, Burnet NG, Twyman N, Taylor RE, et al. Qualitative assessment of inter-clinician variability of target volume delineation for medulloblastoma: a quality assurance for the SIOP PNET 4 trial protocol. *Radiother Oncol* 2003;69:189–94.
- Mitchell DM, Perry L, Smith S, Elliott T, Wylie JP, Cowan RA, et al. Assessing the effect of a contouring protocol on postprostatectomy radiotherapy clinical target volumes and interphysician variation. *Int J Radiat Oncol Biol Phys* 2009;75:990–3.
- Nijkamp J, de Haas-Kock DFM, Beukema JC, Neelis KJ, Wouterensen D, Ceha H, et al. Target volume delineation in radiotherapy for early stage rectal cancer in the Netherlands. *Radiother Oncol* 2012;102:14–21.
- Khoo EL, Schick K, Plank AW, Poulsen M, Wong WW, Middleton M, et al. Prostate contouring variation: can it be fixed? *Int J Radiat Oncol Biol Phys* 2012;82:1923–9.
- Dewas S, Bibault JE, Blanchard P, Vautravers-Dewas C, Pointreau Y, Denis F, et al. Delineation in thoracic oncology: a prospective study of the effect of training on contour variability and dosimetric consequences. *Radiat Oncol* 2011;6:118.
- Gwynne S, Spezi E, Staffurth J, Palaniappan N, Wills L, Hurt C, et al. Inter-observer variation in outlining of pre-trial test case in the SCOPE1 trial: a United Kingdom definitive

- chemoradiotherapy trial for esophageal cancer. *Int J Radiat Oncol Biol Phys* 2012;84:1037–42.
21. Glynne-Jones R, Harrison M, Gollins S, Harte R, Maggs R, Ghuman S, et al. The development of a conformal radiotherapy protocol for the Phase III ARISTOTLE rectal cancer trial. *Radiother Oncol* 2012;103:S307.
 22. Valicenti RK, Sweet JW, Hauck WW, Hudes RS, Lee T, Dicker AP, et al. Variation of clinical target volume definition in three-dimensional conformal radiation therapy for prostate cancer. *Int J Radiat Oncol Biol Phys* 1999;44:931–5.
 23. Senan S, van Sörssen de Koste J, Samson M, Tankink H, Jansen P, Nowak PJ, et al. Evaluation of a target contouring protocol for 3D conformal radiotherapy in non-small cell lung cancer. *Radiother Oncol* 1999;53:247–55.
 24. Petersen RP, Truong PT, Kader HA, Berthelet E, Lee JC, Hiltz ML, et al. Target volume delineation for partial breast radiotherapy planning: clinical characteristics associated with low interobserver concordance. *Int J Radiat Oncol Biol Phys* 2007;69:41–8.
 25. Clark CH, Miles EA, Guerrero Urbano MT. Pre-trial quality assurance processes for an intensity-modulated radiation therapy (IMRT) trial: PARSPORT, a UK multicentre Phase III trial comparing conventional radiotherapy and parotid-sparing IMRT for locally advanced head and neck cancer. *Br J Radiol* 2009;82:585–94.
 26. Szumacher E, Harnett N, Warner S, Kelly V, Danjoux C, Barker R, et al. Effectiveness of educational intervention on the congruence of prostate and rectal contouring as compared with a gold standard in three-dimensional radiotherapy for prostate. *Int J Radiat Oncol Biol Phys* 2010;76:379–85.
 27. Lawton CAF, Michalski J, El-Naqa I, Kuban D, Lee WR, Rosenthal SA, et al. Variation in the definition of clinical target volumes for pelvic nodal conformal radiation therapy for prostate cancer. *Int J Radiat Oncol Biol Phys* 2009;74:377–82.
