IOP Publishing Phys. Med. Biol. 65 (2020) 055014 (12pp) https://doi.org/10.1088/1361-6560/ab6e54

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

PAPER

RECEIVED
2 July 2019

REVISED
7 January 2020

ACCEPTED FOR PUBLICATION
21 January 2020

PUBLISHED
6 March 2020

Data clustering to select clinically-relevant test cases for algorithm benchmarking and characterization

Sarah Weppler[1,2,5], Colleen Schinkel[2,3], Charles Kirkby[1,3,4] and Wendy Smith[1,2,3]

[1] Department of Physics and Astronomy, University of Calgary, 2500 University Dr NW, Calgary, Alberta, T2N 1N4, Canada
[2] Department of Medical Physics, Tom Baker Cancer Centre, 1331 29 St NW, Calgary, Alberta, T2N 4N2, Canada
[3] Department of Oncology, University of Calgary, 2500 University Dr NW, Calgary, Alberta, T2N 1N4, Canada
[4] Department of Medical Physics, Jack Ady Cancer Centre, 960 19 St S, Lethbridge, Alberta, T1J 1W5, Canada
[5] Author to whom any correspondence should be addressed.

E-mail: Sarah.Weppler@albertahealthservices.ca, Colleen.Schinkel@albertahealthservices.ca, Charles.Kirkby@albertahealthservices.ca and Wendy.Smith@albertahealthservices.ca

Keywords: algorithm benchmarking, data clustering, test case selection

Abstract

Algorithm benchmarking and characterization are an important part of algorithm development and validation prior to clinical implementation. However, benchmarking may be limited to a small collection of test cases due to the resource-intensive nature of establishing 'ground-truth' references. This study proposes a framework for selecting test cases to assess algorithm and workflow equivalence. Effective test case selection may minimize the number of ground-truth comparisons required to establish robust and clinically relevant benchmarking and characterization results.

To demonstrate the proposed framework, we clustered differences between two independent workflows estimating during-treatment dose objective violations for 15 head and neck cancer patients (15 planning CTs, 105 on-unit CBCTs). Each workflow used a different deformable image registration algorithm to estimate inter-fractional anatomy and contour changes. The Hopkins statistic tested whether workflow output was inherently clustered and k-medoid clustering formalized cluster assignment. Further statistical analyses verified the relevance of clusters to algorithm output. Data at cluster centers ('medoids') were considered as candidate test cases representative of workflow-relevant algorithm differences.

The framework indicated that differences in estimated dose objective violations were naturally grouped (Hopkins = 0.75, providing 90% confidence). K-medoid clustering identified five clusters which stratified workflow differences (MANOVA: $p < 0.001$) in estimated parotid gland D50%, spinal cord/brainstem Dmax, and high dose CTV coverage dose violations (Kendall's tau: $p < 0.05$). Systematic algorithm differences resulting in workflow discrepancies were: parotid gland volumes (ANOVA: $p < 0.001$), external contour deformations (t-test: $p = 0.022$), and CTV-to-PTV margins (t-test: 0.009), respectively. Five candidate test cases were verified as representative of the five clusters.

The framework successfully clustered workflow outputs and identified five test cases representative of clinically relevant algorithm discrepancies. This approach may improve the allocation of resources during the benchmarking and characterization process and the applicability of results to clinical data.

Introduction

Algorithm benchmarking and characterization are important processes in medical physics, used in product development and comparison prior to clinical implementation. Comparing algorithms is particularly important when specifying which algorithms are acceptable for use as part of a clinical trial workflow. When centres are considering version updates or algorithm changes, effective assessment of algorithm differences is also relevant.

Quantitative guidelines for algorithm comparisons are often well established. However, for certain applications the labour-intensive nature of establishing expert-defined 'ground truth' may limit benchmarking to a

small collection of test cases, such as dose calculation algorithms (Court *et al* 2010, Fragoso *et al* 2010, Ehler *et al* 2014) and deformable image registration (DIR) (Kumarasiri *et al* 2014, Pukala *et al* 2016, Loi *et al* 2018). For DIR benchmarking and characterization, both algorithm and workflow output may be affected by anatomical site (Kashani *et al* 2008, Hoffmann *et al* 2014), contour size (Kumarasiri *et al* 2014, Mencarelli *et al* 2014), and different implementations of a similar algorithmic framework (Kashani *et al* 2008). Physical phantoms, digital phantoms, and literature case studies are helpful but may not be always representative of application-specific clinical cases. Therefore, it is generally recommended to repeat algorithm comparisons against ground truth for each institution-specific clinical application.

As computing power continues to increase, the resources required for running algorithms is generally diminishing. Scripting and automation capabilities are increasing the ease of producing workflow output. In contrast, the establishment of manual gold standards and expert review of results remain labour-intensive processes (for example, Brock *et al* 2017). We propose that information contained in workflow output can better inform test case selection for benchmarking and comparisons against expert-defined ground truth, as compared to conventional qualitative or random test case selections.

We use unsupervised machine learning techniques to let workflow data inform test case selection. These techniques can identify properties in large datasets by accounting for higher-dimensional associations difficult for individuals to quantify. In particular, data clustering formalizes natural groupings in workflow and algorithm output. Cluster centres and/or outlier cases may serve as candidate test cases. Principal components analysis (PCA) aids in data visualization, and supplemented with conventional statistics, confirms that clusters stratify the differences in algorithm output most relevant to the clinical application. Such an approach is hypothesized to better ensure that test cases are representative of clinical data and may reduce the number of test cases needed to establish robust benchmarking results. Our aim is to provide researchers with a general framework to identify which differences between algorithms are clinically relevant, and to identify representative examples of those differences for benchmarking and ground-truth-based assessments.

To demonstrate the proposed framework, we consider two independent and partially-automated workflows, each based on a different DIR algorithm. Each workflow estimates violations of original dosimetric planning objectives for head and neck cancer patients that may result from anatomical changes (e.g. weight loss, tumor shrinkage) occurring throughout the 6–7 weeks of curative-intent treatment. Estimated planning objective violations exceeding a given allowable margin indicate that the patient should be assessed for treatment replanning. Both algorithms have been validated in the literature in comparative assessments (Kumarasiri *et al* 2014) and independently against expert-based ground-truth (Cline *et al* 2015, Ramadaan *et al* 2015, Pukala *et al* 2016, Loi *et al* 2018) and are expected to produce similar geometric output. However, differences in the geometric output of DIR algorithms that are considered minor in such standalone evaluations (Brock *et al* 2017) may lead to clinically significant differences between workflows due to non-linear error propagation or workflow 'upweighting' of algorithm differences. This study shows that the clinically-relevant differences in algorithm output may be identified by clustering on pairwise differences in workflow output. Data points at cluster centers ('medoids') are considered for their suitability as test cases. Test cases may then be assessed against ground truth and workflow-specific algorithm performance requirements.

