Intra- and Inter-operator variability in manual tumor segmentation: Impact on radionuclide therapy dosimetry

Yuni K Dewaraja (yuni@med.umich.edu)
University of Michigan Michigan Medicine https://orcid.org/0000-0002-3920-6925
Elise C. Covert
University of Michigan Michigan Medicine
Kellen Fitzpatrick
University of Michigan Medicine: University of Michigan Michigan Medicine
Justin K. Mikell
University of Michigan Medicine: University of Michigan Michigan Medicine
Ravi K. Kaza
UT Southwestern Medical School: The University of Texas Southwestern Medical Center Medical School
John D. Millet
University of Michigan Medicine: University of Michigan Michigan Medicine
Daniel Barkmeier
University of Michigan Medicine: University of Michigan Michigan Medicine
Joseph Gemmete
UMHS: University of Michigan Michigan Medicine
Jared Christensen
University of Michigan ULAM: University of Michigan Medical School
Matthew J. Schipper
University of Michigan Michigan Medicine

Research Article

Keywords: Uncertainty analysis, Dosimetry, Segmentation, Radioembolization, Observer studies, Radionuclide therapy

Posted Date: March 9th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1408164/v1

Abstract

Purpose The aim was to quantify inter- and intra-observer variability in manually delineated lesion contours and the resulting impact on radionuclide therapy dosimetry.

Methods Ten patients with hepatocellular carcinoma lesions treated with ^{90}Y radioembolization (RE) and imaged with post-therapy ^{90}Y PET/CT were selected for retrospective analysis. Three radiologists contoured 20 lesions manually on baseline multiphase contrast-enhanced MRIs and two of the radiologists re-contoured at two additional sessions. Contours were transferred to co-registered PET/CT-based ^{90}Y dose-maps. Volume-dependent recovery-coefficients (RCs) were applied for partial volume correction when reporting mean absorbed dose. To understand how uncertainty varies with tumor size, we fit power models regressing relative uncertainty in volume and in mean absorbed dose on contour volume. Finally, we determined effects of uncertainty on tumor control probability (TCP), as calculated using logistic models developed in a previous report for lesions treated with RE.

Results The average lesion volume ranged from 1.8 mL to 194.5 mL and the mean absorbed dose ranged from 23.4 to 1,629.0 Gy. The mean inter-observer Dice coefficient for lesion contours was significantly less than the mean intra-observer Dice coefficient (0.79 vs. 0.85, $p < 0.001$). Uncertainty in volume, as measured by the Coefficient of Variation (CV) ranged from 4.2% to 34.7% with a mean of 17.2%. For lesions > 8 mL, the CV in mean absorbed dose had an average value of 7.2% (range 1.5% to 12.6%) while for smaller lesions it was 21.7% (range 8.4 to 55.2%). The fitted uncertainty curves as a function of volume, v (in mL), were: %CV (volume) = $23.0 * v^{-0.17}$ and %CV (mean dose) = $32.4 * v^{-0.44}$. With this model for uncertainty, the mean change in TCP was 16.2% (maximum 48.5%).

Conclusion Though we find relatively high inter- and intra-observer reliability overall, uncertainty in tumor contouring propagates into non-negligible uncertainty in dose metrics and outcome prediction for individual cases that should be considered in dosimetry-guided treatment.

Introduction

There is much recent interest in dosimetry-guided personalization of radionuclide therapy with the goal of maximizing tumoricidal effect while limiting impact on normal tissue to an acceptable level [1–3]. Including a measure of uncertainty for dosimetry parameters that may be used in clinical decision-making and in dose-response studies is of much importance, but this is often not reported.

Dosimetry is a multi-step process and uncertainties are inherent in many of the steps, namely, serial quantitative imaging and registration, volume-of-interest (VOI) definition, time-activity curve-fitting and integration, and absorbed dose estimation [4–5]. In the conventional MIRD-based approach, mean absorbed dose to a lesion is estimated by the product of the VOI time-integrated activity and a volume dependent dose-factor derived for a unit-density sphere model [6]. Even when voxel-level dosimetry is performed by coupling patient images with direct Monte Carlo radiation transport for example, the lesion contour is applied to the dose-map to derive mean absorbed dose [7]. Therefore, in both the conventional dosimetry approach and with voxel-dosimetry, uncertainties in segmentation propagate to uncertainties in the lesion absorbed dose. Furthermore, partial volume correction (PVC) using volume-dependent recovery coefficients (RCs) is often part of the activity quantification process for dosimetry that is also affected by uncertainty in segmentation [4–5].

Automated segmentation of select organs on anatomical imaging modalities using deep learning and atlas-based methods is now widely available. However, these automated methods are not yet sufficiently developed for segmentation of most tumor types because lesion size, shape, and location are highly variable and tumor-to-normal-tissue contrast is often poor [8]. Tumor segmentation for dosimetry can be performed on emission images (SPECT or PET) or co-registered anatomical images (CT or MR). Automated count thresholding and gradient-based algorithms are often used for SPECT and PET-based segmentation because of their speed and repeatability; however, due to the noise and limited spatial resolution of emission images, the resulting contours can have poor accuracy [9–10]. Manual tumor segmentation on anatomic images exploits the high resolution of CT and MRI, but inter and intra-observer variability inevitably exists, even when performed by imaging specialists [11–12].

Estimating uncertainty in absorbed dose metrics reported for radionuclide therapy can be challenging. Gear et al. [4] and Finnochiaro et al. [13] investigated the uncertainty in each step of the SPECT-based dosimetry process and identified uncertainty in delineation of the VOI as the major factor. They derived an analytical equation that expresses volume uncertainty as a function of spatial resolution and voxel size for VOIs segmented on SPECT images by thresholding. This analytical approach is not suitable when using manual contouring on CT or MRI because other factors that do not enter into this equation, such as impact of contrast, can dominate. An alternative approach to determining segmentation uncertainties and corresponding dose estimates is to perform a multi-operator study. Such multi-operator studies for manual lesion segmentation are rare in the internal dosimetry setting as it is labor intensive and requires participation by multiple imaging experts. One reported study by Meyers et al. [14] investigated inter- but not intra operator variability. Furthermore, that study did not investigate variability associated with individual lesion contours, as the estimated quantity was the dose to the total tumoral liver.

Our study aims to quantify inter- and intra-observer variability in lesion contours delineated manually on anatomical images and the resulting impact on dosimetry. We also propagate the uncertainty to determine its effects on tumor control probability (TCP). We extended our study to also perform a preliminary investigation of operator variability in measurement of lesion diameters, a parameter widely used to assess response in dose-response studies. We focus on the application of Yttrium-90 (^{90}Y) microsphere radioembolization (RE), a promising treatment for hepatocellular carcinoma (HCC) and metastatic liver malignancies [3, 15]. In RE, intra-arterially delivered microspheres become trapped in the arterioles feeding the tumor and do not redistribute. Dosimetry can therefore be performed with a single imaging timepoint under the assumption that only physical decay contributes to the kinetics, thereby eliminating the uncertainty associated with multi-timepoint emission imaging, registration, and curve fitting.

