Journal of Medical Radiation Sciences

Open Access

ORIGINAL ARTICLE

Evaluation of a deformable image registration quality assurance tool for head and neck cancer patients

Molly Mee, BRadTherapy(Hons),[1] Kate Stewart, BAppSc(MRT-RT), MPH,[1,2] Marika Lathouras, BAppSc (MRT-RT),[2] Helen Truong, BRadTherapy, [2] & Catriona Hargrave, BAppSc(MRT-RT), MAppSc(Research), PhD[1,3]

[1]Faculty of Health, School of Clinical Sciences, Queensland University of Technology, Brisbane, Queensland, Australia
[2]Department of Radiation Oncology, Royal Brisbane and Women's Hospital, Metro North Health Service District, Herston, Queensland, Australia
[3]Radiation Oncology PAH – Raymond Terrace Campus, Division of Cancer Services, Metro South Health Service District, South Brisbane, Queensland, Australia

Keywords
deformable image registration, quality assurance, radiation therapy, treatment planning

Correspondence
Molly Mee, Department of Radiation Oncology, Royal Brisbane and Women's Hospital, Metro North Health Service District, Herston, QLD, Australia 4029. Tel: +61 413 157 912. E-mail: molly.mee@health.qld.gov.au

Received: 2 April 2020; Revised: 11 August 2020; Accepted: 12 August 2020

J Med Radiat Sci **67** (2020) 284–293

doi: 10.1002/jmrs.428

Abstract

Introduction: A challenge in implementing deformable image registration (DIR) in radiation therapy planning is effectively communicating registration accuracy to the radiation oncologist. This study aimed to evaluate the MIM® quality assurance (QA) tool for rating DIR accuracy. **Methods:** Retrospective DIR was performed on CT images for 35 head and neck cancer patients. The QA tool was used to rate DIR accuracy as good, fair or bad. Thirty registered patient images were assessed independently by three RTs and a further five patients assessed by five RTs. Ratings were evaluated by comparison of Hausdorff Distance (HD), Mean Distance to Agreement (MDA), Dice Similarity Coefficients (DSC) and Jacobian determinants for parotid and mandible subregions on the two CTs post-DIR. Inter-operator reliability was assessed using Krippendorff's alpha coefficient (KALPA). Rating time and volume measures for each rating were also calculated. **Results:** Quantitative metrics calculated for most anatomical subregions reflected the expected trend by registration accuracy, with good obtaining the most ideal values on average (HD = 7.50 ± 3.18, MDA = 0.64 ± 0.47, DSC = 0.90 ± 0.07, Jacobian = 0.95 ± 0.06). Highest inter-operator reliability was observed for good ratings and within the parotids (KALPA 0.66–0.93), whilst ratings varied the most in regions of dental artefact. Overall, average rating time was 33 minutes and the least commonly applied rating by volume was fair. **Conclusion:** Results from qualitative and quantitative data, operator rating differences and rating time suggest highlighting only bad regions of DIR accuracy and implementing clinical guidelines and RT training for consistent and efficient use of the QA tool.

Introduction

Image registration is a key process widely implemented in radiation therapy treatment planning as it improves the accuracy of tumour volume and organ at risk (OAR) delineation.[1] In Australia, radiation therapists (RTs) most commonly register diagnostic images with planning CT images for radiation oncologists (ROs) to use in the target and OAR delineation process. Diagnostically acquired MRI or PET imaging is utilised for planning in 71% of Australian radiation therapy departments and are increasingly being used to provide functional information about the patient's cancer.[2,3] When registering these diagnostic images with planning CT images, issues present when differences in patient positioning and anatomy are evident, causing anatomy to not align correctly after the registration is complete.