 28. Michalski JM, Lawton C, El Naqa I, Ritter M, O'Meara E, Seider MJ, et al. Development of RTOG consensus guidelines for the definition of the clinical target volume for post-operative conformal radiation therapy for prostate cancer. *Int J Radiat Oncol Biol Phys* 2010;76:361–8.
 29. Allozi R, Li XA, White J, Apte A, Tai A, Michalski JM, et al. Tools for consensus analysis of experts' contours for radiotherapy structure definitions. *Radiother Oncol* 2010;97:572–8.
 30. Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Trans Med Imaging* 2004;23:903–21.
 31. Musat E, Roelofs E, Bar-Deroma R, Fenton P, Gulyban A, Collette L, et al. Dummy run and conformity indices in the ongoing EORTC low-grade glioma trial 22033-26033: first evaluation of quality of radiotherapy planning. *Radiother Oncol* 2010;95:218–24.
 32. Royal College of Radiologist. Imaging for Oncology. Collaboration between Clinical Radiologists and Clinical Oncologists in Diagnosis, Staging and Radiotherapy Planning. Board of the Faculty of Clinical Oncology, The Royal College of Radiologists. London, UK: Royal College of Radiologists; 2004.
 33. Gwynne S, Spezi E, Hurt C, Falk S, Gollins S, Joseph G, et al. Importance of the reference volume in assessing outlining performance for the purpose of training and revalidation. NCRI Cancer Conference. Liverpool, UK: NCRI, 2012.
 34. Pooler AM, Mayles HM, Naismith OF, Sage JP, Dearnaley DP. Evaluation of margining algorithms in commercial treatment planning systems. *Radiother Oncol* 2008;86:43–7.
 35. Hanna GG, Hounsell AR, O'Sullivan JM. Geometrical analysis of radiotherapy target volume delineation: a systematic review of reported comparison methods. *Clin Oncol* 2010;22:515–25.
 36. Muijs CT, Schreurs LM, Busz DM, Beukema JC, van der Borden AJ, Pruim J, et al. Consequences of additional use of PET information for target volume delineation and radiotherapy dose distribution for esophageal cancer. *Radiother Oncol* 2009;93:447–53.
 37. Jena R, Kirkby NF, Burton KE, Hoole AC, Tan LT, Burnet NG. A novel algorithm for the morphometric assessment of radiotherapy treatment planning volumes. *Br J Radiol* 2010;83:44–51.
 38. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960;20:37–46.
 39. Fotina I, Lütgendorf-Caucig C, Stock M, Pötter R, Gerog D. Critical discussion of evaluation parameters for inter-observer variability in target definition for radiation therapy. *Strahlenther Onkol* 2012;188:160–7.
 40. Gwynne S, Hurt C, Spezi E, Crosby T, Staffurth J, Jena R. Mean distance to conformity as a tool for assessing outlining variation in the UK SCOPE 1 oesophageal chemoradiotherapy trial. *Radiother Oncol* 2012;103:S259.
 41. Mukherjee S, Halle C, Spezi E, Joseph G, Branagan JMR, Hurt C, et al. Comparison of investigator-delineated GTV in chemoradiotherapy (CRT) for locally advanced non-metastatic pancreatic cancer (LANPC): analysis of the pre-trial test case for the scalop trial. ASCO, GI, San Francisco; 2013.
 42. Cole N, Gwynne S, Spezi E, Maggs R, Sebag-Montefiore D, Adams R. Quality assurance of target volume definition in the ARISTOTLE Phase III rectal cancer trial: initial assessment. *Radiother Oncol* 2012;103:S380.
 43. Burton K, Jefferies S, Jena R, Estall V, Burnet N. Inter and intra observer variation in the gross tumour volume (GTV) delineation for glioblastoma (GBM). *Radiother Oncol* 2008;88:S27.
 44. Gwynne S, Spezi E, Joseph G, Hurt C, Staffurth J, Crosby T. Is there a role for more than one test case in pre-trial outlining assessment? NCRI Cancer Conference. Liverpool, UK: NCRI; 2012.