IOP Publishing Phys. Med. Biol. 69 (2024) 025025 https://doi.org/10.1088/1361-6560/ad19ef

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

OPEN ACCESS

RECEIVED
11 September 2023

REVISED
27 November 2023

ACCEPTED FOR PUBLICATION
2 January 2024

PUBLISHED
22 January 2024

PAPER

More than one way to skin a dose volume: the impact of dose-surface map calculation approach on study reproducibility

Haley M Patrick and John Kildea

Medical Physics Unit, McGill University, Montreal, QC, H4A3J1, Canada

E-mail: haley.patrick@mail.mcgill.ca

Keywords: dose surface map, dose–volume analysis, quality control, reproducibility

Supplementary material for this article is available online

Abstract

Objective. Dose-surface maps (DSMs) provide spatial representations of the radiation dose to organ surfaces during radiotherapy and are a valuable tool for identifying dose deposition patterns that are predictive of radiation toxicities. Over the years, many different DSM calculation approaches have been introduced and used in dose-outcome studies. However, little consideration has been given to how these calculation approaches may be impacting the reproducibility of studies in the field. Therefore, we conducted an investigation to determine the level of equivalence of DSMs calculated with different approaches and their subsequent impact on study results. *Approach.* Rectum and bladder DSMs were calculated for 20 prostate radiotherapy patients using combinations of the most common slice orientation and spacing styles in the literature. Equivalence of differently calculated DSMs was evaluated using pixel-wise comparisons and DSM features (rectum only). Finally, mock cohort comparison studies were conducted with DSMs calculated using each approach to determine the level of dosimetric study reproducibility between calculation approaches. *Main results.* We found that rectum DSMs calculated using the planar and non-coplanar orientation styles were non-equivalent in the posterior rectal region and that equivalence of DSMs calculated with different slice spacing styles was conditional on the choice of inter-slice distance used. DSM features were highly sensitive to choice of slice orientation style and DSM sampling resolution. Finally, while general result trends were consistent between the comparison studies performed using different DSMs, statisitically significant subregions and features could vary greatly in position and magnitude. *Significance.* We have determined that DSMs calculated with different calculation approaches are frequently non-equivalent and can lead to differing conclusions between studies performed using the same dataset. We recommend that the DSM research community work to establish consensus calculation approaches to ensure reproducibility within the field.

1. Introduction

Proper understanding of the dose-outcome responses of normal tissues is essential in order to be able to design radiotherapy treatment plans that minimize the likelihood of radiation toxicity. Traditionally, dose–volume histograms (DVHs) have been the primary tool used to derive dose-outcome relationships and dosimetric constraints for organs at risk (OARs) in radiotherapy research studies. These constraints may end up used in clinical practice to guide and evaluate the quality of individual treatment plans (Emami *et al* 1991, Bentzen *et al* 2010). However, DVH-based dose-outcome models lack spatial information and assume OARs have homogenous radiation sensitivities, potentially masking the existence of important radiosensitive subregions (Jaffray *et al* 2010, Acosta *et al* 2013). Therefore, for certain OARs, alternative dose-outcome analysis tools are of interest to the radiation oncology community.

One alternative to the DVH that preserves spatial information is the dose-surface map (DSM): a 2D projection of the dose to an organ's 3D surface. DSMs have mainly been used to study dose to the rectum and bladder (Buettner *et al* 2009, Palorini *et al* 2016, Shelley *et al* 2017), though several studies have also been published for other hollow organs such as the vagina, esophagus, duodenum, and heart (Witztum *et al* 2016, McWilliam *et al* 2020, Serban *et al* 2021). To date, DSMs have been used to identify spatial dose features and organ subregions predictive of early and late toxicities. In some cases, DSMs have been shown to be more predictive of radiation toxicities than DVHs (Buettner *et al* 2011, Acosta *et al* 2013, Palorini *et al* 2016, Mylona *et al* 2020).

Although promising as a dosimetric tool, it is important to note that the published methods of calculation and analysis of DSMs are much more diverse than is the case for DVHs. While nearly all DSMs are created by (1) defining slices of the organ of interest, (2) defining points around the surface of each slice to sample dose at, and (3) cutting open and unfurling the surface to create a 2D dose map, individual DSM implementations may use different approaches for each step. For instance, the DSM slices may all be oriented parallel to those of the treatment planning image (planar slicing) (Buettner *et al* 2009, Moulton *et al* 2017, Shelley *et al* 2017) or individually angled such that each slice is orthogonal to the organ's central axis path (non-coplanar slicing) (Heemsbergen *et al* 2005, Wortel *et al* 2015). Slices may also be separated using a set spacing for all patients (fixed spacing) (Palorini *et al* 2016), or with different spacing for each patient to ensure all DSMs contain the same number of slices (scaled spacing) (Buettner *et al* 2009, Mylona *et al* 2020). Analysis techniques are similarly diverse, with different groups comparing DSMs either in a pixel-wise manner or based on features. This diversity of calculation and analysis approaches can make it difficult to compare results between research studies and may be impacting the reproducibility of results in the field.

To date, the only DSM-based toxicity metrics that have been reproduced in the literature have been for late rectal bleeding (Buettner *et al* 2009, Heemsbergen *et al* 2020, Shelley *et al* 2020) and late bladder dysuria (Yahya *et al* 2017, Mylona *et al* 2020), despite many unconfirmed reports of other predictive metrics. Although cohort effects may play a role in the lack of reproducibility across studies, it is possible that variations in DSM calculation approaches may also be responsible. Determining the influence of calculation approaches on DSM-based findings is important, not only to help facilitate the consolidation of findings across DSM studies to firmly establish spatially-informed dosimetric constraints, but also to determine how dependent the clinical validity of these constraints is on the level of concordance between the DSM calculation approaches used in the clinic and in the research that was used to derive the constraints in the first place. With this in mind, the purpose of the present study was to determine the impact of DSM calculation approach on DSM topography and analysis for bladder and rectum structures. Specifically, we aimed to determine if:

(1) Choice of slice orientation style has a statistically significant effect on DSM topography and features (rectum only);

(2) Choice of slice spacing style has a statistically significant effect on DSM topography and features (rectum and bladder);

(3) The results and conclusions of a DSM-based cohort study are dependent on the DSM calculation approach used for analysis (rectum and bladder).

2. Methods

2.1. Patient cohort

To evaluate the effect of DSM calculation methodology on DSM topology, our analyzes for aims #1 and #2 were conducted at a population level, using a benchmark cohort of patients (to represent a retrospective research study), and at an individual representative patient level (to represent a single clinical case). The treatment plans of 20 moderate-risk prostate cancer patients treated at our centre between 2016 and 2017 were used as our retrospective patient cohort. One patient from the cohort with a rectum of median length was chosen as the representative patient. Simulation CT images, acquired on a Philips Big Bore CT scanner using a 3.0 mm slice thickness, were contoured according to RTOG guidelines for the male pelvis (Gay *et al* 2012). All patients were prescribed 60 Gy in 20 fractions to the prostate CTV alone, plus 7.0 mm isotropic PTV margins, using a two-arc VMAT approach. Plans were generated in the Eclipse treatment planning system (Varian Medical Systems, Palo Alto, CA) using previously-published treatment-planning constraints (Barbosa Neto *et al* 2015).

In order to facilitate the investigation of aim #3, a second comparison cohort was artificially created by calculating new dose distributions for each patient using smaller, 5.0 mm isotropic PTV margins. This yielded a paired cohort with predictable dose distribution differences from the benchmark cohort, which make it easier to assess how the comparison of two cohorts is affected by DSM calculation approach.