Methods

Imaging protocol

We used retrospective images from a prior IRB approved study at University of Michigan where ^{90}Y PET/CT imaging was performed after RE with glass microspheres for the purposes of lesion dosimetry [7]. From this larger data set, all patients that had a pre-treatment multiphase contrast-enhanced MRI and HCC lesions that appeared to be greater than approximately 2 mL in volume were selected. Nine patients with a total of 20 lesions were selected based on these criteria. MRI scans were performed on a 1.5T GE Healthcare, 1.5T Philips Healthcare, or a 3T Siemens Healthineers scanner. In-plane resolution ranged from 0.69–1.37 mm and slice thickness ranged from 2–3 mm. Acquisition and reconstruction protocols varied because some scans were obtained from outside hospitals.

Contouring protocol

Three board certified, subspecialty trained abdominal radiologists with 20, 7, and 8 years of experience, respectively were asked to contour the 20 lesions manually on baseline dynamic post-contrast T1-weighted fat-saturated MR images in the phase of contrast enhancement that maximized lesion visualization. Two of the radiologists (A and B) were asked to re-contour the same lesions at two additional rounds separated by one-month intervals to assess intra-observer variability. This gave a total of 140 observations – seven reads per lesion. At each round, the radiologists also recorded the longest lesion diameter in the axial plane according to RECIST criteria [16]. Each radiologist was provided a PowerPoint file with images indicating the general location of the lesions (for example, Fig. 1) and which phase to contour on. Lesion outlines were not specifically identified to minimize bias and PET/CT images were not provided. All contouring and measuring were performed on MIMcloud version 7.1.3 (MIM Software, Cleveland). Radiologists were free to use any of the available tools with the software but were encouraged to contour on the axial slices because of the higher resolution. The radiologists saved their data to a common cloud location and were not able to access data from any previous sessions.

The most experienced radiologist (A) classified lesion boundaries as well-defined or poorly defined. Well-defined lesions were those which showed enhancement on the arterial phase and high contrast to the background liver parenchyma. Lesions were further classified as small or large, with small lesions defined as those with a mean volume of 8 mL or less across all reads.

Dose metrics

^{90}Y dose maps were available from our prior study [7]. Generation of patient-specific dose maps using the in-house developed Dose Planning Method (DPM) Monte Carlo (MC) code is described in detail in that study and is briefly summarized here. The inputs to DPM were the patient's CT-derived density map and the quantitative ^{90}Y PET image acquired on a Siemens Biograph mCT with time-of-flight and reconstructed with 21 iterations, 1 subset of 3D-OSEM with resolution recovery and a 5 mm Gaussian post-filter. The output was the voxel-level dose-rate map, which was converted to a dose map assuming physical decay only.

The baseline MRI was registered to the ^{90}Y PET/CT-derived dose map, and the lesion contours were transferred. Visually, if the automatic rigid registration was deemed unacceptable, manual fine-tuning of the alignment was performed using PET intensity as a guide, which is the process we use in our clinical dosimetry studies. The registration was performed once for each case and saved so that contours of subsequent rounds could be imported without having to re-register the images.

Several dosimetry metrics were recorded for each of the 140 lesion contours transferred to the dose maps – mean absorbed dose, D10, D90, V50, and V100. The D10 and D90 represent the minimum dose, in Gy, delivered to 10% and 90% of the target volume, respectively. The V50 and V100 are the volumes, in mL, which received at least 50 and 100 Gy, respectively. A mean value PVC to correct for resolution effects was only applied to the mean absorbed dose by scaling the value by a volume-dependent RC. The RC versus volume relationship, $RC=-0.934*v^{-0.573}+0.883$, used for this correction came from a previously reported phantom experiment using multiple spheres filled with known activity [7]. Ideally, a voxel-level PVC should be applied to the image to improve accuracy of voxel dosimetry metrics, but such a correction is methodologically challenging, and there is not yet a well-validated practical method of doing so.

Inter- and intra-observer variability

We started by comparing variability among reads conducted by the same radiologist (intra-observer) to variability among reads conducted by two different radiologists (inter-observer). Intuitively, we expected inter-observer variability to exceed intra-observer variability; put simply, contours drawn by the same radiologist should be more similar to each other than contours drawn by different radiologists. We verified this assumption using a two-sample t-test comparing inter- and intra-observer mean Dice coefficient, a measure of spatial overlap. Inter-observer Dice coefficients were calculated for each pair of reads executed by different radiologists on the same tumor during the same round, and intra-observer Dice coefficients were calculated for each pair of reads executed by the same radiologist on the same tumor across rounds.

We performed variance components analysis (VCA) by fitting a two-factor random effects models for repeated measures of volume, mean absorbed dose, and RECIST diameter measurements. Lesion and reader terms were treated as random effects (as opposed to fixed effects) because we take them to be randomly selected from a larger population of interest. That is, the specific raters in this study were not of primary importance; rather, the target of inference was the set of all possible clinicians who may contour HCC lesions. The same is true for the selected lesions themselves. Overall variability in measurements was partitioned into three sources: differences among the lesions themselves, differences between observers (inter-observer error), and differences within observers (intra-observer error). Each source of variability is assumed to have mean zero and an associated variance: σ^2_{lesion}, σ^2_{inter}, and σ^2_{intra}, due to lesion, inter-observer, and intra-observer differences, respectively. These three components sum to the total variance, σ^2_{total}. Models were fit for the whole set of observations, as well as for subsets by size and boundary definition. Subsets jointly defined by size and boundary definition could not be analyzed due to small sample size.

To assess the reliability of the measurements made by the three radiologists, inter- and intra-observer reliability coefficients were calculated [12, 16]. Such reliability coefficients are a form of intraclass correlation coefficients (ICCs) consistent with the random effects model defined above. The inter-observer reliability coefficient describes the consistency of measurements between readers and is expressed as:

$$\rho_{inter} = \frac{\sigma^2_{lesion}}{\sigma^2_{total}}$$

A value closer to 1 indicates that the readers are more interchangeable; that is, more of the variability is attributable to the lesion and not the readers.

The intra-observer reliability coefficient describes the consistency and reproducibility of measurements within a single reader and is expressed as.:

$$\rho_{intra} = \frac{\sigma^2_{lesion} + \sigma^2_{inter}}{\sigma^2_{total}} = 1 - \frac{\sigma^2_{intra}}{\sigma^2_{total}}$$

A value closer to 1 indicates that an increased portion of the total variance is due to differences between lesions and differences between observers; that is, less variance in the outcome is due to random error within one observer.

Uncertainty as a function of tumor volume

To understand how tumor volume impacts uncertainty in volume and dosimetry, we fit models regressing relative uncertainty on contour volume, *v*. Uncertainty was measured by the coefficient of variation (CV), defined by the SD across the seven reads of the same lesion, scaled by the mean of those reads. Initial graphs indicated that a linear model would not be an appropriate fit for the data, and power models of the form $CV = \alpha v^{\beta}$ were found to be a good fit based on examination of residuals.