Materials and methods

Image data, algorithms and clinical workflow

We retrospectively assessed the image data from a cohort of 15 head and neck cancer patients. Each patient received curative-intent VMAT chemoradiotherapy (70 Gy in 33 fractions). Treatment plans were designed to meet institutional dosimetric planning objectives including: high dose clinical target volume (CTV) D99%, D95%, D2%; low dose CTV D99%, D95%, D20%; high dose planning target volume (PTV) D99%, D95%, D2%; low dose PTV D99%, D95%, D20%; brainstem Dmax; spinal cord Dmax; and ipsilateral and contralateral parotid gland D50% dose parameters (Weppler *et al* 2018). Planning was performed in the Eclipse Treatment Planning System, Version 11 (Varian Medical Systems, Palo Alta, CA) with the anisotropic analytical algorithm (AAA). Throughout treatment, each patient received a pre-treatment CT simulation (pCT) as well as approximately weekly on-unit cone-beam CT (CBCT) imaging, totaling 15 pCTs and 105 CBCTs for the cohort.

The two compared workflows ('workflow #1' and 'workflow #2') indicated the need for a replan assessment according to differences in planned and delivered doses. For each patient, two commercial DIR algorithms, B-spline-based Velocity™ Version 3.2.0 (Rueckert *et al* 1999, Lawson *et al* 2007) ('DIR#1') and demons-based SmartAdapt® (Thirion *et al* 1998, Wang *et al* 2005) ('DIR#2') (Varian Medical Systems, Palo Alto, CA), independently deformed copies of the pCT over the field of view of CBCTs. The DIR#1 'CBCT Corrected Deformable' multi-pass option provided the most flexibility in CT-to-CBCT deformation. The 'Structure-Guided' multi-pass option was used for a patient with bolus digitally rendered on the pCT and physical bolus applied during treatment. DIR#2 had only one deformation option available. After pCT to CBCT deformation, the resulting

'synthetic CTs' ('sCT#1' and 'sCT#2') had the original pCT HU calibration curve, larger pCT field of view, and CBCT-based anatomical changes. The original clinician-delineated target and organ-at-risk contours were propagated from the pCT to the sCTs according to DIR deformation vector fields. Several sCT#2's had obvious local artifacts in the external contour and these errors were corrected as part of a reasonable clinical workflow; sCT#2 CTVs and PTVs were cropped 3 mm from the corrected external contour as needed. It was uncommon to observe external contour artifacts in sCT#1 images and corrections were generally not required.

The original treatment plan was reapplied to synthetic CTs in the treatment planning system to estimate the dose delivered on the day of CBCT-acquisition. Instances where delivered doses violated scaled institutional planning objectives (total plan ÷ 33 fractions) were tabulated for each patient (p), fraction with CBCT-acquisition (f), and dose parameter (d), in addition to whether the violation ($v_{p,f,d}$) exceeded $\{\tau\}_{j=1}^{5} = \{1\%, 2\%, 3\%, 5\%, 10\%\}$ of the scaled planning objective (motivated by Weppler *et al* 2018):

$$\widetilde{w}_{p,f,d} = \underset{\{\tau_j |\ \tau_j \leqslant v_{p,f,d}\}}{\arg\min} (v_{p,f,d} - \tau_j). \tag{1}$$

Therefore, quantized differences, $\widetilde{w}_{p,f,d}^{sCT\#1} - \widetilde{w}_{p,f,d}^{sCT\#2}$, in workflow output emphasized large discrepancies in estimated violations (e.g. the change from 5% to 10% is likely more important than the change from 2% to 3%), according to sample cut-off values that may be used to assess replan need in practise.

Data clustering on workflow output

Differences in workflow output, $\widetilde{w}_{p,f,d}^{sCT\#1} - \widetilde{w}_{p,f,d}^{sCT\#2}$, were input for data clustering as a matrix with pf rows and d columns; each analyzed patient and fraction was assigned to a data cluster. Data was centered but unscaled to preserve the relative magnitude of $\widetilde{w}_{p,f,d}^{sCT\#1} - \widetilde{w}_{p,f,d}^{sCT\#2}$ differences among dose parameters. The Hopkins statistic (Lawson *et al* 1990) tested for the presence of natural data groupings in workflow output differences. k-Medoid clustering (PAM-k) (Kaufman *et al* 1990) heuristically assigned paired data points to k clusters where k was estimated by the Duda-Hart test (Duda *et al* 1973) and average silhouette width (Hennig *et al* 2013). Euclidean and Manhattan distances were tested as similarity metrics. PCA projected the results of the d-dimensional clustering into two dimensions for visualization and qualitative inference of factors contributing to cluster differences.

Assessing differences in algorithm output

We verified whether workflow clusters also grouped patients/fractions according to algorithm differences, $a_{p,f,g}^{sCT\#1} - a_{p,f,g}^{sCT\#2}$, where g denotes a geometric parameter directly assessing algorithm output. Differences in centers-of-mass (CoM) between sCT#1 and sCT#2 paired contours were calculated. In addition, we assessed differences in contour volumes and CTV-to-PTV planning margins (CTV-to-PTV mean distance to agreement). Differences in external body contour were measured on the CT axial image slice containing the high dose CTV CoM. All algorithm differences except CoM (in mm) were considered as a percentage difference relative to values at planning.

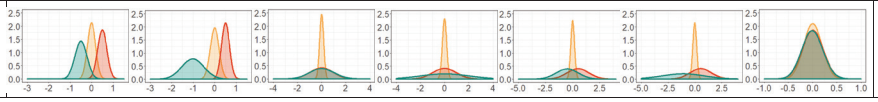
MANOVA and ANOVA statistical tests determined whether there were statistically significant differences ($p < 0.05$) in $a_{p,f}^{sCT\#1} - a_{p,f}^{sCT\#2}$ between clusters, and with respect to which geometric parameters. This ensured that workflow-based clusters did not simply result from random algorithm discrepancies. For ANOVA results with $0.05 < p < 0.10$ for a given parameter g, paired t-tests (accounting for heteroscedasticity) were performed between the cluster with the largest mean absolute difference and aggregate of the remaining clusters.

Correlation tests verified that statistically significant differences in g between clusters were relevant to the workflow output. Pearson correlation coefficient identified correlations between algorithm output parameters. Kendall's rank correlation coefficient assessed the correlation of aggregate $a_{p,f,g}^{sCT\#1} - a_{p,f,g}^{sCT\#2}$ differences with $\widetilde{w}_{p,f,d}^{sCT\#1} - \widetilde{w}_{p,f,d}^{sCT\#2}$ differences for each g and d. Benjamini–Hochberg p values adjustments managed the false discovery rate under multiple testing (Benjamini *et al* 1995).

Simulated data

In order to facilitate interpretation of clustering and PCA results for the more complex patient data, we created a simple simulation. We simulated the output of two algorithms, denoted as Algorithm A and Algorithm B, in seven independent Gaussian-distributed variables (analogous to the structure-specific dose parameter values) and produced 1000 samples of each. Details are provided in table 1. Systematic differences between Algorithms A and B were repeated in two variables where the simulated workflow upweighted one variable but not the other. We assessed: (1) the detectability of systematic and random differences in workflow output, (2) the detectability of the upweighted algorithm variables, (3) the qualitative characteristics of simulated clusters.

Table 1. Simulated algorithm and workflow data. Algorithm output in each variable was sampled from a Gaussian distribution with mean (μ) and standard deviation (σ) indicated. Each of the seven subsets of samples had a specific discrepancy. The simulated workflow up-weighted select algorithm discrepancies by a factor of two. Applying the proposed framework to the simulated data produced the cluster assignment in the right-most column. Yellow font: Algorithm A. Red font: Algorithm B. Blue font: Workflow Effect.