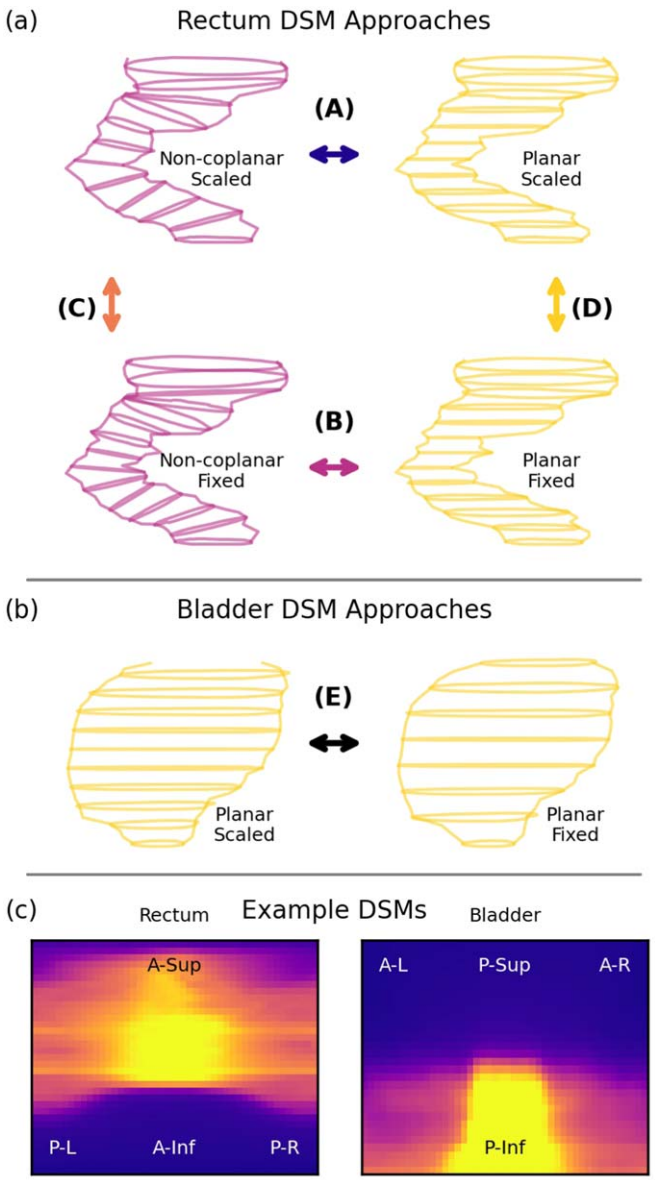

Figure 1. Visualization of all DSM calculation approaches and comparisons used in this work for (a) rectum and (b) bladder structures. Individual comparisons are indicated with capital letters and colors that are consistent across all figures in this article. Examples of typical bladder and rectum DSMs and how to interpret them are shown in (c).

2.2. DSM calculation workflow

As stated in the introduction, the two aspects of DSM calculation approach examined in this study were choice of slice orientation (planar or non-coplanar) and choice of slice spacing (scaled or fixed). Rectum DSMs were calculated using all four possible combinations of these aspects (figure 1(a)), whereas bladder DSMs were calculated with planar slice orientation only and the two different slice spacing approaches (figure 1(b)). This allowed us to reproduce the breadth of calculation approaches present in current DSM literature. All DSM calculations were performed using *rtdsm*, a recently-developed open-source Python package for DSM calculation and analysis (Patrick and Kildea 2022). *rtdsm* can calculate planar or non-coplanar DSMs using the standard RT-Structure and RT-Dose files from a DICOM-RT-compliant radiotherapy treatment planning system as input. In this work, voxel resolutions for these files were $1 \times 1 \times 3$ mm^3 and 2.5 mm^3 isotropic, respectively. For the calculation of fixed-spacing DSMs, a slice separation of 3.0mm (CT slice thickness) was used, whereas scaled-spacing DSMs fixed the total number of slices to the median number of CT slices (n_{slices}) the organ spanned for all patients in the cohort ($n_{\text{slices}} = 35$ for rectum, $n_{\text{slices}} = 25$ for bladder). When unfurling the surface doses to form the DSMs, rectum DSMs were cut open on the posterior side, and bladder DSMs on the anterior. These cut locations are typical in DSM research as they allow for the anticipated dose hotspots of these organs to be centered in their DSMs (figure 1(c)). All DSMs used a sampling resolution of 45 equiangular points per slice.

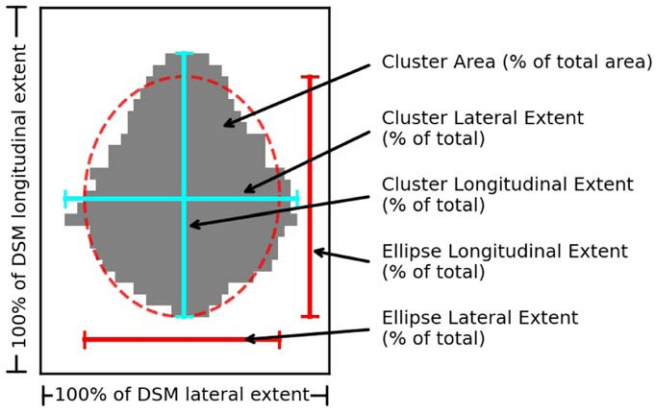

Figure 2. Visualization of the DSM features used in this work, calculated as originally defined in the literature (Buettner *et al* 2009, Moulton *et al* 2017). As is convention, each feature is reported as a percentage of either the total area or total dimensional extent (lateral or longitudinal) of the DSM used to calculate it.

2.3. DSM analysis technique

The average DSM of the benchmark cohort and the representative patient's DSM for each calculation approach were calculated and compared between approaches. To enable direct visual comparison of the effects of DSM calculation approach, dose difference maps (DDMs) were calculated for each comparison shown in figure 1 by subtracting the comparator DSMs. Because the DSMs in the benchmark cohort did not all contain the same number of slices when using fixed slicing, the average DSMs and DDMs in the population-level comparisons were truncated to the height of the shortest DSM in the cohort. Differences between DSMs owing to the different calculation approaches were quantified in two ways: (1) using pixel-wise comparisons through multiple comparisons permutation (MCP) testing, which is a commonly-used method to compare dose maps, and (2) using feature-based comparisons, as is popular for rectum DSMs. This dual analysis was performed in order to make findings easily translatable to the existing body of literature.

Pixel-wise comparison has been used in both bladder and rectum DSM research to identify subregions of either organ where statistically-meaningful differences in dose exist between two cohorts. While pixel-wise DSM comparisons are possible with pixel-wise t-tests and can be used to identify general areas where dose differences exist, it is good practice to apply a correction for multiple comparisons, such as with MCP testing, to reduce sensitivity to false positives (Chen *et al* 2013). The standard MCP test determines the similarity of two unpaired image-type datasets of the same anatomy and identifies pixels that vary significantly between the datasets while accounting for pixel-wise variance. As the datasets in our study were paired, we developed a paired implementation of the MCP test by modifying the permutation process to keep the labels of data pairs linked. Full details are provided in Supplement A.

Feature-based comparison is an analysis technique used to identify statistically-meaningful differences in isodose cluster characteristics that was developed for rectum DSMs (Buettner *et al* 2009, Moulton *et al* 2017). Features are derived by first creating a mask of a cluster of pixels for a given dose level and then extracting size, position, and shape metrics from either the mask itself or from an ellipse fitted to it, as initially performed by Buettner *et al* (2009). For this study, we opted to calculate the five most common features that have been reported in the literature: cluster area, cluster lateral and longitudinal extent, and ellipse lateral and longitudinal extent (figure 2) for four dose levels: 15, 35, 45, and 55 Gy. These dose levels were selected as they covered the full dose range of our data, and they matched the dose levels of toxicity-predictive features reported in other 2 Gy-per-fraction (or equivalent) studies (Buettner *et al* 2009, Moulton *et al* 2017, Onjukka *et al* 2019). Once calculated, we compared features between the cohorts using Wilcoxon signed-rank testing.

Statistical significance for both types of analysis was defined as $p \leqslant 0.05$. A Bonferroni correction of 4.0 was applied when comparing pairs of rectum DSM features to reduce false positives. In addition to comparing DSMs between calculation approaches for the same cohort, the analysis was also employed to compare average DSMs between the benchmark cohort and the artificially-generated comparison cohort (7.0 mm versus 5.0 mm PTV margins, respectively) for each of the DSM calculation approaches investigated. For each calculation approach, the average DSMs of the two cohorts were compared to assess how DSM calculation approach affects the ability to examine dosimetric differences between distinct cohorts.