TCP example

We extended our study to determine the extent to which uncertainty in manual lesion contouring propagates to uncertainty in absorbed dose and probability of tumor control for an example data set. Volume and mean absorbed dose measurements for 89 lesions (from 28 patients with primary and secondary hepatic malignancies) treated with ^{90}Y radioembolization were obtained from a prior study by Dewaraja et al. [7], where tumor control probability (TCP) models were developed for these lesions. Models used a logit link, with ^{90}Y PET/CT based mean absorbed dose as the covariate and binary tumor-level response classification at first follow-up, defined by lesion shrinkage criteria, as the outcome. The prior study did not include any uncertainty estimates. For each of the 89 lesions, we applied mean dose uncertainty (computed from the power model developed in the present study) to determine its effect on TCP using the following steps:

1. Utilizing the volume vs. variability in mean dose function fitted to our original 20 lesions, predict the relative mean dose uncertainty for the given lesion based on its contour volume. Scale by measured mean dose to obtain expected standard deviation (SD). This value represents the SD we would expect to see for this lesion, given inter- and intra-observer uncertainty in volume contouring.
2. Compute measured mean dose ± 2 SD.
3. Plug mean dose ± 2 SD into the previously derived TCP model (logit function).
4. Compute Δ_{TCP} = TCP(mean dose + 2SD) - TCP(mean dose − 2SD). This quantity represents the plausible range of TCP values we would expect to see for this lesion, given the uncertainty in volume contour.

Using Gaussian kernel smoothing, we fit a density estimate to the distribution of Δ_{TCP} and computed area under the density curve to characterize how volume uncertainty translates into uncertainty in tumor control outcomes.

All analyses were performed using R version 4.1.1.

Results

Descriptive statistics and overall uncertainty

Example contours are shown in Fig. 2 and example (longest) diameters according to RECIST criteria are shown in Fig. 3. Individual values of lesion volume, mean absorbed dose and diameters corresponding to each reader and round are plotted in Supplemental Fig. 1–3. Table 1 displays descriptive statistics for lesion measurements and absorbed dose metrics by lesion aggregated across all seven repeated contours. The lesions selected for this study covered a wide range in volume, with average lesion volume ranging from 1.8 mL to 194.5 mL (interquartile range 3.7 to 33.6 mL). Nine lesions were classified as large and 11 small; the most experienced participating radiologist deemed 12 lesions to have well-defined margins and 8 to have poorly-defined margins. While subsets jointly defined by size and boundary definition could not be analyzed due to small sample size, we kept the joint distribution in mind while interpreting the variance components analysis. Mean absorbed dose with RCs applied ranged from 23.5 to 1,629.0 Gy (interquartile range 142.3 to 441.7 Gy).

Table 1
Descriptive statistics by lesion, summarized by mean (SD across the 7 reads).

Lesion Code	Well-Defined?	Size Classification	Volume (mL)	Mean Absorbed Dose (Gy)*	RECIST Diameter (mm)	D10 (Gy)	D90 (Gy)	V50 (mL)	V100 (mL)
Patient 19									
19.2	Yes	Large	54.6 (2.9)	167.1 (5.2)	51.8 (1.3)	196.7 (1.9)	64.6 (5.5)	52.5 (2.4)	38.4 (1.8)
19.3	No	Small	6.1 (1.1)	183.0 (22.5)	25.5 (2.5)	161.2 (5.5)	52.3 (3.7)	5.6 (0.8)	2.6 (0.2)
19.4	No	Small	5.1 (1.5)	98.6 (20.0)	23.0 (2.9)	61.6 (2.7)	35.1 (1.9)	2.0 (0.6)	0.02 (0.05)
19.5	Yes	Small	3.0 (0.6)	144.5 (34.7)	20.9 (2.8)	70.3 (1.8)	34.5 (4.0)	1.6 (0.1)	0.0 (0.0)
Patient 39									
39.1	Yes	Small	1.8 (0.3)	819.4 (280.8)	18.3 (1.3)	205.8 (1.7)	109.8 (5.2)	1.8 (0.3)	1.7 (0.2)
39.3	Yes	Small	2.5 (0.2)	193.0 (23.9)	18.5 (0.6)	110.4 (2.0)	22.5 (2.6)	1.4 (0.1)	0.4 (0.01)
Patient 46									
46.1	Yes	Large	57.8 (2.4)	246.0 (3.7)	54.6 (1.8)	351.2 (1.3)	54.9 (3.7)	52.9 (1.7)	43.6 (1.3)
46.2	Yes	Small	6.2 (0.8)	135.5 (12.1)	29.0 (1.6)	104.9 (1.7)	44.2 (3.1)	5.1 (0.5)	0.9 (0.004)
Patient 49									
49.1	No	Large	194.5 (22.2)	168.1 (12.2)	103.4 (5.1)	284.0 (7.8)	26.5 (10.4)	156.9 (6.9)	114.1 (3.1)
49.2	Yes	Small	3.9 (0.9)	383.4 (64.6)	22.1 (1.2)	238.0 (6.8)	96.5 (10.1)	3.9 (0.8)	3.5 (0.7)
Patient 55									
55.1	No	Large	26.6 (5.1)	87.0 (11.0)	45.7 (6.2)	145.0 (5.3)	4.0 (3.3)	14.4 (4.2)	6.8 (1.2)
55.2	Yes	Large	24.6 (3.3)	68.0 (2.6)	41.3 (1.4)	75.4 (1.3)	23.2 (1.5)	11.7 (1.4)	0.6 (0.0)
55.3	Yes	Large	10.4 (0.9)	23.5 (2.5)	37.0 (2.2)	34.2 (1.5)	2.3 (0.6)	0.03 (0.02)	0.0 (0.0)

*With PVC

Lesion Code	Well-Defined?	Size Classification	Volume (mL)	Mean Absorbed Dose (Gy)*	RECIST Diameter (mm)	D10 (Gy)	D90 (Gy)	V50 (mL)	V100 (mL)
55.4	No	Small	2.1 (0.7)	1,629.0 (899.0)	17.3 (1.7)	394.4 (9.7)	264.1 (23.3)	2.1 (0.7)	2.1 (0.7)
Patient 61									
61.1	No	Large	9.8 (1.9)	263.9 (26.6)	32.9 (2.0)	245.2 (4.2)	83.0 (8.8)	9.5 (1.9)	8.2 (1.5)
Patient 62									
62.1	Yes	Large	58.5 (11.5)	538.5 (44.6)	57.0 (9.8)	727.2 (25.5)	153.4 (18.4)	58.4 (11.4)	56.8 (10.5)
Patient 69									
69.1	No	Small	6.8 (1.9)	566.8 (115.4)	28.0 (1.2)	616.3 (38.9)	93.1 (12.5)	6.7 (1.9)	5.9 (1.5)
69.2	No	Small	3.0 (0.7)	1,284.6 (324.3)	20.7 (1.9)	699.9 (24.5)	260.2 (34.1)	3.0 (0.7)	3.0 (0.7)
69.3	Yes	Small	7.4 (1.0)	383.4 (32.3)	24.8 (2.2)	325.4 (7.2)	138.3 (10.9)	7.4 (1.0)	7.2 (1.0)
Patient 74									
74.1	Yes	Large	125.4 (20.2)	409.4 (29.9)	86.5 (7.2)	578.8 (17.3)	105.1 (19.9)	121.2 (18.6)	112.8 (16.0)

*With PVC

There was a considerable range of uncertainty (CV% across seven measurements) in volume, ranging from 4.2–34.7% with a mean of 17.2%. Uncertainty in RECIST diameter measurement within a given lesion was lower than that of contour volume. The uncertainty in diameter measurements ranged from 2.6–17.2%, with a mean of 7.7%. Uncertainty in RC-corrected mean absorbed dose varied from 1.5–55.2% with a mean of 15.1%. For large (> 8 mL) lesions, the uncertainly in mean absorbed dose ranged from 1.5–12.6% with a mean value of 7.2%. For small lesions, uncertainty in mean absorbed dose ranged from 8.4–55.2% with a mean value of 21.7%. Regarding dose-volume histogram metrics, uncertainty in D10 ranged from 0.4–6.3% with a mean of 2.7%. Variability in D90 was much higher, with CV% ranging from 4.7–83.1%, with a mean of 15.7%.