Until recently, rigid image registration (RIR) has been the most widely implemented technique in the clinical setting, which has limitations when registering images

with positional and anatomical differences. Deformable image registration (DIR) is increasingly being implemented to correct for these differences, by transforming individual voxels by differing magnitudes from one image to align with those in another.[4] DIR is of particular value in the head and neck region as differences between the diagnostic CT and planning CT often result from changes in neck flexion, shoulder position, inclusion of immobilisation devices, changes in patient size and tumour growth.[5,6] Although DIR can correct for these differences, this method is more difficult to visually assess for accuracy after registration than RIR, creating challenges around communicating registration accuracy efficiently. Identifying a robust method for evaluating the accuracy of the deformation would enhance appropriate clinical decision-making, particularly when delineating tumour and OAR volumes in planning with the aim of avoiding tumour and OAR localisation errors.

When evaluating clinically acceptable and accurate regions of alignment after DIR for tumour delineation purposes, the American Association of Physicists in Medicine (AAPM) TG132 Report[7] recommends a qualitative and quantitative approach. This study aims to evaluate a quality assurance (QA) traffic light tool in the MIM Maestro®version 6.7.6 (MIM Software, Ohio, USA) (MIM6.7.6) that allows RTs to rate subregions of deformed images with respect to good, fair or bad registration accuracy. The study's primary aim was to compare the quantitative metrics of the commonly used Hausdorff Distance (HD) metric, as well as the Mean Distance to Agreement (MDA), Dice Similarity Coefficient (DSC) and Jacobian determinants for structures as recommended in the AAPM TG132 Report with corresponding overlapping regions of good, fair or bad qualitative ratings of DIR accuracy.[7] The secondary aims of the study were to 1) assess the level of inter-operator agreement of qualitative ratings when using this tool to evaluate DIR accuracy and 2) to assess the time taken to perform the ratings and the volume of each rating applied to the deformed images. This study assessed the potential efficacy of this QA tool prior to its clinical implementation at the Royal Brisbane and Women's Hospital (RBWH) Department of Radiation Oncology.

Methods

Patient data

Ethics approval was granted (HREC no. LNR2018QRBW4700) to retrospectively access the data of 35 patients who had completed their radiation therapy course at the RBWH between 1 July 2018 and 31 December 2018. All patients were >18 years and had carcinoma in the base of tongue or tonsillar fossae. Each patient had a diagnostic CT scan, acquired for staging purposes where RT treatment positioning was not considered, as well as their planning CT scan imported into MIM6.7.6. This patient data were anonymised and allocated a patient ID between 1 and 35. The planning target volume (PTV) was contoured by the RO and three anatomical structures, the left and right parotids and mandible, were contoured by RT A on both datasets. An initial RIR was performed followed by DIR on MIM6.7.6 using an automatic workflow without any user refinement. This differed from the clinical workflow where user refinement is used after both the RIR and DIR processes, thus facilitating a range of registration accuracy levels across the patient cases included in the study.

Qualitative ratings

Five RTs with varying levels of radiotherapy experience rated the accuracy of the deformed images using the new MIM6.7.6 QA traffic light tool as indicated in Table 1.

Using the Reg Reveal® tool in MIM6.7.6, each RT rated the deformed images using red, yellow and green colours to indicate what they identify as regions of low, medium or high registration accuracy equating to the bad, fair or good traffic light tool ratings respectively. Prior to performing the ratings, the following criteria were established: a good registration was where there was <2 mm deviation between features on the two images, a fair registration was a 3–5 mm deviation and a bad registration was >5 mm deviation. Ratings were performed on the transverse slices of the deformed images encompassing the most superior volume of the parotid to the most inferior volume of the PTV.

Of the 35 patients included, patients 1–30 were assessed using the MIM6.7.6 QA tool by three RTs (A, B, C) where each RT assessed 10 different patients resulting in 30 different ratings. For patients 31–35, all five RTs

Table 1. Summary of the varying levels of radiotherapy experiences across RTs A-E.

RT	Level of clinical knowledge of DIR	Level of MIM6.7.6 experience
A	Advanced	Advanced
B	Intermediate	Advanced
C	Advanced	Minimal
D	Basic	Minimal
E	Basic	Minimal

Abbreviations: RT = Radiation Therapist, DIR = Deformable Image Registration.