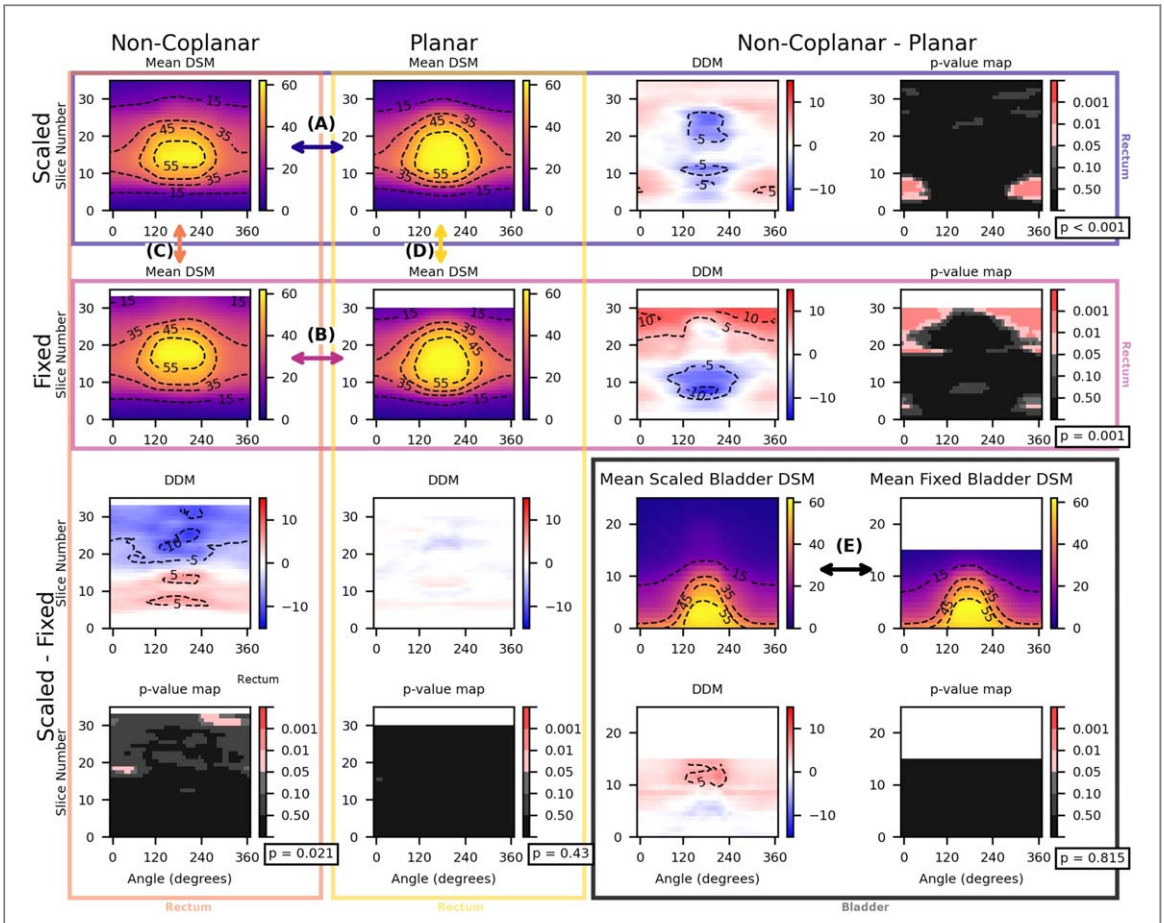

Figure 3. Quantitative comparisons of DSM calculation approaches for rectum (colored boxes) and bladder (black box) structures. Each comparison is indicated as a row or column with its letter label and arrow from figure 1, and a correspondingly colored box. Each comparison box contains: the cohort-mean DSMs of the two approaches being compared, their DDM, and the map of significantly different pixels between the two as determined by MCP testing. All maps are shown in units of Gy except for the MCP test result maps, which show *p*-values. Blank areas are due to truncating to the height of the shortest DSM and no calculations are made in these areas.

3. Results

3.1. Influence of calculation approach on pixel-wise comparisons

3.1.1. Aim 1: influence of slice orientation style

Side-by-side comparisons of the average rectum DSMs of the benchmark cohort produced using non-coplanar and planar slicing are shown in the first two rows of figure 3. Differences in the shapes of the 35, 45 and 55 Gy isodose clusters are visually observable between the two slice orientation styles (first column compared to second column), illustrating the influence of slice angling on DSMs. When scaled spacing is used (i.e. constant number of slices, row A, purple), the two DSM styles disagree significantly on dose to the posterior-inferior wall as shown in the corresponding DDMs and *p*-value maps. Similarly, significant disagreement is also observed when using fixed spacing (row B, magenta), only this time it occurs over a much larger area of the superior-posterior wall. In both cases, as indicated by the DDMs, the regions of significant disagreement appear to be caused by the non-coplanar DSMs measuring higher doses than their planar counterparts.

Similar patterns of disagreement are observable in the DSMs of the representative patient (figure 4) and provide further insight into probable underlying causes. There is a clear fundamental difference in the shape of the moderate (35 Gy) and high dose (55 Gy) regions between the non-coplanar and planar slicing methods (first column compared to second column in figures 3 and 4), which is likely related to slice angling and may explain the inferior differences. However, in the case of fixed spacing, an additional factor is introduced: difference in DSM length (i.e. number of slices) between the non-coplanar and planar calculation methods. The longer sampling path of the non-coplanar method requires more slices than the planar, stretching out the DSM and effectively desynchronizing information between the two styles the more superior in the DSM we go from the common inferior-most starting point. This may explain the difference between the non-coplanar and planar DSMs at the superior end when fixed spacing is used.

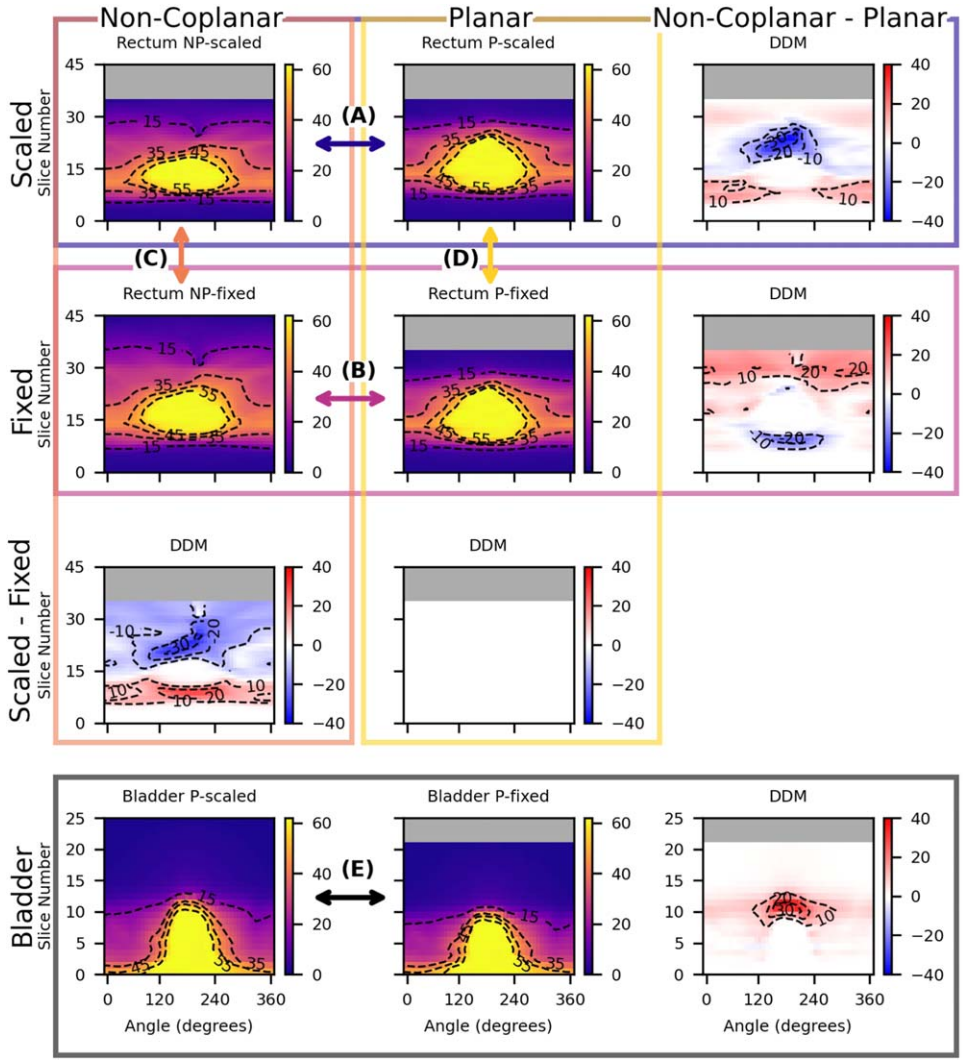

Figure 4. Comparisons of DSM calculation approaches for rectum (colored boxes) and bladder (black box) structures for our representative patient. Each comparison is indicated with its respective label, arrow, and colored box and includes the individual DSMs calculated by each approach along with their DDM. All maps are in units of Gy and truncated to the height of the shorter DSM where required.