Inter- and intra-observer variability

Dice coefficient values for the contours had a left skewed distribution and ranged from 0.48 to 0.95 (mean = 0.82, median = 0.84). Mean inter-observer Dice coefficient was 0.79 (SD = 0.09), while mean intra-observer Dice coefficient was 0.85 (SD = 0.07). Histograms of Dice coefficients are plotted in supplemental Fig. S4. A two-sample t-test confirmed that the mean Dice coefficient was significantly higher for pairs of contours traced by the same reader than for pairs of contours traced by different readers ($T = -4.80$, $P < 0.001$).

Figure 4 illustrates the breakdown of total variance in volume, mean absorbed dose, and RECIST into lesion, inter-observer, and intra-observer components. Output in table form, including specific percent values, is available in the supplementary materials. Corresponding intra- and inter-observer reliability ICCs are presented in Table 2.

Table 2
inter- and intra-observer reliability ICCs for all outcomes.

	All Lesions (n = 140)	Small Lesions (n = 77)	Large Lesions (n = 63)	Well-Defined Lesions (n = 84)	Poorly Defined Lesions (n = 56)
Volume					
ρ_{intra}	0.992	0.878	0.989	0.983	0.997
ρ_{inter}	0.967	0.735	0.954	0.956	0.974
Mean Dose					
ρ_{intra}	0.818	0.788	0.993	0.957	0.790
ρ_{inter}	0.754	0.714	0.980	0.853	0.732
RECIST Diameter					
ρ_{intra}	0.991	0.898	0.985	0.994	0.989
ρ_{inter}	0.966	0.743	0.939	0.947	0.984

ρ_{intra}: intra-observer reliability coefficient; ρ_{inter} = inter-observer reliability coefficient.

Uncertainty as a function of tumor volume

For each lesion, uncertainty in volume and mean absorbed dose are plotted versus lesion volume in Fig. 5, with the functional form of the fitted curves displayed.

TCP example

Figure 6 illustrates the propagation of volume and mean dose uncertainty into estimation of probability of tumor control. Panel A presents the procedure for finding Δ_{TCP} for an example patient, represented by the red bar along the y-axis. Repeating the same procedure, we overlay the Δ_{TCP} values for all 89 lesions in Panel B. Recall that the length of the bars is determined by volume; thus, even two lesions with similar mean dose can have error bars of noticeably different sizes. Panel C displays empirical and density smoothed histograms of the distribution of Δ_{TCP}. Among the lesions included (empirical histogram), mean Δ_{TCP} was 16.2% and maximum Δ_{TCP} was 48.5%. Based on the density smoothed curve, 27.0% of lesions have TCP differences of at least 25% when accounting for standard uncertainty.

Discussion

Uncertainty in delineating the VOI is a primary source of error along the radionuclide therapy dosimetry chain [4, 13]. The degree of contrast enhancement, spatial resolution, and tumor volume are main factors that restrict the precision with which the observer can assess the lesion boundary on anatomical imaging modalities. Ideally, we would create an average VOI boundary across multiple observers, but this is an impractical use of resources in clinical practice. Our study simulated that ideal situation by having three radiologists repeatedly outline the same tumors on a historical data set. Leveraging these repeated measurements, we have quantified how observer Effects contribute to uncertainty in VOI delineation and the corresponding absorbed dose estimates and provided a model that can be used in future studies to estimate uncertainty. To our knowledge, this is the first study to assess the impact of manual lesion contouring variability on dosimetry results via an inter- and intra-operator study.

Dice coefficients revealed that the mean intra-observer spatial overlap (0.85) is significantly greater than the mean inter-observer overlap (0.79) in contours, which substantiates the assumption that operators tend to agree with themselves more than they agree with other operators. Based on variance components analysis, the vast majority of overall variance in volume, mean absorbed dose, and RECIST diameter is attributable to inherent differences between the lesions such as differences in volume and enhancement (represented by the darkest blue bars in Fig. 4). This observation makes sense in the context of the study because the lesions in our sample varied greatly in size. We anticipated that inter-observer effects would account for a greater proportion of variability than intra-observer effects, which is true in most cases. However, this hypothesis did not hold for mean absorbed dose overall, in small lesions, or in poorly defined lesions. Sensitivity analyses revealed this result to be largely attributable to one very small, poorly defined lesion (Fig. 2(d), lesion code 55.4). Radiologist A defined Lesion 55.4 to have a volume of 1.8mL on one read and 3.5mL on another; this nearly two-fold difference was greatly magnified by the application of RC correction at small volumes, thereby creating large variation within Radiologist A's mean dose measurements. When excluding Lesion 55.4, inter-observer variability is larger than intra-observer variability across all outcomes and subgroups, as expected. We also anticipated that variability attributable to the reader would be greater in small versus large lesions and in poorly-defined versus well-defined lesions. Indeed, observer effects (the two bars in lighter shades of blue in Fig. 4) account for a greater proportion of the total variance in small lesions than in large lesions. Surprisingly, however, observer effects are not greater in poorly defined lesions compared to well-defined lesions. One possible explanation is that the subgroup analyses by boundary definition did not account for lesion volume, which we know to be an important factor in variability.

The inter- and intra-observer reliability coefficients presented in Table 2 suggest substantial agreement both between and within readers. Most intra-observer reliabilities are greater than 0.9, reinforcing the conclusion that observations of the same case made by the same reader are generally consistent and reproducible. Encouragingly, inter-observer reliabilities are nearly all above 0.8, reflecting substantial agreement *between* readers, as well. These findings are consistent with the findings of Meyers et al. [14], who report an inter-observer ICC of 0.94 for volume and 0.73 for mean absorbed dose for delineation on contrast enhanced CT in a similar cohort of HCC patients treated with ^{90}Y RE. Similarly, McErlean et al. [11] determined intra- and inter-reliabilities of 0.957 and 0.954, respectively, for RECIST measurements on CT images, which compare well with our values.

We provide fitted uncertainty curves (Fig. 5) that can potentially be applied to future patient studies to produce an informed estimate of standard uncertainty without implementing the entire error propagation schema. However, it is important to caveat that these values depend heavily on the imaging modality/parameters and contouring method used. We expect that the findings will be also applicable to hepatic lesion types other than HCC, because lesion contouring was done on the contrast enhanced sequences of MRI, which is routinely used for evaluation of any primary or secondary hepatic malignancies. Figure 5 demonstrates that in general, uncertainty is reduced when

progressing from volume to mean dose calculation. This is to be expected because the dose maps are blurred out by motion and the limited spatial resolution of ^{90}Y PET. In Fig. 5, the sharp rise in the mean dose uncertainty at small volumes is partly due to the sharp rise in the RC curve at small volumes.