Table 2. Description of the quantitative metrics, their recommended value and expected trends for the traffic light tool ratings,

Quantitative metric	Metric description	Recommended value for high registration accuracy	Expected trend for good, fair and bad ratings
HD	Point-based measure between the maximum distance of a point in one contoured volume to that of another[8]	<3 mm[8]	Value to increase from good to bad ratings
MDA	Point-based measure where the distances between a series of points in one contour to a series of points in a second contour are calculated and averaged[7]	<2–3 mm[7]	Value to increase from good to bad ratings
DSC	Measures the volume of overlap between two contours, with 0 indicating no overlap and 1.0 perfect overlap[7]	>0.8–0.9[7]	Value to decrease from good to bad ratings
Jacobian determinant	Identifies the local volume change across the two images after DIR[7]	>0 (0–1 for reasonable volume reduction, >1 for reasonable volume expansion)[7]	Value to be close to 1 for good ratings and values to increase or decrease in magnitude in a positive or negative direction away from 1 progressively for the fair to bad ratings

Abbreviations: HD = Hausdorff Distance, MDA = Mean Distance to Agreement, DSC = Dice Similarity Coefficient.

performed the ratings independently on these five patients resulting in a total of 25 patient ratings. This data was used to establish inter-operator reliability. Once each RT had completed their ratings, the two RTs with the highest level of experience (A and C) used the tool to rate patients 31–35 together. It is important to note that a decision was made by RTs (A and C) when completing the consensus ratings that regions of artefact would be identified as a bad registration. These ratings are referred to as consensus ratings.

Quantitative data

A description of the quantitative metrics used in this study to evaluate DIR accuracy and the MIM6.7.6 QA tool is provided in Table 2. Table 2 also presents recommended values for each quantitative metric[7,8] and the expected trend of these values within the good to bad ratings.

Three structures (mandible, left parotid and right parotid) outlined on the primary planning CT and secondary diagnostic CT prior to DIR were transferred to the deformed diagnostic CT using the registration link (shown in Figure 1). These structures were chosen due to their proximity to the PTV as well as to provide soft-tissue and bony structures to compare after DIR. The QA tool enabled different regions of the same anatomical structure to be highlighted with all three ratings as shown in Figure 2. Using the Boolean operations tool on MIM6.7.6, these three structures were then segmented into subregions where the structures were divided by more than one of the three different ratings as shown in Figure 2.

The mean of the contour comparison metrics (HD, MDA and DSC) for each of these three structures and for their segmented subregions coinciding with more than one of the three different ratings were exported from MIM6.7.6. The mean, median, minimum and maximum

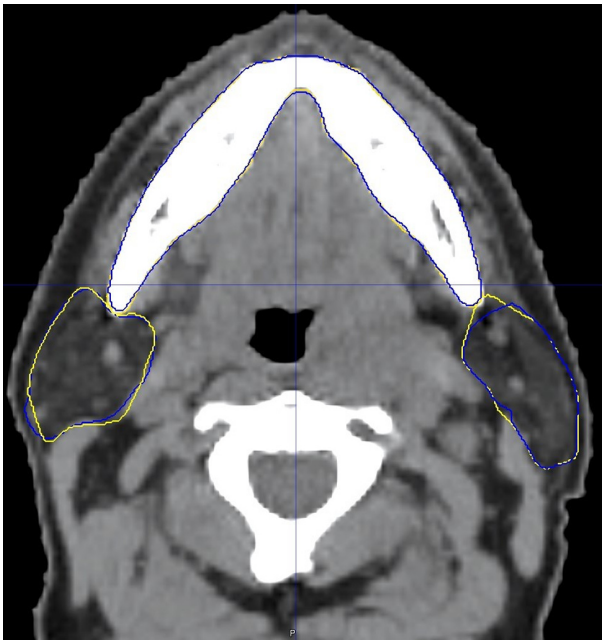

Figure 1. Example of a deformed image with transferred structures.Transverse slice of a patient deformed image showing the three structures (mandible, left parotid and right parotid) that were transferred from the primary planning CT (blue) and secondary diagnostic CT (yellow) images.