3.1.2. Aim 2: influence of slice spacing style

Comparisons of rectum DSMs created with fixed and scaled spacing are presented in columns C (orange) and D (yellow) of figure 3, as well as for the representative patient in figure 4. Corresponding DDMs (scaled minus fixed) and *p*-value maps are found in the rows beneath. Overall, slice spacing style has much less of an effect on rectum DSM topography than slice orientation style as evidenced by the lack of significant regions in the *p*-value maps. For the non-coplanar DSMs (column C, orange), the DDM suggests a shift or rescaling of the DSM topography in the superior-inferior direction between the two spacing styles. This rescaling effect is also visible in non-coplanar scaled-minus-fixed DDM of the representative patient (figure 4(C)), which also clearly demonstrates how fixed slice spacing requires more slices to represent the entire rectum. Despite this clearly introducing a desynchronization effect, similar to that between non-coplanar and planar fixed spacing DSMs, significant disagreement is only found for two small patches of the superior rectum. Planar DSMs are even less affected by choice of slice spacing style, as no significant sites of disagreement were observed between the planar scaled and planar fixed DSMs. We note that, based on the DDMs of the representative patient, this result may have occurred due to our choice to use an n_{slices} value for scaled DSMs that equaled the number of slices in our median rectum, causing the median effective slice distance to be approximately equal to the fixed DSM spacing (3 mm).

In the case of bladder DSMs, those created with scaled spacing contain noticeably more slices than those created with fixed spacing (figures 3 and 4, box E), which can be attributed to the need to truncate the fixed spacing DSMs to the height of the shortest bladder in the cohort. Nevertheless, while the DDM suggests slice desynchronization akin to the rectum DSMs, no significant disagreement was observed between the two DSM

slice spacing approaches across the slices they had in common. Once again, this is most likely attributed to our choice of n_{slices} for scaled spacing being similar to the number of slices required to construct a fixed spacing DSM for the average patient.

3.2. Influence of calculation approach on DSM features (aims 1 and 2)

DSM features were calculated and compared for rectum DSMs for all four calculation approaches, and the results for the 15 and 55 Gy clusters are shown in figure 5 (figures of the 35 and 45 Gy clusters, and the representative patient are available in supplement B). Features differ between calculation approaches, particularly between the planar (yellow) and non-coplanar (magenta) slicing styles. Longitudinal features are generally larger for planar DSMs at higher dose levels (figures 5(c), (e)), consistent with the pixel-wise findings when comparing planar and non-coplanar mean DSMs (figure 3), whereas they are larger for non-coplanar DSMs at the 15 Gy isodose level. Interestingly, this trend of features being larger for non-coplanar DSMs at low doses also extends to ellipse lateral extent (figure 5(d)), potentially as a consequence of how ellipses were fitted to larger clusters. While pixel-wise comparisons of mean DSM isodose regions do not suggest particularly notable differences in lateral spans between slice orientation approaches (figure 3), it is possible that choice of slice orientation approach subtly influences the shape of the clusters and fitted ellipses used in feature-based analysis.

While scaled and fixed DSM features are generally in more agreement with one another for a given slice orientation style, disagreements do occur for certain area and ellipse-based features (figures 5(a), (d), (e)). This is somewhat unexpected as DSM features are conventionally reported as percentages and thereby should not disagree significantly when DSMs are only effectively rescaled. As the primary difference between our scaled and fixed DSMs was longitudinal sampling resolution (fixed: 3 mm, scaled: variable per patient), these results may indicate that DSM features may not be stable between different sampling resolutions.

3.3. Influence of calculation approach on the conclusions of a cohort comparison (aim 3)

Rectum and bladder doses were compared between the benchmark cohort with 7.0 mm margins and the comparison cohort with 5.0 mm margins using DSMs calculated with all calculation approach variants, and the consistency of the results was assessed across approaches. As designed, the cohorts were found to be statistically dissimilar from one another when using all styles of DSMs. However, these dissimilarities varied in their magnitude and spatial localization according to the DSM calculation style used.

For rectum DSMs, the higher dose ring present in all DDMs (7.0 mm minus 5.0 mm margins) changed in both shape and magnitude depending on calculation approach (figures 6(a)–(h)). MCP testing p-value maps also indicated that the locations of statistically significant subregions (SSRs) depended on DSM type (figure 6(a)). Rectum DSM features were relatively consistent between calculation approaches (figure 57), with all four styles of DSM generally agreeing on whether or not a feature differed significantly between the two cohorts. However, feature magnitudes did differ between approaches, especially for the longitudinal features. For example, the mean difference in the 15 Gy cluster longitudinal extent for planar scaled DSMs was notably different in magnitude from the other DSM approaches (figure 7(c)). Similarly, there was reduced consistency in ellipse longitudinal extent mean differences between DSM calculation approaches at the 55 Gy dose level (figure 7(e)).

Bladder DSM findings were much less consistent between DSM styles when the two artificially different cohorts were compared. DDMs show cold spots in the fixed spacing comparison that are not present in the scaled spacing comparison (figures 6(i)-(l)). While MCP testing does identify significantly differing pixels in the right and left inferior bladder for both DSM styles, the size and shape of these regions are different for the two calculation approaches. Tests conducted with the fixed-spacing dataset also failed to identify the SSR found by the scaled-spacing dataset in the superior bladder, likely a product of DSM truncation (figures 6(j), (l)).

4. Discussion

DSM calculation approaches are diverse and can differ considerably between research groups. Although all DSM research studies share the same general goal of identifying dosimetric spatial factors that are predictive of radiation toxicities, little to no work has been done to assess the reproducibility of these spatial factors between different DSM calculation approaches. To the best of our knowledge, this paper is the first quantitative investigation of analysis sensitivity to DSM calculation approach. We have identified that significant disagreement between DSMs can occur when different calculation approaches are used. In the discussion of our findings below, we refer to figures 8 and 9, in which we have attempted to graphically illustrate the influences of the DSM calculation approaches we examined in this work.