Effect	Systematic offset (× 1)	Systematic offset (× 2)	Difference in variability (× 1)	Difference in variability (× 2)	Systematic offset and difference in variability (× 1)	Systematic offset and difference in variability (× 2)	Random noise	
Samples ($n = 1000$)	v_1	v_2	v_3	v_4	v_5	v_6	v_7	Cluster assignment
1 to $\lfloor\frac{n}{7}\rfloor$	$\mu = 0, \sigma = 0.1$ $\mu = 0.5, \sigma = 0.1$ × 1	$\mu = 0, \sigma = 0.1$						a
$\lfloor\frac{n}{7}\rfloor$ to $\lfloor\frac{2n}{7}\rfloor$	$\mu = 0, \sigma = 0.1$	$\mu = 0, \sigma = 0.1$ $\mu = 0.5, \sigma = 0.1$ × 2	$\mu = 0, \sigma = 0.1$					b
$\lfloor\frac{2n}{7}\rfloor$ to $\lfloor\frac{3n}{7}\rfloor$	$\mu = 0, \sigma = 0.1$		$\mu = 0, \sigma = 0.1$ $\mu = 0, \sigma = 0.5$ × 1	$\mu = 0, \sigma = 0.1$				a
$\lfloor\frac{3n}{7}\rfloor$ to $\lfloor\frac{4n}{7}\rfloor$	$\mu = 0, \sigma = 0.1$			$\mu = 0, \sigma = 0.1$ $\mu = 0, \sigma = 0.5$ × 2	$\mu = 0, \sigma = 0.1$			a c d
$\lfloor\frac{4n}{7}\rfloor$ to $\lfloor\frac{5n}{7}\rfloor$	$\mu = 0, \sigma = 0.1$				$\mu = 0, \sigma = 0.1$ $\mu = 0.5, \sigma = 0.5$ × 1	$\mu = 0, \sigma = 0.1$		a e
$\lfloor\frac{5n}{7}\rfloor$ to $\lfloor\frac{6n}{7}\rfloor$	$\mu = 0, \sigma = 0.1$					$\mu = 0, \sigma = 0.1$ $\mu = 0.5, \sigma = 0.5$ × 2	$\mu = 0, \sigma = 0.1$	a f
$\lfloor\frac{6n}{7}\rfloor$ to n	$\mu = 0, \sigma = 0.1$						$\mu = 0, \sigma = 0.1$ $\mu = 0, \sigma = 0.1$ × 1	a
Histograms								

Framework overview

In summary, the above framework (figure 1): (1) produces data clusters based on differences in workflow output; (2) assesses algorithm differences among clusters; and (3) verifies that those algorithm differences are the most relevant to the workflow. The result is a process that clusters data based on the algorithm differences that are most relevant to the workflow output. Lastly, (4) data points at the centre of each cluster (cluster 'medoids') are considered as paired sCT#1 and sCT#2's representative of 'average' algorithm differences. These are assessed as candidate test cases. For general differences in algorithm output associated with each cluster, the mean values of test cases are compared with cluster averages.

All clustering, PCA, and statistical analyses were performed in R (R Version 3.5.1, The R Foundation for Statistical Computing, Vienna, Austria).

Results

Simulated data

The results of our simulation demonstrate how the framework detects algorithm discrepancies and workflow effects. The Hopkins statistic strongly indicated that simulated data was naturally clustered (Hopkins statistic = 0.13, sampling rates from 5% to 10%). Seven ground truth clusters were present (v_{1-7}), while PAM-k assigned data into six subgroups (clusters 'a'–'f'). Cluster assignment is summarized in table 1 and PCA projections are shown in figure 1.

Cluster shape and location corresponds to various systematic differences in algorithm output (figure 2). Qualitatively, as algorithm discrepancies or workflows produce larger systematic offsets in data (μ_i^A versus μ_i^B for some variable v_i), the cluster will move farther from the origin in a PCA projection. The number of clusters and corresponding test cases are also dependent on σ. As variability increases, clusters will generally elongate about the data origin and may separate to produce a band of smaller clusters. Excess noise in the algorithm and workflow data may lead to fewer clusters in general and medoids occurring nearer to mean noise values.

Variable v_6 had a systematic offset and difference in variability between Algorithms A and B and was upweighted in importance by the workflow. As expected, data clustering and PCA identified this discrepancy as having the greatest workflow effect, shown through the alignment of cluster 'f' and the v_6 loading and its strong association with PCA dimension 1. Qualitatively, this cluster was systematically offset from, but still contained, the data origin. However, when compared to ground truth, v_6 data about the origin was partially assigned to the 'random variation' cluster, as further examined below. Variable v_4 simulated Algorithm A versus Algorithm B differences in variability alone, with differences upweighted by the workflow. The framework produced two clusters symmetric about the origin and aligned with PCA dimension 2. This indicates that manual corrections may

Objective: Identify test cases for algorithm benchmarking. Test cases are chosen to be representative of discrepancies in algorithm output that correspond to the largest discrepancies in workflow output.

1.) Cluster data based on differences between workflow output, $w_{\#1} - w_{\#2}$.
2.) Assess whether workflow clusters also stratify $a_{\#1} - a_{\#2}$ differences between algorithm output using e.g., MANOVA, ANOVA, Levene's test
   a. If $a_{\#1} - a_{\#2}$ are significantly different between clusters, proceed to 3.
   b. If $a_{\#1} - a_{\#2}$ are not significantly different between clusters, increase sample size and consider whether p values decrease. If p values decrease, iteratively increase sample size until an acceptable significance is achieved and proceed to 3. If not, another workflow component may more directly contribute to workflow differences. Consider applying the framework to other components. Otherwise, clusters may result from random workflow discrepancies and random test cases can be selected by proceeding to step 4.
3.) Assess whether algorithm differences, $a_{\#1} - a_{\#2}$, are correlated with differences in workflow output, $w_{\#1} - w_{\#2}$, using e.g., Pearson, Spearman's ρ, or Kendall's τ correlation tests .
   a. If algorithm differences characterizing each cluster are significantly correlated with workflow differences, proceed to 4.
   b. If algorithm differences characterizing each cluster are not significantly correlated with workflow differences, associations may be too subtle to be significant according to conventional statistics or algorithm difference averages may be skewed by outliers. Assess correlation coefficients or measures of association and/or visually review results. If correlations are considered acceptable, proceed to 4. Otherwise, clusters may result from random workflow discrepancies; random test cases can be selected by proceeding to step 4.
4.) Extract representative cases from clusters as candidate test cases. Representatives may be cluster centres ("medoids") or outliers.

Figure 1. Summary of the proposed test case selection framework.