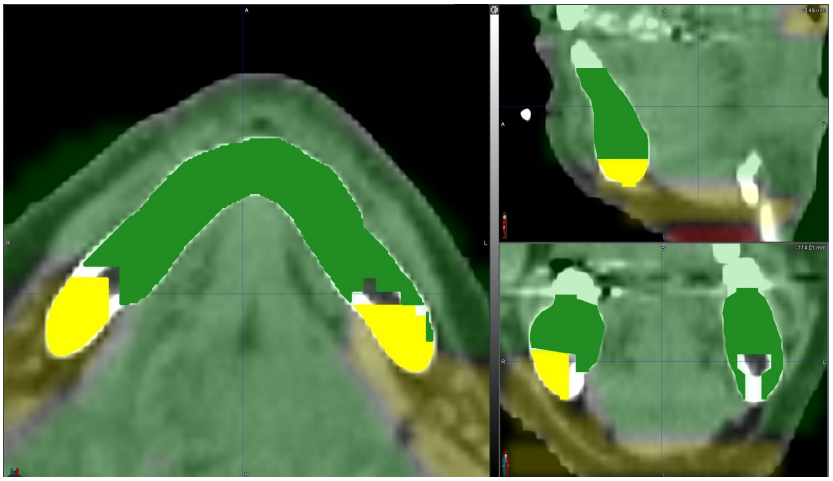

Figure 2. Example of the mandible segmented into subregions by rating. As the mandible had two different traffic light tool ratings applied to it, it was segmented into subregions of good (green) and fair (yellow) ratings on the transverse (left), sagittal (top right) and coronal (bottom right) slices of a patient deformed image.

Jacobian determinant values coinciding with the ratings were extracted through the Jacobian map created using the deformation grid from the deformed CT. Three subregions of the Jacobian map were segmented using thresholds based on the 95% interval of all the MIM6.76 calculated means of the Jacobian determinant values coinciding with fair ratings (rounded to the nearest 0.5 increment). The first subregion threshold applied to the Jacobian map was less than the lowest value of the 95% interval, the second was between 95% interval values and the third was greater than the highest 95% interval value. For patients 31–35 and RTs A-E, the frequency that each rating spatially coincided with the different thresholded subregions of the Jacobian map was recorded for each slice of the deformed image.

Tool efficiency

For all patients rated by RTs A, B and E (n = 35 ratings), the time (in minutes) taken to perform a qualitative rating assessment was recorded. As well as this, the volume of each rating highlighted was extracted from MIM6.7.6 in cm.[3]

Statistical analysis

All statistical analysis was performed in the R statistical package (https://www.r-project.org/) Descriptive statistics were used to evaluate the quantitative metrics (HD, MDA, DSC and Jacobian determinants) associated with each qualitative rating (good, fair and bad). Krippendorff's alpha reliability coefficients[9] were calculated to evaluate the reliability between all RT ratings for patients 31–35. This test is commonly used to determine inter-operator reliability for coders of survey and interview data as it takes missing data into account rather than removing subject and rating data from the calculations as is the case with the intra-class coefficient test. It is important to note that in the context of this study, missing data relates to a patient where a RT did not rate any regions coinciding within the parotids or mandible with the same rating(s) as the other RTs. Poor reliability is indicated by coefficient values <0.5, moderate reliability is indicated by 0.5–0.75, good reliability is indicated by 0.75–0.9 and excellent reliability is indicated by> 0.9.[10]

Results

Quantitative metrics for each qualitative rating

Contour comparison metrics

Table 3 presents the contour comparison metrics (HD, MDA and DSC) for each qualitative rating. For patients 1–30, the left parotid was the only structure that followed the expected trend (Table 2) for HD and MDA, whilst the left parotid and mandible followed the expected trend for DSC. When all structures were considered together, DSC was the only metric that consistently followed the expected trend of values. Unexpected trends were only seen in the fair rating.