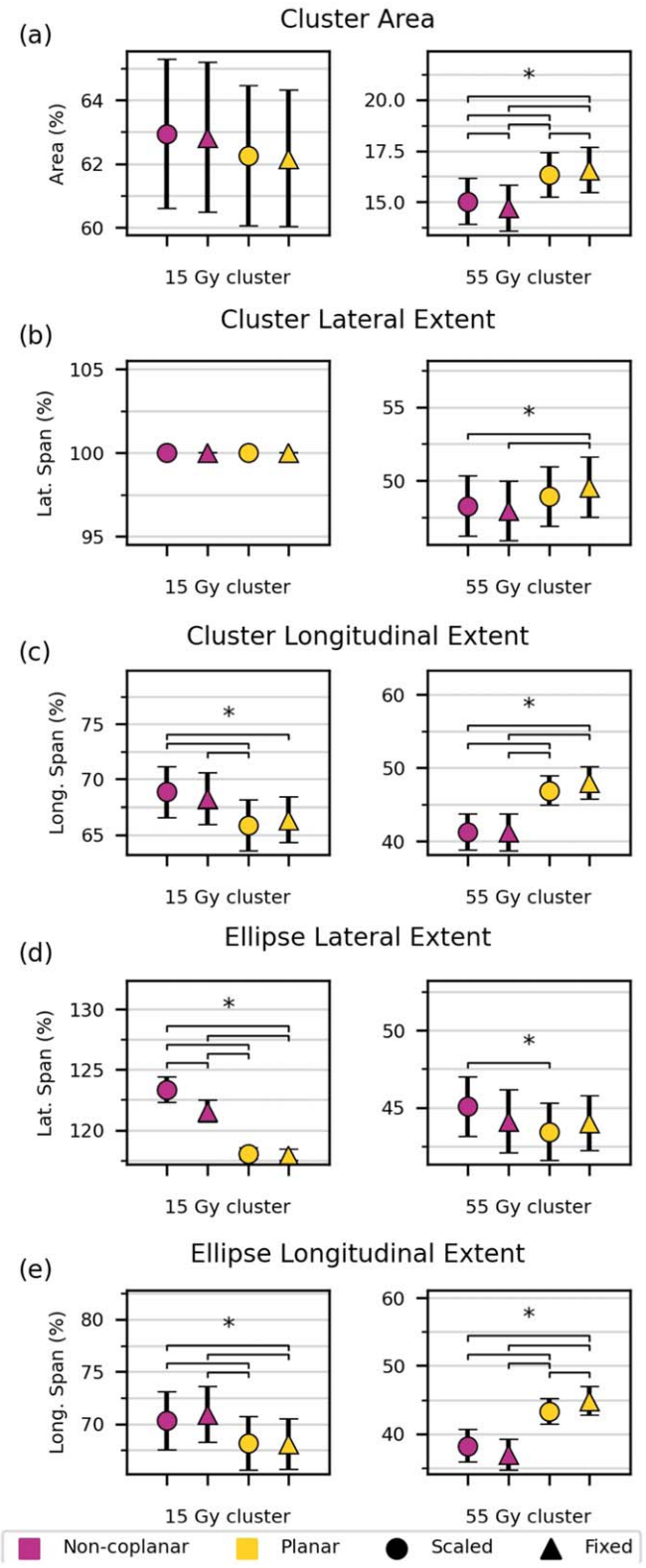

Figure 5. Mean DSM feature values (and standard uncertainty of the mean) from the four rectum DSM calculation approaches. In order, they are (a) cluster area, (b) cluster lateral extent, (c) cluster longitudinal extent, (d) ellipse lateral extent, and (e) ellipse longitudinal extent. Significantly different pairs of features are indicated with a bracket and an asterisk. Figures of the other dose levels are provided in the supplemental material.

4.1. Aim 1: Equivalence of planar and non-coplanar DSMs

Although they have existed side-by-side in the literature for nearly two decades, we found that rectum DSMs calculated using planar and non-coplanar slicing approaches are non-equivalent in two specific regions. Namely, the inferior–posterior wall when using scaled slice spacing and the superior-posterior wall when using

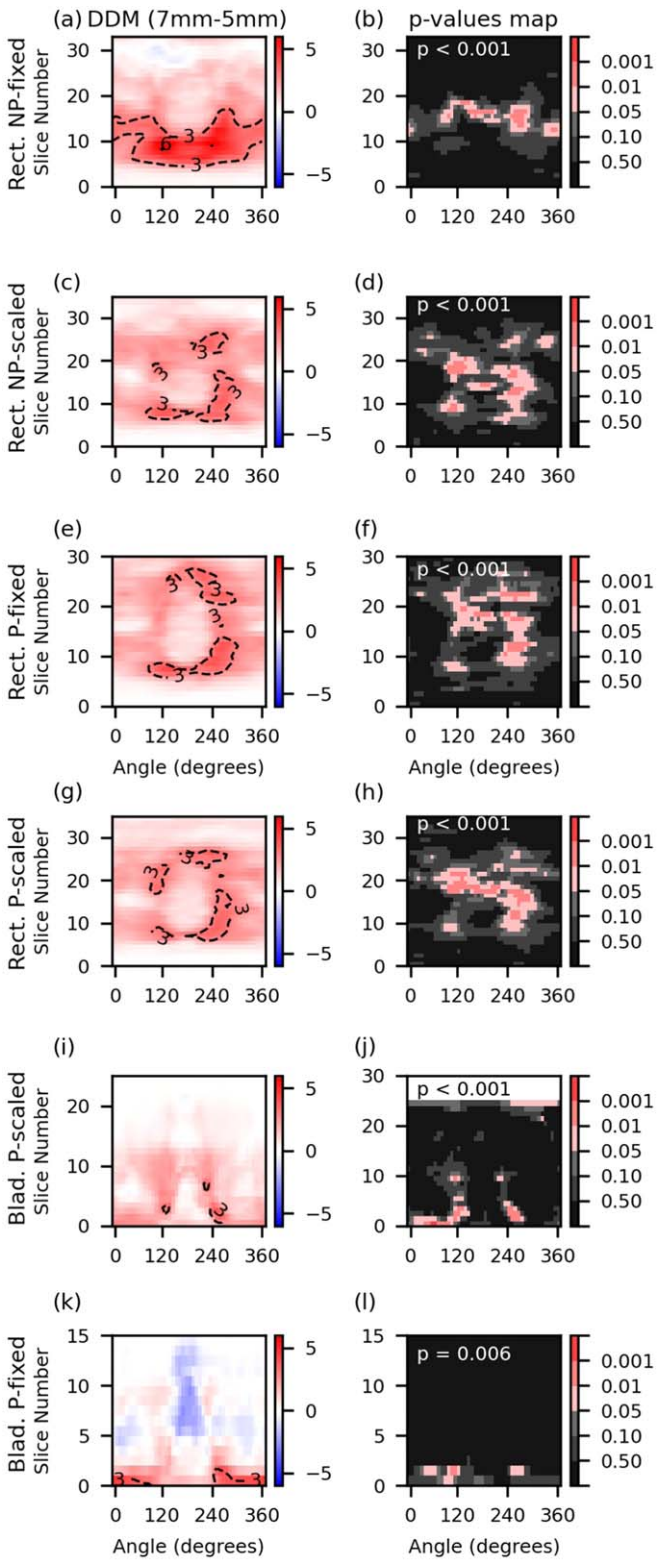

Figure 6. DDMs (left) and maps of significantly different pixels (right) between the benchmark and comparison cohorts for each of the DSM calculation approaches tested for the rectum (a)–(h) and bladder (i)–(l). 'P' refers to planar slice orientations and 'NP' to non-coplanar. DDMs are in units of Gy.

fixed spacing (figure 3). These non-equivalent regions stem from differences in sampling point locations introduced by the slicing methods themselves, causing the same slice to sample dose in different locations between the two approaches (as illustrated in figures 8(a)–(b)).

An ideal DSM calculation technique would yield the most easily interpretable representation of dose to an organ's surface. The most intuitive representation of a rectum's surface is a cylinder sampled at regular grid intervals that is then bent to match the rectum's shape. Non-coplanar sampling attempts to recreate this,