The relationship between volume and dose uncertainty ascertained by our empirical approach can be compared with results presented by Finnochiaro et al. [13], who used an analytical equation that captures uncertainty in volume as a function of image resolution. Although the pattern is the same, our estimates of uncertainty are much lower. For example, at a volume of 100 mL, Fig. 5 estimates just over 10% uncertainty in volume and about 5% uncertainty in mean dose. In contrast, Finnochiaro et al. estimate over 30% uncertainty in volume and over 25% uncertainty in mean dose. This difference can be mostly attributed to the fact that we used MRI for tumor segmentation, which is much higher resolution than the SPECT imaging used in the comparison study. Furthermore, although uncertainty in segmentation was the dominant factor, they included other sources of uncertainty from the dosimetry chain, which were beyond the scope of our study.

We applied our model of segmentation uncertainty from the present study to determine how it impacts a model of probability of tumor control published previously by our group. We found the largest impact on TCP among lesions with intermediate mean dose values, which is attributable to the shape of the logistic curve (Fig. 6). Overall, our analysis predicts that approximately one in four lesions would have Δ_{TCP} of at least 25% when accounting for standard uncertainty. Although TCP is not presently formally utilized as a clinical decision-making aide, it reflects expected treatment efficacy and patient outcomes. A clinician might make different treatment decisions given 50% probability of tumor control compared to 75% probability. Similarly, our characterization of the segmentation uncertainty on absorbed dose reporting could be used to design processes to reduce its contribution to treatment failures. One solution is to devise a more reproducible and repeatable segmentation method. Another is to plan RE infusions with enough additional dose (“dosimetric margin”) to lesions such that the segmentation uncertainty has little effect on outcome; for example, forcing tumor absorbed doses deeper into the plateau region of a dose-response curve. Normal liver parenchyma must be considered as well in such a plan. Nevertheless, demonstrated benefit of personalized dosimetry in a recent trial [3] suggests that such escalation is feasible in select cases.

Although volume uncertainty is expected to be the greatest contributor to dose uncertainty in radionuclide therapies, there are other potential sources of error along the dosimetry chain [4, 5, 13] that we did not investigate, which is a limitation of our study. Some of these, such as those associated with the need for multi-time point imaging of the activity distribution, are not relevant to the ^{90}Y RE application. However, uncertainties due to image mis-registration when transferring the contours defined on baseline MR images to the co-registered PET/CT dose maps are relevant. This could be estimated empirically by intentionally introducing mis-registrations in various directions and calculating the effect on the lesion absorbed dose. Quantifying this effect was beyond the scope of our study aims, but it motivates future work to integrate the analysis conducted here with other sources of variation. Similarly, it is important to note that our analysis of change in TCP only captures uncertainty arising from VOI delineation. Future studies could also benefit from the inclusion of more than three radiologists, which was another limitation of our study. Additionally, in the absence of ground truth, our study was constrained to assessing observer variability and was not equipped to ascertain accuracy. A previous study of accuracy and reproducibility using synthetic brain MR images reported that manual tracers tend to overestimate lesion margins compared to automated techniques [18]. Regardless, reduction of inconsistencies among radiologists reduces variability in dosimetry and has potential to reduce variability in patient outcomes when using dosimetry-guided radionuclide therapy.

Conclusion

Uncertainty in tumor contouring propagates into non-negligible uncertainty in dosimetry metrics and outcome prediction. This is especially true for small (< 8 mL) lesions, where the coefficient of variation in mean absorbed dose was on average 21.7% (range 8.4–55.2%), Understanding how uncertainty propagates from target delineation to dosimetry metrics to treatment decision-making is of critical importance in the quest to personalize therapeutic regimens.

Declarations

Acknowledgments

This work was funded by National Institute of Biomedical Imaging and Bioengineering, Grant/Award Number: R01 EB022075; National Cancer Institute, Grant/Award Number: R01 CA240706. Software support from MIM Software is acknowledged.

Ethical approval.

The research imaging studies involving human participants were approved by the internal review board (IRB) at University of Michigan.

Consent to participate

All patients gave their informed consent to participate to the study.

Conflict of interest

Yuni Dewaraja is a consultant for MIM Software. All the other authors have no conflicts of interest related to the present paper to disclose.

References

1. Konijnenberg M, Herrmann K, Kobe C. *et al.* EANM position paper on article 56 of the Council Directive 2013/59/Euratom (basic safety standards) for nuclear medicine therapy. Eur J Nucl Med Mol Imaging. 2021;48, 67–72.
2. Pandit-Taskar N, Iravani A, Lee D, et al. Dosimetry in Clinical Radiopharmaceutical Therapy of Cancer: Practicality Versus Perfection in Current Practice. J Nucl Med. 2021;62(Suppl 3):60S-72S.
3. Garin E, Tselikas L, Guiu B, el al. Personalized versus standard dosimetry approach of selective internal radiation therapy in patients with locally advanced hepatocellular carcinoma (DOSISPHERE-01): a randomized, multicentre, open-label phase 2 trial. Lancet Gastroenterol Hepatol. 2021;6(1):17-29.2.
4. Gear JI, Cox MG, Gustafsson J, Gleisner KS, Murray I, Glatting G, et al. EANM practical guidance on uncertainty analysis for molecular radiotherapy absorbed dose calculations. Eur J Nucl Med Mol Imaging. 2018;45:2456-

2474.

5. Gustafsson J, Brolin G, Cox M, Ljungberg M, Johansson L, Gleisner KS. Uncertainty propagation for SPECT/CT-based renal dosimetry in (177)Lu peptide receptor radionuclide therapy. Phys Med Biol. 2015;60(21):8329-46.
6. Stabin MG, Sparks RB, Crowe E. OLINDA/EXM: the second-generation personal computer software for internal dose assessment in nuclear medicine. J Nucl Med. 2005 ;46(6):1023-7.
7. Dewaraja YK, Devasia T, Kaza RK, et al.Prediction of tumor control in ^{90}Y Radioembolization by logit models with PET/CT-based dose metrics. J Nucl Med. 2020;61:104-111.
8. Chlebus G, Schenk A, Moltz JH,et al. Automatic liver tumor segmentation in CT with fully convolutional neural networks and object-based postprocessing. *Sci Rep. 2018;* **8:**15497. https://doi.org/10.1038/s41598-018-33860-7
9. Hatt M, Lee JA, Schmidtlein CR, et al. Classification and evaluation strategies of auto-segmentation approaches for PET: Report of AAPM task group No. 211. Medical physics. 2017;44(6):e1-42.
10. Mikell JK, Kaza RK, Roberson PL, et al. Impact of 90 Y PET gradient-based tumor segmentation on voxel-level dosimetry in liver radioembolization. EJNMMI physics. 2018;5(1):1-7.
11. McErlean A, Panicek DM, Zabor EC, et al. Intra- and interobserver variability in CT measurement in oncology. Radiology. 2013;269(2):451-458.
12. 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. Radiation Oncology Biol. Phys. 2007;68(3):763-770
13. Finnochiaro D, Gear JI, Fioroni F, et al. Uncertainty analysis of tumour absorbed dose calculations in molecular radiotherapy. EJNMMI Physics. 2020;7(63):1-16.
14. Meyers N, Jadoul A, Bernard C, et al. Inter-observer variability of 90 Y PET/CT dosimetry in hepatocellular carcinoma after glass microspheres transarterial radioembolization. EJNMMI Physics. 2020;7:1-12.
15. Weber M, Lam M, Chiesa C, et al. EANM procedure guideline for the treatment of liver cancer and liver metastases with intra-arterial radioactive compounds. Eur J Nucl Med Mol Imaging. 2022 Feb 11. doi: 10.1007/s00259-021-05600-z. Epub ahead of print.
16. Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer. 2009;45(2):228-47.
17. Eliasziw M, Young SL, Woodbury MG, Fryday-Field K. Statistical methodology for the concurrent assessment of interrater and intrarater reliability: using goniometric measurements as an example. Physical Therapy. 1994;74(8);777-788.
18. Ashton EA, Takahashi C, Berg MJ, et al. Accuracy and reproducibility of manual and semiautomated quantification of MS lesions by MRI. J. Magn. Reson. Imaging. 2003;17:300-308.