Across all three structures and ratings, the mandible had the highest HD and MDA values, followed by the

Table 3. Summary of MIM6.7.6 calculated quantitative metrics or each anatomical structure and their spatial overlap with the different traffic light tool ratings.

					Structure by rating		
Metric	Structure	Patient	Operator	Entire structure	Good rating (mean ± SD)	Fair rating (mean ± SD)	Bad rating (mean ± SD)
HD (mm)	Left parotid	1–30	A, B, C	8.21 ± 3.45	7.36 ± 2.06	8.13 ± 5.75	12.69 ± 7.21
		31–35	A-E	8.09 ± 2.27	8.28 ± 2.94	9.26 ± 5.29	12.67 ± 8.15
		31–35	Consensus	8.09 ± 2.27	7.88 ± 2.48	*6.63* ± *1.93*	14.70 ± 10.44
	Right parotid	1–30	A, B, C	7.90 ± 4.44	7.62 ± 4.52	*15.71* ± *18.74*	9.86 ± 9.35
		31–35	A-E	8.28 ± 2.72	7.98 ± 2.60	8.18 ± 5.70	9.91 ± 8.71
		31–35	Consensus	8.28 ± 2.72	4.90 ± 3.14	7.92 ± 5.76	11.53 ± 13.07
	Mandible	1–30	A, B, C	6.39 ± 3.12	5.60 ± 3.50	*18.80* ± *31.68*	12.43 ± 9.09
		31–35	A-E	6.72 ± 3.68	6.29 ± 3.54	8.39 ± 9.52	10.12 ± 7.02
		31–35	Consensus	6.72 ± 3.68	3.71 ± 1.46	5.30 ± 4.56	11.13 ± 8.70
	Average across all structures	1–30	A, B, C	7.50 ± 3.80	6.85 ± 3.63	*15.09* ± *24.05*	11.80 ± 8.78
		31–35	A-E	7.70 ± 3.03	7.50 ± 3.18	8.56 ± 7.27	10.71 ± 8.04
		31–35	Consensus	7.70 ± 3.03	5.36 ± 2.97	6.61 ± 4.68	12.11 ± 11.05
MDA (mm)	Left parotid	1–30	A, B, C	0.84 ± 0.41	0.84 ± 0.65	1.49 ± 1.16	2.40 ± 1.80
		31–35	A-E	1.09 ± 0.60	0.93 ± 0.51	1.23 ± 1.03	1.76 ± 0.94
		31–35	Consensus	1.09 ± 0.60	0.88 ± 0.52	0.96 ± 0.50	1.96 ± 0.98
	Right parotid	1–30	A, B, C	0.82 ± 0.49	0.75 ± 0.47	*5.11* ± *12.82*	2.60 ± 2.90
		31–35	A-E	0.95 ± 0.63	0.73 ± 0.42	0.81 ± 0.50	1.55 ± 1.30
		31–35	Consensus	0.95 ± 0.63	0.39 ± 0.28	0.99 ± 0.68	1.30 ± 1.44
	Mandible	1–30	A, B, C	0.30 ± 0.14	0.27 ± 0.14	*5.47* ± *17.83*	1.21 ± 1.64
		31–35	A-E	0.39 ± 0.24	0.29 ± 0.12	0.51 ± 0.65	0.62 ± 0.42
		31–35	Consensus	0.39 ± 0.24	0.22 ± 0.04	0.27 ± 0.15	0.66 ± 0.43
	Average across all structures	1–30	A, B, C	0.65 ± 0.45	0.62 ± 0.53	*4.43* ± *13.90*	1.90 ± 2.19
		31–35	A-E	0.81 ± 0.61	0.64 ± 0.47	0.82 ± 0.79	1.25 ± 1.07
		31–35	Consensus	0.81 ± 0.61	0.47 ± 0.43	0.73 ± 0.60	1.21 ± 1.16
DSC	Left parotid	1–30	A, B, C	0.89 ± 0.05	0.85 ± 0.21	0.58 ± 0.35	0.31 ± 0.33
		31–35	A-E	0.84 ± 0.09	0.86 ± 0.08	0.67 ± 0.20	0.49 ± 0.25
		31–35	Consensus	0.84 ± 0.09	0.88 ± 0.08	0.68 ± 0.14	0.65 ± 0.16
	Right parotid	1–30	A, B, C	0.90 ± 0.05	0.90 ± 0.05	*0.42* ± *0.38*	0.50 ± 0.40
		31–35	A-E	0.86 ± 0.09	0.89 ± 0.07	0.74 ± 0.16	0.58 ± 0.25
		31–35	Consensus	0.86 ± 0.09	0.94 ± 0.04	*0.73* ± *0.15*	0.76 ± 0.16
	Mandible	1–30	A, B, C	0.95 ± 0.02	0.95 ± 0.02	0.73 ± 0.37	0.70 ± 0.30
		31–35	A-E	0.93 ± 0.04	0.95 ± 0.02	0.84 ± 0.22	0.75 ± 0.28
		31–35	Consensus	0.93 ± 0.04	0.96 ± 0.01	0.92 ± 0.01	0.83 ± 0.10
	Average across all structures	1–30	A, B, C	0.91 ± 0.05	0.90 ± 0.13	0.59 ± 0.39	0.52 ± 0.38
		31–35	A-E	0.88 ± 0.09	0.90 ± 0.07	0.76 ± 0.21	0.63 ± 0.28
		31–35	Consensus	0.88 ± 0.09	0.93 ± 0.06	0.78 ± 0.15	0.76 ± 0.16