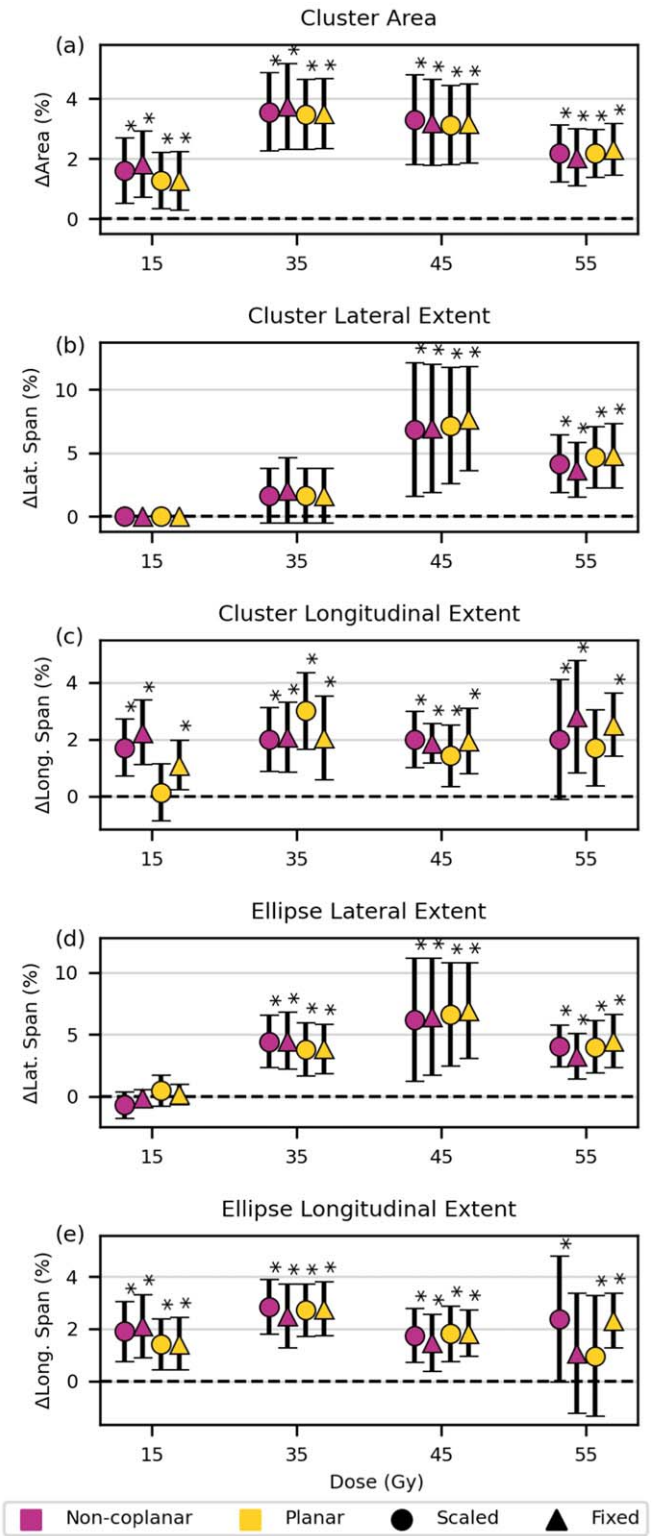

Figure 7. Mean differences in DSM feature values between the benchmark and comparison cohorts for each of the four rectum DSM calculation approaches. Error bars indicate 95% confidence intervals and statistically significant differences between the cohorts are indicated with asterisks.

estimating where these grid intervals lie and producing a surface map that represents the rectum deformed back to a basic cylindric representation. Planar sampling however, simply assumes the rectum is a cylinder, regardless of how non-linear its path is. While this assumption may hold for very straight tubular organs, our findings demonstrate it does not for rectums. Consequently, surface doses presented by planar DSMs will increasingly distort relative to our intuitive spatial understanding of the rectum's surface the less linear a rectum's path is. For

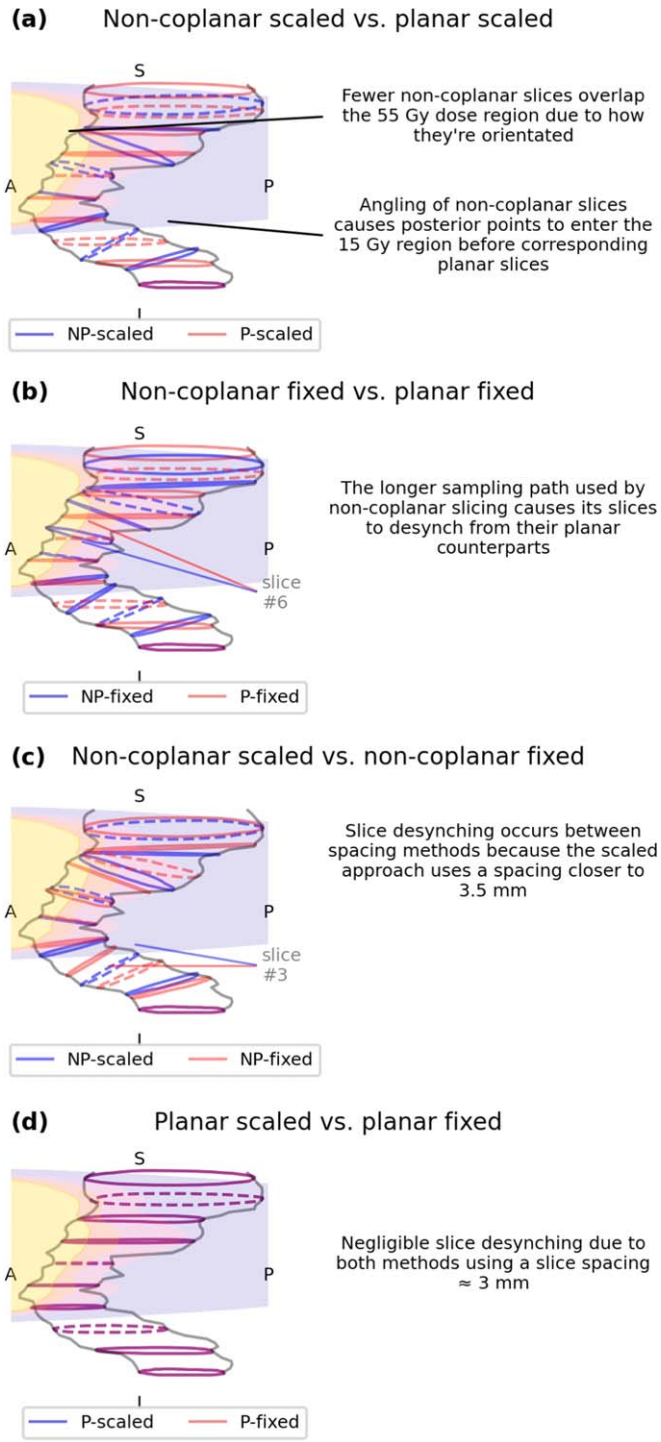

Figure 8. Examples of how the choice of DSM calculation approach influences the slices of the rectum (and subsequent appearance of the DSM) of the representative patient. Subfigures cover each comparison made in this work. The 15, 35, and 55 Gy isodose regions are included as purple, pink, and yellow overlays, respectively, for dosimetric context. Additionally, for visual clarity, only every fourth slice is shown. Dashed lines are used to represent every 12th slice to make visual comparisons of relative slice positions between slicing styles easier.

this reason, we recommend the use of non-coplanar DSM slicing approaches where possible and to be aware and investigate the influence of the planar approach on DSM topology before proceeding otherwise.

4.2. Aim 2: Equivalence of scaled and fixed slice spacing DSMs

In our testing, we found that scaled and fixed slice spacing approaches were roughly equivalent to one another for our chosen comparison scenarios, wherein the fixed and median effective scaled slice spacings were both approximately equal to 3 mm. As some evidence of slice desynching was still observed between spacing approaches (figures 3(C), (E)), we suspected that these results may be conditional on our choice of matching slice

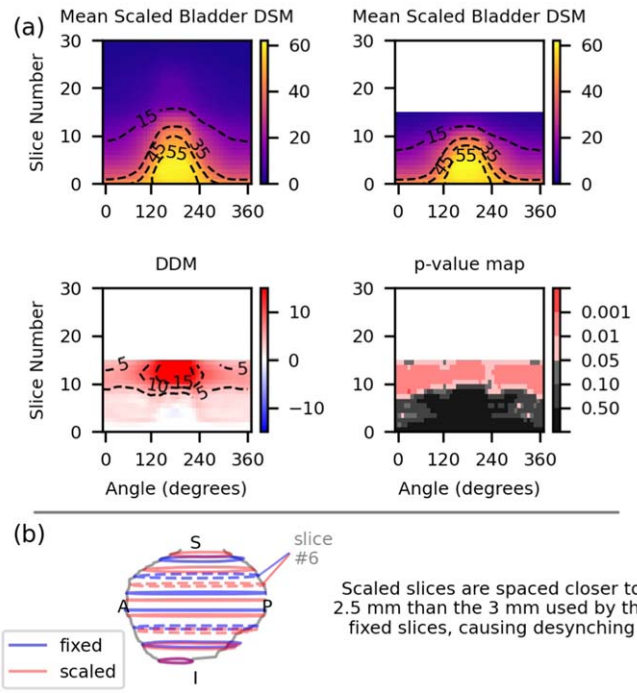

Figure 9. (a) Comparison of scaled bladder DSMs calculated with 30 scaled spacing slices (median 2.5 mm) and fixed slice spacing DSMs. Mean DSMs are shown for both approaches along with their DDM and *p*-value map. (b) Visual demonstration of the slice desynching that occurs between scaled and planar bladder DSMs. Only every 4th slice is shown for visual clarity, with every 12th slice represented with dashed lines.