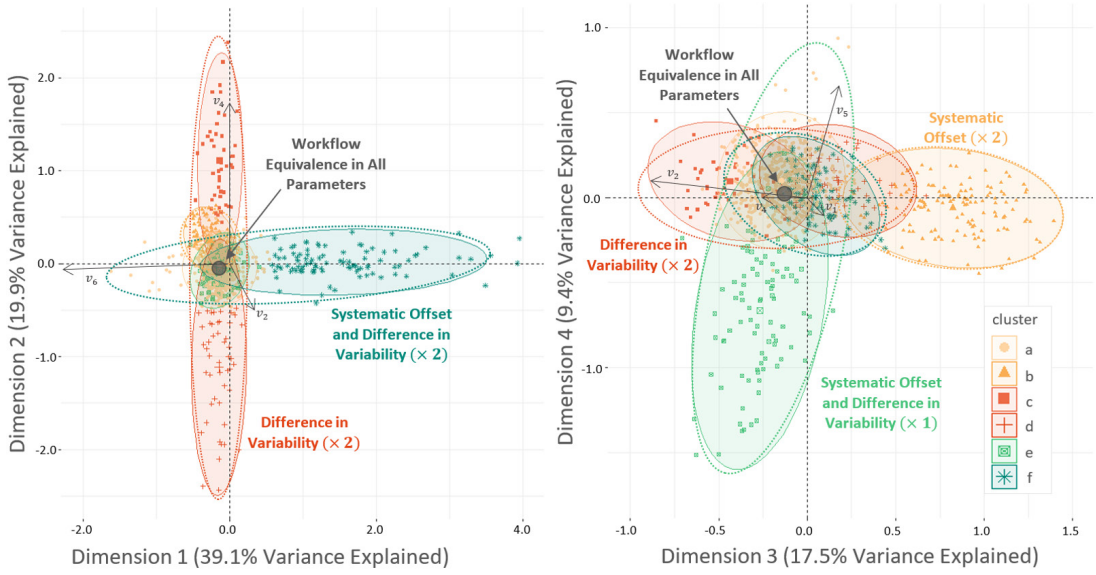


Figure 2. Data clustering (coloured ellipses with solid boundaries) and PCA results for the simulated algorithm and workflow. The first three principal component dimensions aligned with the algorithm differences upweighted by the workflow (v_6, v_4, and v_2) as indicated by the principal component loadings. More subtle algorithm differences were consolidated with the random noise effects (v_7) in cluster 'a'. Ground truth clusters are indicated by ellipses with dotted boundaries; ground truth clusters correspond to the discrepancies simulated in each row of table 1. For example, samples 715–857 in row 6 of table 1 have a ground truth cluster assignment of v_6, indicated by the dark green dotted line aligning with PCA dimension 1.

be required to consolidate clusters to better reproduce ground truth. However, the framework output provided improved sampling of the variability (medoids for clusters 'c' and 'd') compared to the ground truth medoid which aligned with that of the random grouping. The variable with the workflow-upweighted systematic offset (v_2) was shifted from the data origin, dominating PCA dimension 3.

Non-upweighted workflow effects, such as v_1 and v_3, and the random variable (v_7) were assigned exclusively to a random variation cluster (cluster 'a'), approximately centered about the original data origin. As data must be centered prior to clustering and PCA, there is some misalignment of the original data origin and the PCA origin.

All variables except that with the workflow-upweighted systematic offset (v_2) had some elements assigned to the random variation group. Clustering should be visually reviewed and adjusted if needed.

MANOVA confirmed that clusters stratified workflow differences ($p < 0.001$). ANOVA identified statistically significant differences in algorithm output according to cluster stratification for variables: v_1, v_5, and v_6 ($p < 0.05$). Weak statistical differences were indicated for v_2 and v_4. If differences in algorithm variability in a given parameter are indicated, such as those due to workflow upweighting, statistical tests may be supplemented by tests assessing differences in standard deviation (e.g. Levene's test). Weak associations may still be informative as statistical significance can be limited for variables where systematic algorithm discrepancies partially overlap with random noise, such as for v_2 and v_4. In general, statistical significance may also be limited by sample size.

Data clustering on workflow output

For the clinical data, the Hopkins statistic showed that workflow differences, $\widetilde{w}_{p,f,d}^{sCT\#1} - \widetilde{w}_{p,f,d}^{sCT\#2}$, had a naturally clustered structure. Statistic values were approximately 0.75 under various sampling rates from 5% to 10%. Therefore, we rejected the null hypothesis that the data did not have a natural clustering tendency with 90% confidence (Lawson *et al* 1990). PAM-k clustered the data into five groups; figure 3 shows the projection of clusters into the first two principal component dimensions. 84.3% of the data variance was explained by the first four principal components (30.1%, 22.3%, 18.9%, and 13.0%, respectively). Variance explained by the fifth principal component reduced to 3.8%. As a result, we reviewed projections into the first four principal component dimensions. Qualitatively, cluster 1 contained the original $\widetilde{w}_{p,f,d}^{sCT\#1} - \widetilde{w}_{p,f,d}^{sCT\#2}$ data origin, corresponding to identical workflow output in all parameters, and consisted of relatively small discrepancies between workflows. All other clusters were offset from the origin, indicating potentially significant systematic offset discrepancies in workflow output as indicated by the data simulation. Based on PCA, inferences about workflow differences defining each systematic cluster are indicated in figure 3. These hypotheses were confirmed by statistical analysis below.

Table 2 summarizes the differences in workflow output among the clusters. For patients/fractions assigned to cluster 2, mean and median workflow #2 estimates of ipsilateral parotid gland D50% were much larger than for workflow #1. For cluster 3, mean and median low dose target coverage violations estimated by workflow #2 were larger than workflow #1. Mean high dose CTV hotspot, brainstem and spinal cord violations estimated by workflow #2 were larger for cluster 4 data. In contrast, for cluster 5, mean and median estimated target coverage violations were small for workflow #2 relative to that of workflow #1. Median workflow differences for cluster 1 were identically zero across all parameters. These results are consistent with the qualitative indications from PCA.

Various modifications of the analysis were tested to assess framework robustness. As closed-form sample size calculations do not generally exist in the machine learning setting, varying the input sample size provided a measure of cluster stability. Ninety-six or more randomly selected samples gave comparable clustering and PCA. Clusters 2, 3, and 4 demonstrated some variability so that method stability may further improve by increasing sample size. Although workflows were partially automated, manual image import/export, basic contour corrections and others imposed sample size limitations. However, it is helpful to know which parameters may vary as a result of practical constraints on sample size. When using Euclidean and Manhattan distances, cluster assignment was identical for 94 of the 105 data points. Variations occurred in the definition of clusters 2 and 4, which were adjacent and overlapping in projections onto PCA dimensions 1 and 2. PAM-k with Manhattan distance introduced a single-datapoint outlier cluster. For the remainder of this analysis, we elected to use the results obtained with the more conventional Euclidean distance.

Assessing differences in algorithm output

Differences in algorithm output, $a_{p,f,g}^{sCT\#1} - a_{p,f,g}^{sCT\#2}$, for patients/fractions within each cluster are included in table 3. MANOVA confirmed that algorithm output was significantly different between clusters ($p < 0.001$), with ANOVA identifying significant differences in estimations of: target, spinal cord and parotid gland volumes; and high dose CTV-to-PTV margins. *T*-tests found statistically significant differences between cluster 5 and the aggregation of clusters 1–4 for DIR#1 and DIR#2 low dose CTV-to-PTV margin estimates ($p = 0.009$); external contour estimates showed a larger discrepancy for cluster 4 than the remaining pooled data ($p = 0.022$). Cluster 1 variability was largest for spinal cord CoM differences with $p = 0.017$, again based on *t*-test results, possibly due to differences in algorithm processing of minor CT-to-CBCT misalignments of bony anatomy. All other algorithm differences in cluster 1 images were relatively small; we considered cluster 1 to be analogous to cluster 'a' in the simulation.