Italics and underline indicate when the fair rating did not show progressive increase (HD and MD) or progressive decrease (DSC) from good to bad ratings.

Abbreviations: HD = Hausdorff Distance, MDA = Mean Distance to Agreement, DSC = Dice Similarity Coefficient.

right parotid and left parotid respectively. The lowest DSC values were in the mandible followed by the right parotid and left parotid respectively. It is also evident that the lowest HD and MDA values were in the mandible subregions coinciding with good ratings, and the highest HD and MDA values were in the mandible subregions coinciding with bad ratings. For DSC, the mandible obtained the highest DSC values within the good rating, and the left parotid obtained the worst DSC within the bad rating.

Table 4. Summary of the average MIM6.7.6 calculated statistics for the Jacobian determinant values coinciding with the traffic light tool ratings.

Statistic	Patient	Operator	Good rating (mean ± SD)	Fair rating (mean ± SD)	Bad rating (mean ± SD)
Mean	1–30	A, B, C	0.94 ± 0.05	0.88 ± 0.21	1.12 ± 0.68
	31–35	A-E	0.95 ± 0.06	0.91 ± 0.14	0.87 ± 0.12
	31–35	Consensus	0.94 ± 0.01	0.95 ± 0.06	0.81 ± 0.10
Median	1–30	A, B, C	0.98 ± 0.03	0.90 ± 0.26	1.05 ± 0.50
	31–35	A-E	1.00 ± 0.04	0.93 ± 0.20	0.91 ± 0.14
	31–35	Consensus	0.98 ± 0.02	1.00 ± 0.06	0.86 ± 0.14
Minimum	1–30	A, B, C	−0.04 ± 0.09	0.02 ± 0.10	0.09 ± 0.40
	31–35	A-E	0	0.04 ± 0.16	−0.10 ± 0.63
	31–35	Consensus	0	0	−0.24 ± 0.48
Maximum	1–30	A, B, C	2.66 ± 1.53	2.01 ± 0.61	3.28 ± 2.73
	31–35	A-E	2.20 ± 0.60	2.12 ± 0.55	3.11 ± 2.34
	31–35	Consensus	1.70 ± 0.32	1.9 ± 0.29	3.39 ± 2.22

Jacobian determinant data

Descriptive statistics for the Jacobian determinants of each rating are presented in Table 4. The expected trend (Table 2) was evident for patients 1–30 for the mean but not the median statistic values. It was also expected that bad ratings would obtain the lowest minimum values and highest maximum values for Jacobian determinants. However, this trend was not reflected in the mean of the minimum statistic values for patients 1–30 (Table 4). The good ratings obtained the lowest minimum values followed by fair and bad respectively. For the mean of the MIM6.7.6 maximum statistic, the trend was again not as expected, as the bad ratings obtained the highest maximum values whilst fair obtained the lowest maximum values.