spacing distances and could change if we used a different combination (figures 8(c)–(d)). To demonstrate the possible influence of this, figure 9 shows what the results of a comparison of fixed bladder DSMs (with a spacing of 3 mm) versus scaled bladder DSMs with 5 additional slices (meaning a median effective slice spacing 2.5 mm instead of 3 mm) would be. As shown, when effective slice spacing resolutions are not equal slice desynching is much greater, causing the DSMs to be more dissimilar. Based on this apparent sensitivity, we would heavily advise against direct comparisons of DSMs that use different slice spacing resolutions and encourage relative, scaled comparisons instead. We would also like to further stress the importance of using scaled DSMs when working with organs that exhibit significant fluctuations in size in the superior-inferior dimension such as the bladder. When fixed spacing DSMs are created for these organs, dose to the superior-most region of smaller structures is compared to dose to central regions of larger ones (as can be noted by the degree of bladder DSM truncation in figures 3 and 9). In general, we recommend fixed DSMs should be avoided unless (1) superior-inferior spans are consistent across a structure's population and (2) there is an explicit need to study spatial dose in absolute distance (e.g. dose to interior-most 3 cm).

4.3. Aim 3: DSM Feature robustness again DSM calculation method

Rectal DSM features were originally designed for planar DSMs and their specific patterns of dose topography (Buettner *et al* 2009). As such, it is not surprising that features calculated from non-coplanar DSMs are non-equivalent to their planar counterparts (figure 5). Due to their sampling approach, non-coplanar DSMs have smaller high dose isodose clusters and less elliptical low dose isodose clusters (figure 4), impacting the calculation of ellipse-based features. For these reasons, we recommend that DSM features continue to be calculated from planar DSMs only to facilitate feature-specific reproducibility. However, slice spacing approach must also be considered. We observed significant differences between features calculated with fixed and scaled slice spacing approaches, suggesting a possible resolution effect as well. This is concerning, as DSM resolution is one of the most variable factors between studies. From what we have observed in the literature, reported resolutions vary from 21 × 21 to 200 × 200 pixels (Buettner *et al* 2009, Onjukka *et al* 2019) and can be achieved through either direct sampling (Palorini *et al* 2016, Shelley *et al* 2017) or interpolation methods (Casares-Magaz *et al* 2017, Moulton *et al* 2017, Heemsbergen *et al* 2020), which may introduce their own effects. Currently there is little to no discussion on what the optimal resolution for a DSM is or how it should be determined, though we expect it to be likely related to the resolution of the voxels of the CT image and RT-dose grid used to create the DSM. Because of this, we strongly recommend that new feature-based studies choose DSM sampling resolutions

that are consistent with the previous studies they plan to compare to until a DSM resolution standard is established.

4.4. Relation to current state of reproducibility in the literature

To date, few studies in the DSM literature have agreed on what DSM information is predictive of radiation toxicities. Most discussions focus more on general trends that persist across studies, such as increased dose to the posterior rectum (Wortel *et al* 2015, Moulton *et al* 2017) and the bladder trigone (Palorini *et al* 2016) causing increasing toxicity risk, and usually point to cohort effects to explain differences (Mylona *et al* 2020). While the impacts of cohort, fractionation scheme, and analysis techniques on reproducibility cannot be discounted, our findings highlight that is also important to consider DSM calculation approaches when trying to understand the similarities (or dissimilarities) between published results.

Reproduced rectal DSM-toxicity results exist only for rectal bleeding and are limited to the 51 Gy cluster area (Buettner *et al* 2009, Moulton *et al* 2017), the 40–60 Gy lateral ellipse extents (Buettner *et al* 2009, Shelley *et al* 2017), and dose to the inferior quarter of the rectum (Moulton *et al* 2017, Heemsbergen *et al* 2020, Shelley *et al* 2020). Corroborated feature results were all obtained using planar DSMs with resolutions between 21 $\times$ 21 and 51 $\times$ 45 pixels, whereas papers using higher resolution DSMs ($\geqslant$200 $\times$ 200) reported no reproduced features (Casares-Magaz *et al* 2017, Onjukka *et al* 2019). It is also worthwhile to note that the reproduced rectal bleeding SSR was located inferiorly, where slice desynching effects are expected to be minimal between the two different slice orientation styles used by these authors (planar and non-coplanar). In contrast, other non-reproduced SSRs, like those for proctitis (Wortel *et al* 2015, Moulton *et al* 2017, Shelley *et al* 2020) and incontinence (Onjukka *et al* 2019, Heemsbergen *et al* 2020, Shelley *et al* 2020), were distributed in different locations in the superior half of the rectum, which we have observed to be more prone to slice desynching between calculation methods (figure 8).

Bladder DSM SSR reproducibility has been limited to dose to the inferior-anterior bladder being predictive of late dysuria and is the subject of a study by Mylona *et al* (2020). Once again, we note that this SSR is located in the inferior organ, where we would expect the least discordance between Mylona's planar scaled slice spacing DSMs and the planar fixed slice spacing DSMs of other published studies (Palorini *et al* 2016, Yahya *et al* 2017). However, we would also like to highlight the two SSRs Mylona found for acute and late retention. These were located in the superior half of the bladder, in the region of more significant slice desynching and above the level at which the fixed slice spacing DSM cohorts of the other studies truncated the maximum extent of their bladders. This truncation handicapped the comparability of these studies and is a noteworthy example of why consensus DSM calculation approaches are needed.

Although variations in DSM calculation methodology can help explain the state of reproducibility in our field, we recognize other factors do need to be considered as well. In addition to commonly discussed cohort or analysis differences, it is worth noting the role that different outcome reporting metrics may play. Choice of toxicity scoring instrument also varies greatly between studies (e.g. CTCAE, UCLA-QoL, IPSS, custom patient-reported outcome measures, (Improta *et al* 2016, Moulton *et al* 2017, Heemsbergen *et al* 2020)), as do the timepoints at which outcomes are collected, especially for late effects (first timepoint range: 3–27 months, (Casares-Magaz *et al* 2017, Moulton *et al* 2017)). Considering that toxicity scoring concordance has been shown to be limited between observers and scoring instruments (Denis *et al* 2003, Atkinson *et al* 2016), further investigations into their effect are warranted.

5. Conclusion

We have determined that different DSM calculation approaches produce non-equivalent DSMs that can impact the conclusions of a given study. This has the potential to limit clinical translation of DSM-based research unless measures are taken. Ideally, the community should establish standardized methodologies to calculate DSMs for each organ of interest, and at a minimum better awareness of DSM non-equivalencies is needed. While further discussions within the community are required to establish any sort of consensus, we wish to present the following recommendations for consideration:

(1) **A planar scaled slice spacing approach should be used to calculate DSMs of sphere-like organs.** This is especially true for organs, such as the bladder, that exhibit significant isotropic volume changes between subjects, as the scaling ensures the same anatomical regions are represented by the same DSM slices.

(2) **Non-coplanar slicing should be used for tubular organs with significant curvatures.** This includes organs like the rectum and duodenum where planar slicing cannot accurately account for the organ's trajectory. Scaled slice spacing should be used unless sufficient use-case-specific justification for fixed spacing is given.

(3) **The planar slicing approach is acceptable for straight tubular organs or organs with curvature in specific use cases.** These include the esophagus or spinal cord, or rectum DSMs in the context of feature-based analysis.

(4) **Consensus calculation approaches should be developed for each organ by the DSM research community.** While some approaches may be more straightforward to decide, such as for the bladder, discussions will be necessary for organs with more complex geometries like the rectum.

(5) **Data sharing should be encouraged within the community to better evaluate study reproducibility between different DSM code implementations.** This could be facilitated through either the sharing of anonymized DICOM files, or arrays of surface vertices and dose matrices.