Kendall's τ test identified significant correlations ($p < 0.05$ after Benjamini–Hochberg adjustment) in $\widetilde{w}_{p,f,d}^{sCT\#1} - \widetilde{w}_{p,f,d}^{sCT\#2}$ for ipsilateral parotid gland D50% and $a_{p,f,g}^{sCT\#1} - a_{p,f,g}^{sCT\#2}$ for target, parotid gland, and external contour volumes. These associations most clearly coincided with the above observations for cluster 2. High dose target hotspot violation differences correlated with CTV CoM, coinciding with observations for cluster 4

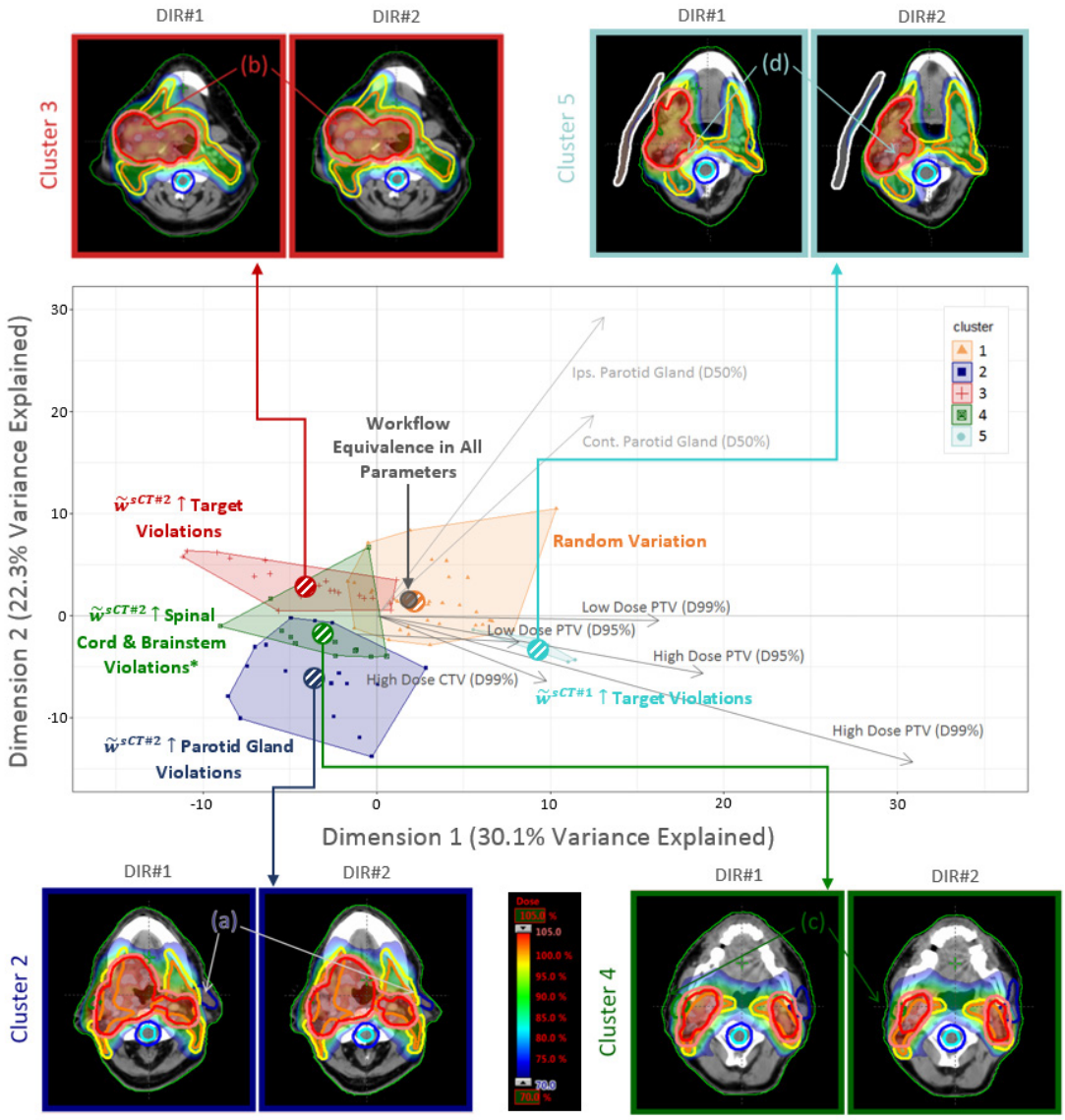


Figure 3. Data clustering of patients/fractions based on differences in the output between two workflows ($\widetilde{w}^{sCT\#1}$, $\widetilde{w}^{sCT\#2}$). PCA indicates the presence of one random grouping (cluster 1) and four kinds of systematic discrepancies (clusters 2–5) between the two workflows. Four candidate test cases (cluster 2–5 'medoids') are shown. Contours include: high dose CTV (red); high dose PTV (pink); low dose CTV (orange); low dose PTV (yellow); spinal cord (cyan); spinal cord plus margin (blue); left parotid gland (dark blue); right parotid gland (green); and bolus (white). In general, workflow-relevant algorithm discrepancies occurred in the processing of: (a) parotid glands, (b) high dose CTV, (c) external contour, (d) CTV-to-PTV margin. *Based on projection of clusters onto dimensions 1 and 3 (not shown). Reproduced with permission from COMP65 Annual Scientific Meeting: Young Investigators Symposium - 'Artificial Intelligence to Advance Adaptive Radiation Therapy' by S Weppler *et al.*

(although CoM differences were not found to be statistically significantly different by ANOVA). High dose coverage was correlated with high dose CTV-to-PTV margin, as captured by cluster 5. No significant correlations were identified to confirm that $a_{p,f,g}^{sCT\#1} - a_{p,f,g}^{sCT\#2}$ differences for clusters 1 and 3 were relevant to workflow output. This is expected for a random variations between DIR#1 and DIR#2 output (cluster 1). We further consider cluster 3 in the Discussion.

Often, Kendall's τ tests found significant correlations with multiple target parameters, such as low dose CTV and low dose PTV volume changes, not found by clustering. Pearson correlation coefficient confirmed that for high dose and low dose targets, statistically significant correlations existed between CTV and PTV volume and CoM discrepancies; the single exception was for low dose PTV volume and low dose PTV CoM with $p = 0.156$. Discrepancies in algorithm output for CTV volume and CoM were weakly correlated, all others were moderately to highly correlated (Pearson correlation coefficient of 0.54–0.85).

Cluster medoids were considered as candidate test cases. Axial CT slices of candidate test cases showing the largest discrepancies in algorithm output are included in figure 3 for clusters 2–5. For cluster 2, differences in DIR#1 and DIR#2 estimations of ipsilateral and contralateral parotid gland volumes were 13.2% and 13.5% for the medoid, similar to the cluster averages of 13.7% and 12.3%, respectively. Cluster 4's differences in estimated

Table 2. Cluster-specific differences in $\widetilde{w}_{p,f,d}^{sCT\#1}$ and $\widetilde{w}_{p,f,d}^{sCT\#2}$ expressed as median (mean).