Figures

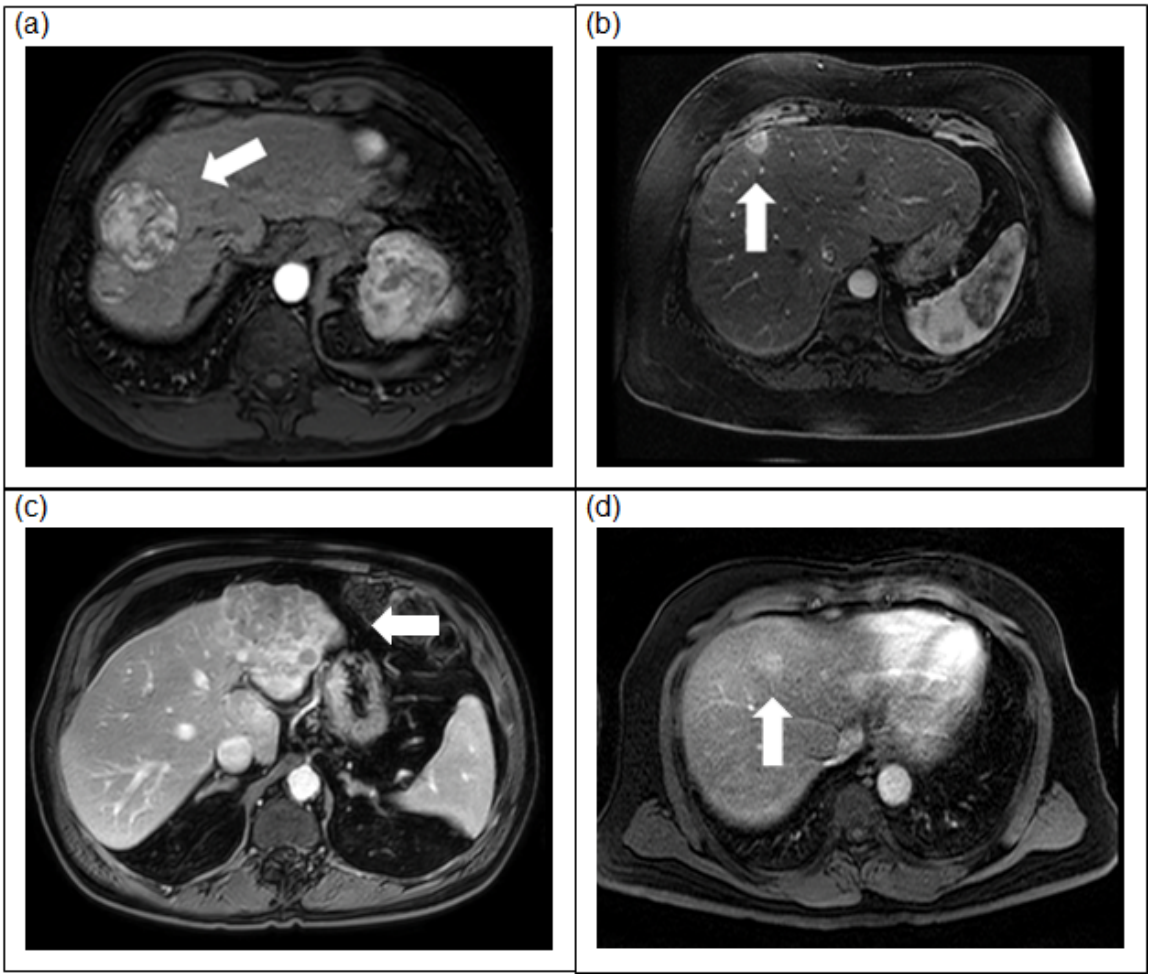

Figure 1

Example HCC lesions as seen on dynamic post-contrast T1-weighted fat-saturated MRI images. (a) Large, well defined (code 46.1, corresponding to patient #46, lesion #1). (b) Small, well defined (39.3). (c) Large, poorly defined (49.1). (d) Small, poorly defined (55.4)