(6) **Open-source DSM calculation codes should be encouraged. This can be facilitated by code-sharing sites such as Github.** The DSM calculation code, *rtdsm*, is an example.

Acknowledgments

HMP acknowledges funding from the Fonds de recherche du Québec—Santé in the form of a Doctoral Training Award, as well as partial support by the CREATE Responsible Health and Healthcare Data Science (SDRDS) grant of the Natural Sciences and Engineering Research Council. JK acknowledges support from from the Canada Foundation for Innovation John R Evans Leaders Fund. Logistical support provided by the Medical Physics Department at the McGill University Health centre was invaluable for this work.

Data availability statement

All data that support the findings of this study are available on reasonable request to the authors (and any supplementary information files).

Ethical statement

This retrospective study was approved by the Research Ethics Board of the McGill University Health Centre (project number 2024-9506) with a waiver of informed consent as the study was retrospective in nature and the collected data was limited to radiotherapy treatment dose distributions and contour structures. We confirm that all work of the study was conducted in accordance with REB guidelines and regulations that adhere to the Declaration of Helsinki.

ORCID iDs

John Kildea https://orcid.org/0000-0002-7084-1425

References

Acosta O, Drean G, Ospina J D, Simon A, Haigron P, Lafond C and De Crevoisier R 2013 Voxel-based population analysis for correlating local dose and rectal toxicity in prostate cancer radiotherapy *Phys. Med. Biol.* **58**(**8**) 2581–95

Atkinson T M *et al* 2016 The association between clinician-based common terminology criteria for adverse events (CTCAE) and patient-reported outcomes (PRO): a systematic review *Support Care Cancer* **24**(**8**) 3669–76

Barbosa Neto O, Souhami L and Faria S 2015 Hypofractionated radiation therapy for prostate cancer: the mcgill university health center experience *Cancer/Radiotherapie* **19** 431–6

Bentzen S M, Constine L S, Deasy J O, Eisbruch A, Jackson A, Marks L B, Ten Haken R K and Yorke E D 2010 Quantitative Analyses of Normal Tissue Effects in the Clinic (QUANTEC): an introduction to the scientific issues *Int. J. Radiat. Oncol. Biol. Phys.* **76** S3–9

Buettner F, Gulliford S L, Webb S and Partridge M 2011 Modeling late rectal toxicities based on a parameterized representation of the 3D dose distribution *Phys. Med. Biol.* **56**(**7**) 2103–18

Buettner F, Gulliford S L, Webb S, Sydes M R, Dearnaley D P and Partridge M 2009 Assessing correlations between the spatial distribution of the dose to the rectal wall and late rectal toxicity after prostate radiotherapy: an analysis of data from the MRC RT01 trial (ISRCTN 47 772 397) *Phys. Med. Biol.* **54**(**21**) 6535–48

Casares-Magaz O, Muren L P, Moiseenko V, Petersen S E, Pettersson N J, Høyer M, Deasy J O and Thor M 2017 Spatial rectal dose/volume metrics predict patient-reported gastro-intestinal symptoms after radiotherapy for prostate cancer *Acta Oncol.* **56**(**11**) 1507–13

Chen C, Witte M, Heemsbergen W and Herk M V 2013 Multiple comparisons permutation test for image based data mining in radiotherapy *Radiat. Oncol.* **8** 293

Denis F *et al* 2003 Late toxicity results of the GORTEC 94-01 randomized trial comparing radiotherapy with concomitant radiochemotherapy for advanced-stage oropharynx carcinoma: comparison of LENT/SOMA, RTOG/EORTC, and NCI-CTC scoring systems *Int. J. Radiat. Oncol. Biol. Phys.* **55**(**1**) 93–8

Emami B, Lyman J, Brown A, Cola L, Goitein M, Munzenrider J E, Shank B, Solin L J and Wesson M 1991 Tolerance of normal tissue to therapeutic irradiation *Int. J. Radiat. Oncol. Biol. Phys.* **21** 109–22

Gay H A *et al* 2012 Pelvic normal tissue contouring guidelines for radiation therapy: a radiation therapy oncology group consensus panel atlas *Int. J. Radiat. Oncol. Biol. Phys.* **83** 1–86

Heemsbergen W D, Hoogeman M S, Hart G A, Lebesque J V and Koper P C 2005 Gastrointestinal toxicity and its relation to dose distributions in the anorectal region of prostate cancer patients treated with radiotherapy *Int. J. Radiat. Oncol. Biol. Phys.* **61**(**4**) 1011–8

Heemsbergen W D, Incrocci L, Pos F J, Heijmen B J and Witte M G 2020 Local dose effects for late gastrointestinal toxicity after hypofractionated and conventionally fractionated modern radiotherapy for prostate cancer in the HYPRO trial *Front. Oncol.* **10** 469

Improta I *et al* 2016 Bladder spatial-dose descriptors correlate with acute urinary toxicity after radiation therapy for prostate cancer *Phys. Med.* **32**(**12**) 1681–9

Jaffray D A, Lindsay P E, Brock K K, Deasy J O and Tomé W A 2010 Accurate accumulation of dose for improved understanding of radiation effects in normal tissue *Int. J. Radiat. Oncol. Biol. Phys.* **76** S135–9

McWilliam A, Dootson C, Graham L, Banfill K, Abravan A and van Herk M 2020 Dose surface maps of the heart can identify regions associated with worse survival for lung cancer patients treated with radiotherapy *Phys. Imaging Radiat. Oncol.* **30** 46–51

Moulton C R, House M J, Lye V, Tang C I, Krawiec M, Joseph D J, Denham J W and Ebert M A 2017 Spatial features of dose-surface maps from deformably-registered plans correlate with late gastrointestinal complications *Phys. Med. Biol.* **62**(**10**) 4118–39

Mylona E *et al* 2020 Local dose analysis to predict acute and late urinary toxicities after prostate cancer radiotherapy: assessment of cohort and method effects *Radiother. Oncol.* **147** 40–9

Onjukka E *et al* 2019 Patterns in ano-rectal dose maps and the risk of late toxicity after prostate IMRT *Acta Oncol.* **58**(**12**) 1757–64

Palorini F, Cozzarini C, Gianolini S, Botti A, Carillo V, Iotti C, Rancati T, Valdagni R and Fiorino C 2016 First application of a pixel-wise analysis on bladder dose-surface maps in prostate cancer radiotherapy *Radiother. Oncol.* **119**(**1**) 123–8

Patrick H M and Kildea J 2022 Technical note: rtdsmAn open-source software for radiotherapy dose-surface map generation and analysis *Med. Phys.* **49** 7327–35

Serban M, de Leeuw A A, Tanderup K and Jürgenliemk-Schulz I M 2021 Vaginal dose-surface maps in cervical cancer brachytherapy: Methodology and preliminary results on correlation with morbidity *Brachytherapy* **20**(**3**) 565–75

Shelley L E *et al* 2017 Delivered dose can be a better predictor of rectal toxicity than planned dose in prostate radiotherapy *Radiother. Oncol.* **123**(**3**) 466–71

Shelley L E, Sutcliffe M P, Thomas S J, Noble D J, Romanchikova M, Harrison K, Bates A M, Burnet N G and Jena R 2020 Associations between voxel-level accumulated dose and rectal toxicity in prostate radiotherapy *Phys. Imaging Radiat. Oncol.* **14** 87–94

Witztum A, George B, Warren S, Partridge M and Hawkins M A 2016 Unwrapping 3D complex hollow organs for spatial dose surface analysis *Med. Phys.* **43**(**11**) 6009

Wortel R C, Witte M G, van der Heide U A, Pos F J, Lebesque J V, van Herk M, Incrocci L and Heemsbergen W D 2015 Dose-surface maps identifying local dose-effects for acute gastrointestinal toxicity after radiotherapy for prostate cancer *Radiother. Oncol.* **117**(**3**) 515–20

Yahya N, Ebert M A, House M J, Kennedy A, Matthews J, Joseph D J and Denham J W 2017 Modeling urinary dysfunction after external beam radiation therapy of the prostate using bladder dose-surface maps: evidence of spatially variable response of the bladder surface *Int. J. Radiat. Oncol. Biol. Phys.* **97**(**2**) 420–6