	Cluster 1 (n = 49)	Cluster 2 (n = 17)	Cluster 3 (n = 21)	Cluster 4 (n = 12)	Cluster 5 (n = 6)
High dose CTV D95%	0% (0.1%)	0% (0.8%)	0% (0.0%)	0% (−0.1%)	**2% (2.5%)**
High dose PTV D95%	0% (0.5%)	−1% (−0.8%)	−2% (−2.7%)	0% (−1.0%)	**2.5% (2.5%)**
High dose CTV D99%	0% (0.3%)	0% (1.1%)	0% (0.3%)	0% (0.1%)	**8% (7.5%)**
High dose PTV D99%	0% (0%)	−1% (−2.1%)	−3% (−4.1%)	−0.5% (−2.4%)	**7.5% (7.5%)**
Low dose CTV D95%	0% (0.1%)	**0% (0.4%)**	0% (0.1%)	0% (0.3%)	0% (0.2%)
Low dose PTV D95%	0% (0.2%)	−1% (−1.2%)	**−2% (−2.2%)**	−1% (−1.1%)	0% (−0.7%)
Low dose CTV D99%	0% (0.4%)	0% (0.4%)	0% (0.1%)	0% (0.2%)	**−2% (−0.8%)**
Low dose PTV D99%	0% (0.2%)	**−2% (−2.1%)**	**−5% (−5.3%)**	−0.5% (−1.6%)	−2% (−1.2%)
High dose CTV D2%	0% (−0.1%)	0% (−0.2%)	**0% (−0.4%)**	**0% (−0.4%)**	0% (−0.2%)
High dose PTV D2%	0% (0%)	0% (0%)	0% (0%)	0% (0%)	0% (0%)
Low dose CTV D20%	0% (−0.3%)	**0 (−0.5%)**	0% (−0.3%)	0%	0% (0%)
Low dose PTV D20%	0% (−0.1%)	**−1% (−0.8%)**	0% (0%)	0% (−0.4%)	0% (0%)
Brainstem Dmax	0% (−0.3%)	0% (0%)	0% (0.1%)	**0% (−2.5%)**	0% (0%)
Spinal cord Dmax	0% (−0.2%)	0% (0.3%)	0% (−0.1%)	**0% (−0.9%)**	0% (−0.3%)
Ips. parotid gland D50%	0% (0.2%)	**−10% (−8.8%)**	0% (0%)	0% (0.8%)	0% (0%)
Cont. parotid gland D50%	0% (0.2%)	0% (−2.9%)	0% (−0.2%)	**−10% (−9.4%)**	0% (0%)

Clusters with the largest and second largest absolute median differences between $\widetilde{w}_{p,f,d}^{sCT\#1}$ and $\widetilde{w}_{p,f,d}^{sCT\#2}$ are highlighted with dark and light colors, respectively (ties are broken according to absolute mean differences).

external contours was 5.7% for both the medoid and cluster average. Cluster 5's high dose CTV-to-PTV margin discrepancy was −15.8% for the medoid and an average of −14.6% for the remainder of the cluster. No clear correspondence between cluster medoid and cluster averages were observed for clusters 1 and 3, as expected for cluster 1. However, including cases from clusters with small discrepancies in workflow output (e.g. Cluster 1) is important as although these cases have similar workflow output, this output may differ from ground truth.

Figure 4 shows the assignment of patient/fraction data to each of the five clusters. Cluster 5 consisted of data exclusively from patient 1, who was the only patient analysed with DIR#2's 'structure-guided' setting. Only three patients were exclusively assigned to one cluster (patient 9 to cluster 1, patients 10 and 12 to cluster 3). Given that some patients were imaged for the first three fractions and then weekly, medoids were generally from the first half of treatment.

Discussion

Test selection and prioritization techniques are well developed in select fields, such as computer science, yet are uniquely tailored to those applications (e.g. Hao *et al* 2016). To our knowledge, the proposed framework provides the first general approach for test case selection in medical physics. As demonstrated for the simulated and DIR-based clinical examples, the framework may be used to identify workflow-relevant differences between algorithms by clustering discrepancies in downstream workflow output. Cluster centres provided representative test cases for further algorithm benchmarking and characterization. Such an approach may offer important improvements on alternative benchmarking approaches that unevenly or randomly sample from the inherent data clusters. For example, to compare the approach with current standards for DIR algorithm benchmarking and characterization, conventional test case selection may assess end-of-treatment images from sequential patients. For our dataset, the first 8 patients of our sequence were needed to get an example from each cluster. Without cluster analysis, stopping criteria on qualitative selection would be unknown. Alternatively, random selection of five images has a 1.3% probability that at least one image is included from each cluster, based on the hypergeometric distribution. At least 20 random cases are required for a probability of 75%. The proposed approach has the potential to select test cases that better represent algorithm and workflow variability, while reducing the number of cases requiring expert review.

Table 3. Geometric differences, $a_{p,f,g}^{sCT\#1} - a_{p,f,g}^{sCT\#2}$, between DIR#1 and DIR#2 algorithm output expressed as mean (standard deviation) for each cluster.

	Cluster 1 (*n* = 49)	Cluster 2 (*n* = 17)	Cluster 3 (*n* = 21)	Cluster 4 (*n* = 12)	Cluster 5 (*n* = 6)	*P* value
High dose CTV ΔVolume	5.2% (6.1%)	7.9% (5.0%)	5.4% (5.3%)	**8.8% (5.7%)**	8.6% (1.1%)	0.145
High dose CTV ΔCoM	2.16 mm (1.28 mm)	2.42mm (0.96mm)	**2.83 mm (1.88 mm)**	2.26 mm (0.80 mm)	2.49 mm (0.52 mm)	0.413
High dose PTV ΔVolume	3.8% (4.5%)	6.5% (4.4%)	3.7% (3.4%)	6.8% (4.4%)	**7.3% (1.8%)**	0.021
High dose PTV ΔCoM	1.89 mm (1.10 mm)	2.32 mm (1.05 mm)	**2.54 mm (1.63 mm)**	2.20 mm (0.89 mm)	2.37 mm (0.54 mm)	0.270
Low dose CTV ΔVolume	4.4% (5.1%)	7.2% (3.5%)	3.4% (3.8%)	**8.6% (3.4%)**	6.3% (2.4%)	0.003
Low dose CTV ΔCoM	2.12 mm (1.17 mm)	**2.70 mm (1.41 mm)**	2.15 mm (0.77 mm)	2.05 mm (0.61 mm)	2.01 mm (0.98 mm)	0.367
Low dose PTV ΔVolume	2.9% (3.7%)	**6.5% (4.4%)**	3.2% (3.2%)	5.5% (2.6%)	4.6% (1.3%)	0.019
Low dose PTV ΔCoM	1.82 mm (0.87 mm)	**2.29 mm (1.40 mm)**	1.88 mm (0.78 mm)	1.79 mm (0.63 mm)	1.37 mm (0.30 mm)	0.248
High dose CTV-to-PTV margin	-1.7% (3.5%)	-2.8% (4.7%)	-0.6% (2.8%)	-1.1% (4.4%)	**-14.8% (3.8%)**	<0.001
Low dose CTV-to-PTV margin	-0.7% (3.1%)	-2.3% (4.8%)	-1.1% (4.0%)	0.5% (2.8%)	**-3.7% (1.7%)**	0.093
Brainstem ΔVolume	1.0% (4.0%)	**2.1% (3.1%)**	0.9% (4.3%)	1.4% (4.4%)	-0.4% (2.8%)	0.734
Brainstem ΔCoM	0.97 mm (0.58 mm)	0.93 mm (0.61 mm)	1.09 mm (0.55 mm)	**1.13 mm (0.69 mm)**	0.80 mm (0.28 mm)	0.700
Spinal cord ΔVolume	-4.7% (8.5%)	-1.3% (6.3%)	**-10.2% (9.1%)**	-3.5% (5.7%)	-1.1% (3.6%)	0.008
Spinal cord ΔCoM	**4.16 mm (2.43 mm)**	3.82 mm (2.21 mm)	2.94 mm (2.03 mm)	2.84 mm (1.08 mm)	2.27 mm (1.25 mm)	0.063
Ips. parotid Gland ΔVolume	4.8% (5.9%)	**13.7% (13.0%)**	4.2% (4.8%)	8.9% (6.9%)	2.5% (6.8%)	< 0.001
Cont. parotid Gland ΔVolume	6.6% (6.8%)	**12.3% (12.2%)**	7.2% (6.1%)	5.3% (4.7%)	-2.4% (4.3%)	0.002
Ips. parotid gland ΔCoM	2.32 mm (2.97 mm)	**3.05 mm (1.78 mm)**	2.34 mm (1.03 mm)	2.44 mm (1.35 mm)	2.06 mm (0.88 mm)	0.813
Cont. parotid Gland ΔCoM	1.96 mm (1.22 mm)	2.52 mm (1.33 mm)	2.42 mm (1.28 mm)	**2.89 mm (1.09 mm)**	2.10 mm (1.20 mm)	0.131
Change in external contour	3.4% (3.9%)	5.2% (3.0%)	3.3% (1.9%)	**5.7% (2.3%)**	4.5% (1.0%)	0.077