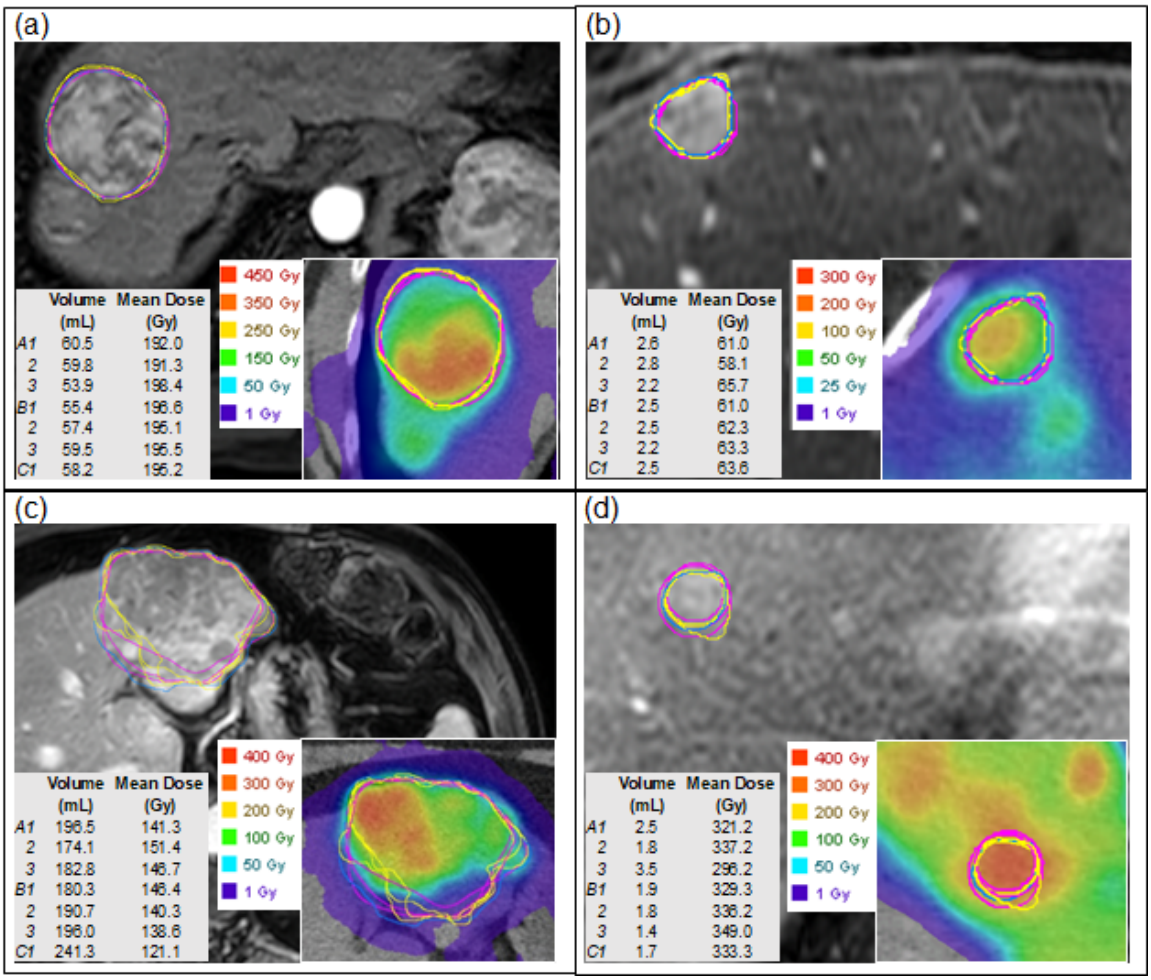

Figure 2

Radiologist-defined tumor contours on MRI corresponding to the four example lesions depicted in Fig. 1. Inserts show contours transferred to co-registered ^{90}Y PET/CT-based dose-maps. Mean dose values are indicated before PVC. (a) Large, well defined (code 46.1). (b) Small, well defined (39.3). (c) Large, poorly defined (49.1). (d) Small, poorly defined (55.1). Contours drawn by the same radiologist are indicated in the same color (Pink = radiologist A, Yellow = radiologist B, Blue = radiologist C).

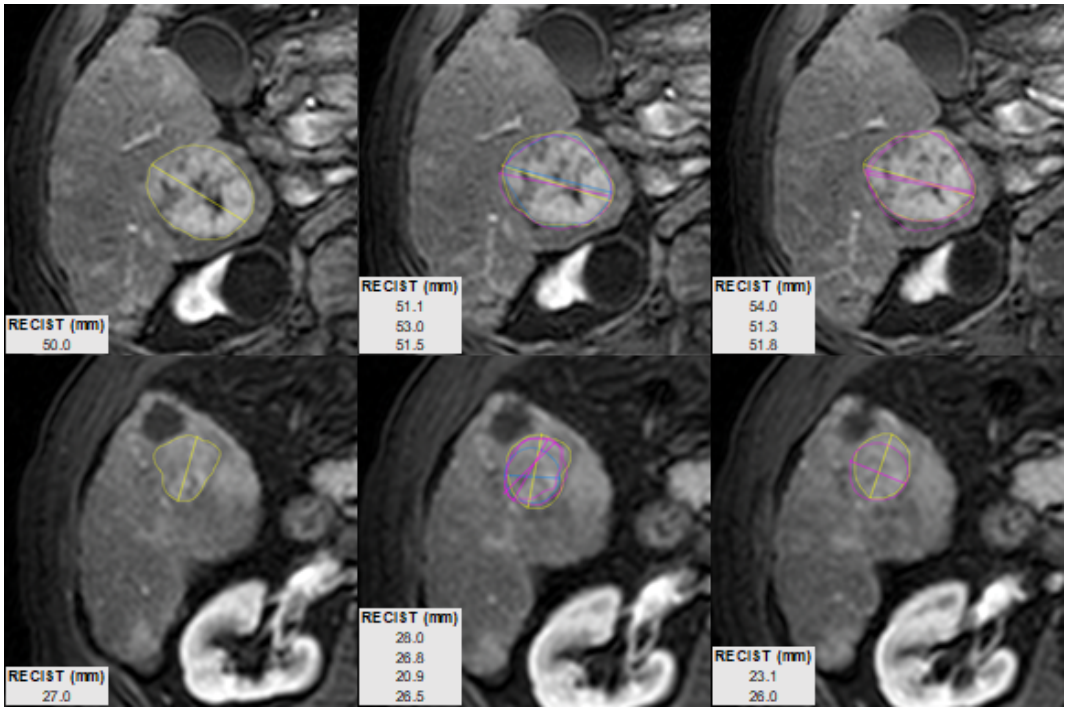

Figure 3

Radiologist-defined RECIST measurements on two example lesions: Lesion code 19.2 (top row) and 19.3 (bottom row). Radiologists were free to choose the MRI slice on which they indicated the diameter. Within each row, each of the three images depicts a different MRI slice. Diameters drawn by the same radiologist are indicated in the same color (Pink = radiologist A, Yellow = radiologist B, Blue = radiologist C). Lesion contours are included for context.