Figure 3 demonstrates that fair ratings more frequently contained minimum and maximum values closest to 1, followed by good and bad respectively. Bad ratings contained the greatest number of high maximum values and the greatest number of negative minimum values which was to be expected. The shaded areas in Figure 3 also identify where 95% of all average Jacobian determinants fall. This range was −0.18–2.34 for bad ratings, 0.51–1.27 for fair ratings and 0.84–1.03 for good ratings. The magnitudes of these ranges were as expected; smallest for good and sitting around 1, and greatest for bad and sitting furthest away from 1.

To determine the frequency that the different ratings coincided slice by slice with the Jacobian determinant values, regions of <0.5 and >1.5 were defined. Bad ratings coincided with the greatest number of counts within the Jacobian threshold levels of <0.5 and >1.5 followed by good then fair (Table 5). As Figure 4 demonstrates, Jacobian determinants <0.5 were commonly located in the oral and tumour regions, whilst values >1.5 were commonly located in the periphery of the patient.

Inter-operator reliability

It is evident in Table 6 that Krippendorff's alpha reliability coefficients for all metrics ranged from poor (<0.5) to excellent reliability (>0.9). Figure 5 also shows this inter-operator variation of the ratings by anatomical structure and quantitative metric. Regions of excellent reliability were recognised in the MDA good subregion and DSC good subregion of the left parotid, and the MDA bad subregion of the right parotid. Good reliability was recognised in the mean Jacobian determinant bad, MDA good and fair subregions in the right parotid, and the DSC good subregion in the right parotid. In general, the lowest reliability was noticed in all mandible subregions which were all recognised as poor inter-operator reliability. The left and right parotids obtained quite similar Krippendorff's alpha reliability coefficients suggesting similar levels of inter-operator reliability when rating these structures.

Efficiency of the qualitative rating tool

The mean and median time to perform the ratings (n = 35) for a single patient was 33 and 45 minutes respectively, across RTs A, B and E. The volume for each rating applied to the deformed images is presented in Table 7. On average, good ratings were applied to the greatest volume of the deformed images and fair ratings were applied to the lowest volume.