Clusters with the largest and second largest mean differences between $a_{p,f,g}^{sCT\#1}$ and $a_{p,f,g}^{sCT\#2}$ are highlighted with dark and light colors, respectively. *P* values < 0.05 (ANOVA statistical tests) indicate that statistically significant differences existed between clusters 1–5 with respect to the corresponding geometric parameter.

K-medoid clustering was the selected clustering method as it is a robust alternative to conventional k-means clustering (Kaufman *et al* 1990) and defines clusters relative to a representative data point, rather than a cluster average of datapoints as for k-means. This reduces the effect of data noise during clustering and provides candidate test cases. Hierarchical clustering is another alternative but may be affected by erroneous merges or splits during clustering (Han *et al* 2012). Density-based methods may filter out outlier cases (Han *et al* 2012) which are of potential value in cluster assignment and possible outlier analysis. The medoids in our framework are pairs of datapoints 'typical' of each cluster, while outliers can also be identified. Our work identified the single outlier patient where DIR#1 used the structure-guided setting as the planning CT had digital bolus and CBCT had physical bolus (cluster 5 in figures 3, 4 and tables 2, 3). Approximately 5%–10% of patients are treated with bolus at our centre; input data for test case selection should be as representative as possible of future clinical workflow inputs.

While correlations between algorithm and workflow discrepancies were clear for clusters 2, 4, and 5 in the patient data example, the statistical significance of these correlations for clusters 1 and 3 could not be confirmed. This was expected for cluster 1 as it contained relatively small discrepancies between DIR algorithms, similar to the 'random' cluster in the simulated data. A distinct type of algorithm discrepancy may not have been observed for cluster 3 as PCA showed cluster 3 differences in workflow output were intermediate to clusters 2 and 4. The heuristic nature of data clustering may also have identified weak associations that were not clear enough to be of statistical significance. While conventional statistics provide a deterministic result, the machine learning

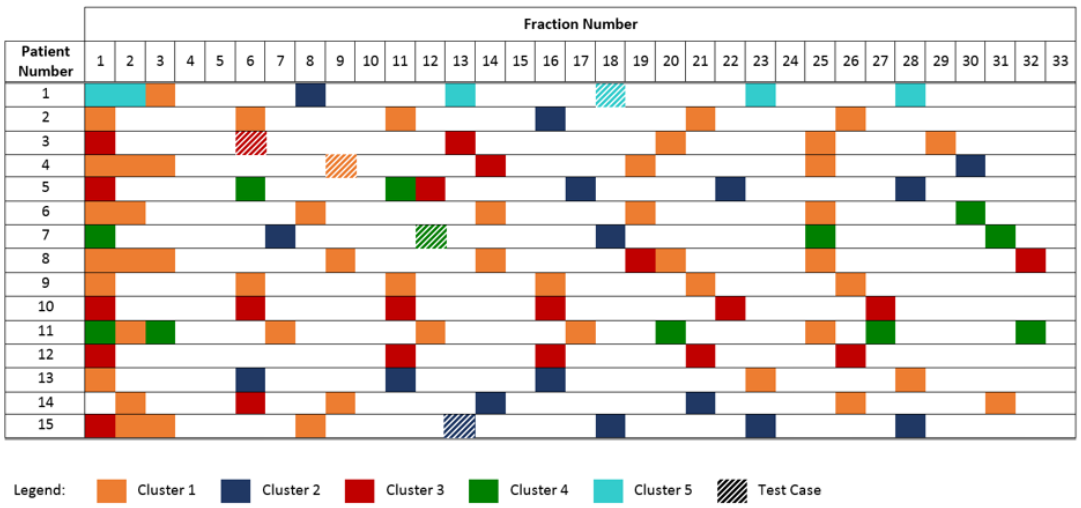


Figure 4. Cluster assignment for each analyzed fraction. In general, patients were not exclusively assigned to one cluster. As identified via the framework, test cases representing discrepancies in DIR#1 and DIR#2 algorithm output that were found most relevant to the clinical application generally occurred mid-course.

framework is intended to guide our interpretation of the data. Data simulations can provide valuable insight into the nature of data clusters and algorithm/workflow discrepancies.

PTV coverage is expected to degrade with daily patient setup uncertainties but was included in the analysis to improve the sensitivity of the framework to algorithm differences. Although not conventionally assessed for DIR benchmarking, differences in DIR#1 and DIR#2 deformations of CTV-to-PTV margins may be used to infer algorithm trade-offs in image similarity and regularization. As CTV-to-PTV margins were generally smaller for sCT#1s, we hypothesize that DIR#1 applied a larger emphasis on regularization; we note that local tumor shrinkage effects are 'blurred' into these planning margins. Differences in CTV-to-PTV margins in the context of the dose violation estimates may be of value for clinicians when deciding to monitor PTV dose objective violations or CTV dose deviations relative to planned. Post-DIR scripts may be considered necessary to automatically re-contour PTV margins (e.g. expand CTV).

Five patients had CBCT images acquired for the first three fractions then weekly, likely contributing to the occurrence of medoids in the first half of treatment. In theory, one could weight certain fractions more heavily if they are anticipated to be more important in the workflow. For example, correct replan indications are more valuable early in treatment when treatment replanning can lead to the greatest improvements in target coverage and healthy tissue sparing.

For the patient demonstration, all images were analyzed in aggregate even though some correlation of algorithm differences could be expected for each patient. Algorithm details are proprietary so it was unclear whether algorithms would process images differently primarily as a result of patient-specific factors (e.g. image quality, proximity of volumes to high contrast regions such as bone) or temporal changes (e.g. extent of weightloss and shrinkage effects). This uncertainty motivated the analysis of these variables in aggregate. More refined multi-level approaches may be used in practise if dominant effects are known beforehand.