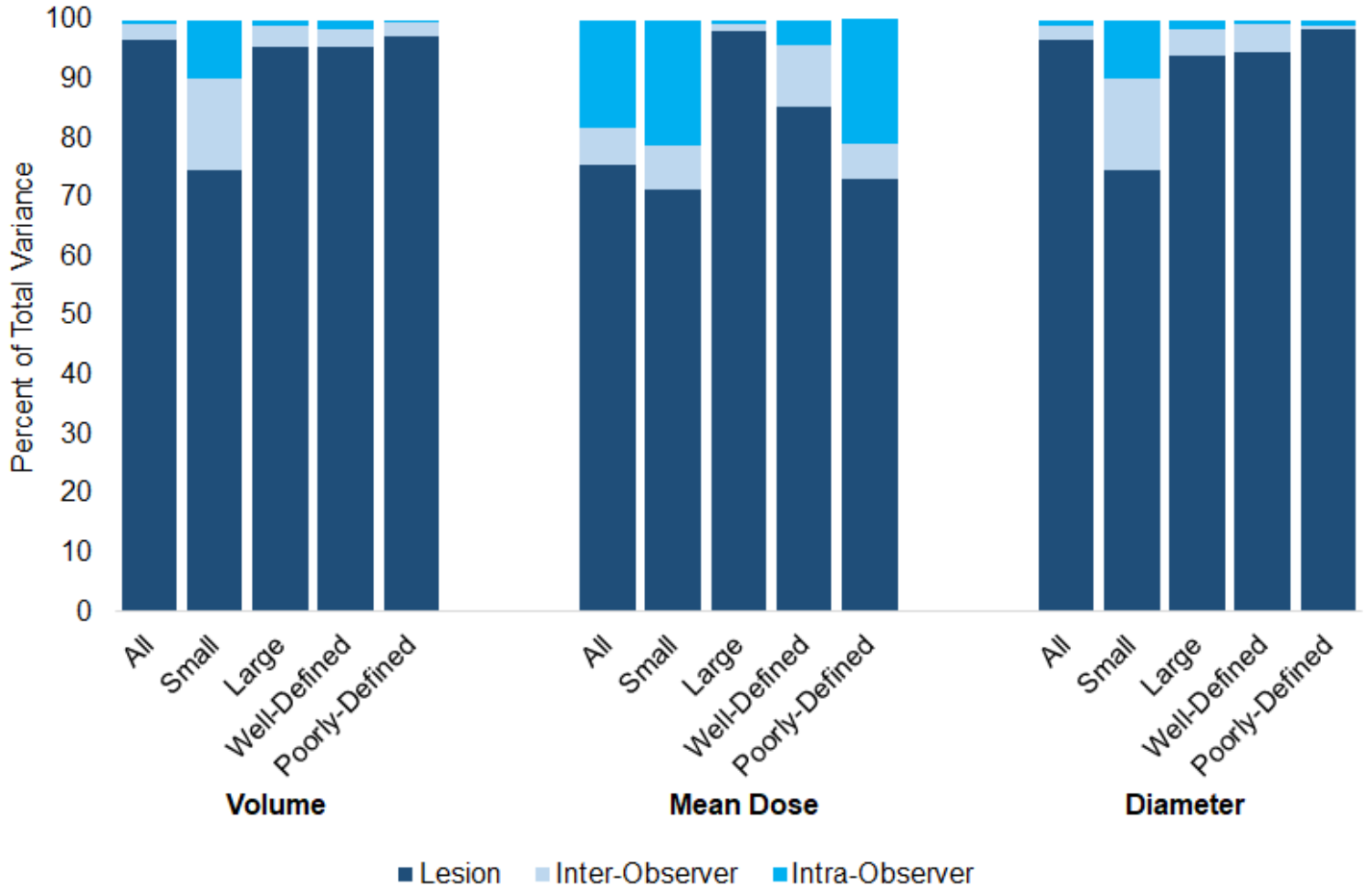

Figure 4

Stacked bar chart illustrating the percentage of total variance attributable to each of three variance components (lesion, inter-observer, intra-observer), overall and within size and boundary definition subgroups. Bars are clustered by outcome of interest (volume, mean absorbed dose, and RECIST diameter).

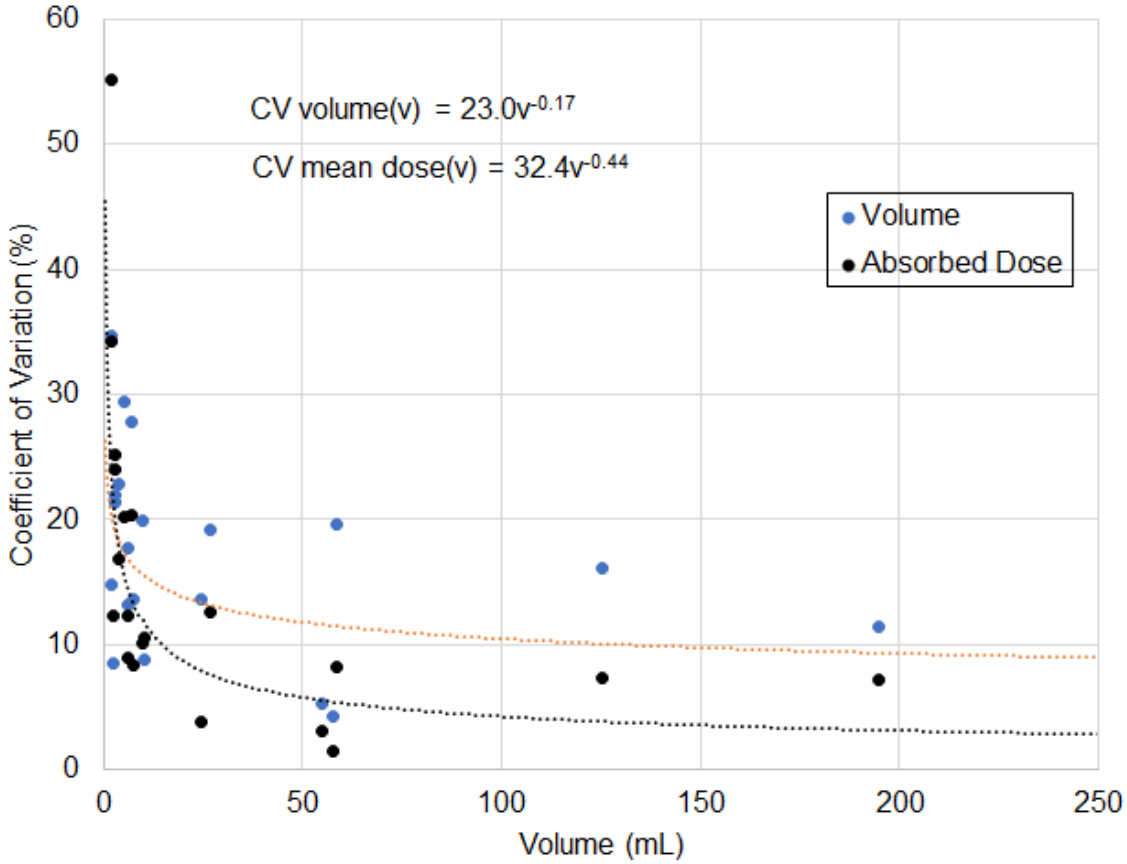

Figure 5

Volume uncertainty (blue points; blue line fitted) and mean absorbed dose uncertainty (black points; black line fitted) as a function of volume.

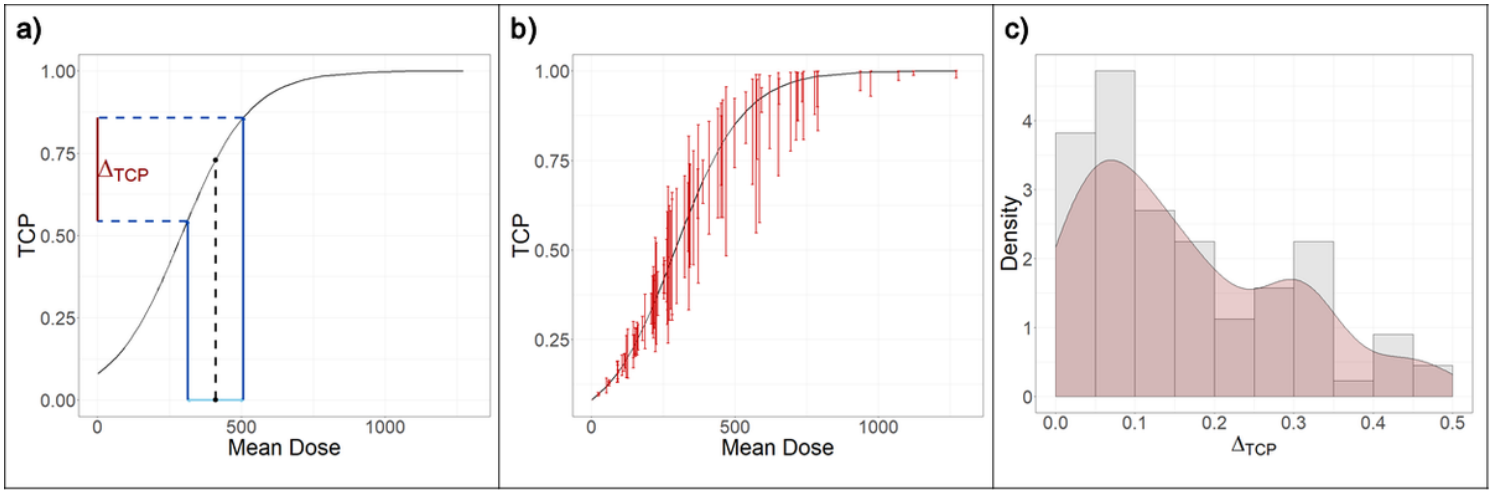

Figure 6

Volume uncertainty propagated into tumor control probability (TCP), applied on the data and TCP curve fitted in the study by Dewaraja et al. [7]. (A) Procedure for computing Δ_{TCP} for an example patient. (B) Original TCP curve (black

line) overlaid by Δ_{TCP} for 89 liver lesions. (C) Empirical and density smoothed histograms for the distribution of Δ_{TCP}.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- covertsupplementaldata.docx