Some DIR implementations are capable of accumulating doses delivered over multiple fractions. Although dose accumulation capabilities were not available for DIR#2 at our centre, the proposed framework may also be applied for workflows based on dose accumulation or dose warping. However, if the nature of algorithm discrepancies are not yet known, it may be beneficial to first assess dose estimates individually. If algorithm differences are patient-specific, dose accumulation may clarify clusters; if dose estimates for a given patient exhibit multiple types of algorithm discrepancies (as observed for the patient demonstration, figure 4), accumulation may have the potential to confound results. For the latter case, we would suggest an iterative approach to assess how nested processes within an algorithm correlate with workflow output discrepancies (figure 1, part 2.) b).

The proposed clustering/PCA/statistical analysis framework is generalizable to a variety of settings. Although it was outside of the scope of this study, analyzing the sequences of cluster assignment may indicate trends in algorithm differences relevant for benchmarking. If correlations between algorithm and workflow variables are unclear, data clustering and PCA can be applied to a combined dataset of workflow and algorithm parameters. While we considered cluster medoids as candidate test cases, algorithm assessments could also be performed for cluster outliers (e.g. on the convex hull of clusters). As PAM-k is a heuristic approach to clustering, future work may examine the selection of test cases according to ideas of mathematical optimality depending on the size of input datasets. While PAM-k provided a suitable number of clusters, k, if constraints on the number of test cases

exist, medoid clustering may be performed for a pre-specified number of clusters. If no clustering tendency is present for a dataset, it may indicate that algorithms are consistent in terms of workflow output or that systematic differences are dominated by random effects. Clustering on random data will still produce a 'Voronoi tessellation' (Hastie *et al* 2009) and partition workflow output for testing. To benchmark more than two algorithms, the framework may be applied to all possible pairs of algorithms. Similarities or differences in workflow clusters and test cases for these pair-wise comparisons can be used to establish a comprehensive test set. These sets could be disseminated with ground truth references to be used by various institutions for algorithm assessments and trial credentialing.

Conclusions

This study proposes a method for selecting test cases for algorithm benchmarking and characterization based on the differences between algorithms that most affect workflow output. This approach is best applied in situations where algorithm output is easy to obtain, but establishing ground truth references is resource-intensive. We compared two workflows, each based on a different DIR algorithm, to estimate the magnitude of during-treatment planning objective violations. Clustering workflow differences on estimated dose violations produced a stratification of algorithm output. Each cluster exhibited a different type of algorithm discrepancy: differences in estimated parotid gland volume; external contour deformation; or CTV-to-PTV margin. Cluster medoids were considered as test cases and found to be representative of algorithm differences defining the clusters.

Acknowledgments

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada—Canada Graduate Scholarship (CGS-D) to SW.

Disclosure of conflicts of interest

The authors have no conflicts of interest to disclose.

References

Benjamini Y and Hochberg Y 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing *J. R. Stat. Soc.* B **57** 289–300

Brock K K, Mutic S, McNutt T R, Li H and Kessler M L 2017 Use of image registration and fusion algorithms and techniques in radiotherapy: report of the AAPM Radiation Therapy Committee Task Group No. 132 *Med. Phys.* **44** e43–76

Cline K *et al* 2015 SU-E-J-89: comparative analysis of MIM and velocity's image deformation algorithm using simulated kV-CBCT images for quality assurance *Med. Phys.* **42** 3282–52

Court L E *et al* 2010 Use of a realistic breathing lung phantom to evaluate dose delivery errors *Med. Phys.* **37** 5850–7

Duda R O and Hart P E 1973 *Pattern Classification and Scene Analysis* (New York: Wiley)

Ehler E D, Barney B M, Higgins P D and Dusenbery K E 2014 Patient specific 3D printed phantom for IMRT quality assurance *Phys. Med. Biol.* **59** 5763–73

Fragoso M *et al* 2010 Dosimetric verification and clinical evaluation of a new commercially available Monte Carlo-based dose algorithm for application in stereotactic body radiation therapy (SBRT) treatment planning *Phys. Med. Biol.* **55** 4445–64

Han J, Kamber M and Pei J 2012 *Data Mining: Concepts and Techniques* (Waltham, MA: Morgan Kaufmann Publishers)

Hao D, Zhang L and Mei H 2016 Test-case prioritization: achievements and challenges *Frontiers Comput. Sci.* **10** 769–77

Hastie T, Tibshirani R and Friedman J 2009 *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* 2nd edn (New York: Springer)

Hennig C and Liao T 2013 How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification *J. R. Stat. Soc.* C **62** 309–69

Hoffmann C *et al* 2014 Accuracy quantification of a deformable image registration tool applied in a clinical setting *J. Appl. Clin. Med. Phys.* **15** 237–45

Kashani R *et al* 2008 Objective assessment of deformable image registration in radiotherapy: a multi-institutional study *Med. Phys.* **35** 5944–53

Kaufman L and Rousseeuw P J 1990 *Finding Groups in Data: an Introduction to Cluster Analysis* (New York: Wiley)

Kumarasiri A *et al* 2014 Deformable image registration based automatic CT-to-CT contour propagation for head and neck adaptive radiotherapy in the routine clinical setting *Med. Phys.* **41** 121712

Lawson J D, Schreibmann E, Jani A B and Fox T 2007 Quantitative evaluation of a cone-beam computed tomography-planning computed tomography deformable image registration method for adaptive radiation therapy *J. Appl. Clin. Med. Phys.* **8** 96–113

Lawson R G and Jurs P C 1990 New index for clustering tendency and its application to chemical problems *J. Chem. Inf. Comput. Sci.* **30** 36–41

Loi G *et al* 2018 Performance of commercially available deformable image registration platforms for contour propagation using patient-based computational phantoms: a multi-institutional study *Med. Phys.* **45** 748–57

Mencarelli A *et al* 2014 Deformable image registration for adaptive radiation therapy of head and neck cancer: accuracy and precision in the presence of tumor changes *Int. J. Radiat. Oncol. Biol. Phys.* **90** 680–7

Pukala J *et al* 2016 Benchmarking of five commercial deformable image registration algorithms for head and neck patients *J. Appl. Clin. Med. Phys.* **17** 25–40

Ramadaan I S et al 2015 Validation of Varian's SmartAdapt deformable image registration algorithm for clinical application *Radiat. Oncol.* **10** 73

Rueckert D, Sonoda L I, Hayes C, Hill D L G, Leach M O and Hawkes D J 1999 Nonrigid registration using free-form deformations: application to breast MR images *IEEE Trans. Med. Imaging* **18** 712–21

Thirion J P 1998 Image matching as a diffusion process: an analogy with Maxwell's demons *Med. Image Anal.* **2** 243–60

Wang H *et al* 2005 Validation of an accelerated 'demons' algorithm for deformable image registration in radiation therapy *Phys. Med. Biol.* **50** 2887–905

Weppler S, Quon H, Banerjee R, Schinkel C and Smith W 2018 Framework for the quantitative assessment of adaptive radiation therapy protocols *J. Appl. Clin. Med. Phys.* **19** 26–34