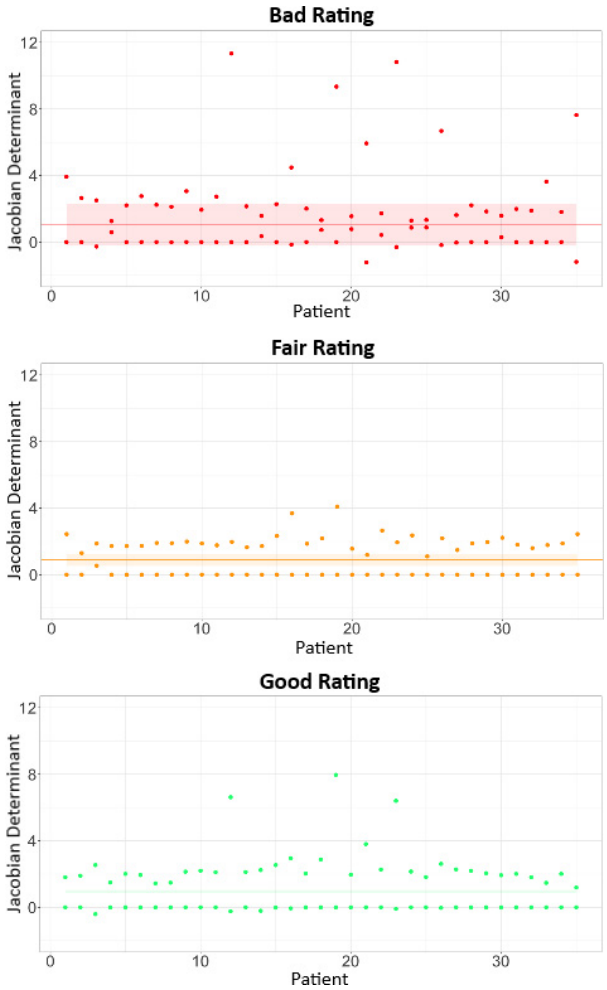

Figure 3. Distribution of minimum and maximum Jacobian determinant values for each traffic light tool rating across patients 1–30 and the consensus ratings of patients 31–35. A Jacobian determinant of 1 indicates no volume change. The dots represent the minimum and maximum MIM6.7.6 values. The line represents the mean for all MIM6.7.6 average values and the shaded area for each rating represents the mean±(1.96*SD) for all MIM6.7.6 average values. This shaded area therefore indicates that 95% of all the mean values for Jacobian determinants fall within −0.18–2.34 for bad, 0.51–1.27 for fair and 0.84–1.03 for good.

Discussion

This study evaluated a new MIM6.7.6 QA traffic light tool allowing RTs to qualitatively evaluate registration accuracy of regions within deformed images. Specific aims included comparing quantitative metrics recommended in the AAPM TG132 Report[7] to qualitative ratings and assessing inter-operator reliability of these ratings. By assessing rating time and volume measures of each rating, the tool's clinical efficiency was also examined. It was found that for structures overlapping the ratings, the quantitative metrics followed the expected trends. Inter-operator agreement was highest between the good and bad ratings, whilst agreement was lowest for the fair rating. Time and volume data provided directions for future clinical use of the QA tool.

Table 5. Frequency in counts of traffic light tool ratings coinciding slice by slice with Jacobian determinant threshold regions of < 0.5 and> 1.5 for patients 31–35 and RTs A-E.

	Good rating		Fair rating		Bad rating	
Threshold	<0.5	>1.5	<0.5	>1.5	<0.5	>1.5
Count	211	64	122	55	224	186
Total count	275		177		410	

In light of the AAPM TG132 Report[7] recommendations and the good rating values obtained in this study, the left parotid, right parotid and mandible achieved clinically acceptable results for DSC; however, only the mandible achieved clinically acceptable values for MDA. Because the AAPM TG132 Report[7] does not specify HD nor a clinically acceptable HD value, it was difficult to report on the acceptability of HD in this study. A study by Hvid et al.[11] does however recommend HD <3 mm classifying all structures evaluated in this study as not clinically acceptable. It has been highlighted that issues with outliers do present when evaluating HD; thus, perhaps HD should be eliminated for future DIR analysis.[12] For DSC, the good values obtained were comparable to those in a study by Latifi et al.[4] where the mandible obtained the highest DSC (0.88) followed by the right parotid (0.82) and left parotid (0.80) respectively. This is however different to findings by Varadhan et al.[12] as the mandible obtained the lowest DSC (0.63) out of the three structures in their study. Both studies[4,12] mentioned did follow a DIR method utilising only CT imaging similar to the methods in the current study. As the good values mostly fell within the clinically acceptable levels of DIR accuracy for all structures whilst the bad values always fell within the non-clinically acceptable levels, it indicates that the ratings were in agreement with the quantitative metrics obtained.

The QA tool used is similar to that used in studies by Guy et al.[13] and Hardcastle et al.[14]. Contrastingly, these studies do not indicate the specific measured deviation between the two registered images as conducted in the current study. It was however difficult and time consuming to measure this deviation on the deformed images to decide on one rating level. As the fair rating was less frequently used, the results imply that operators found it difficult to visualise a 2–5 mm deviation defined

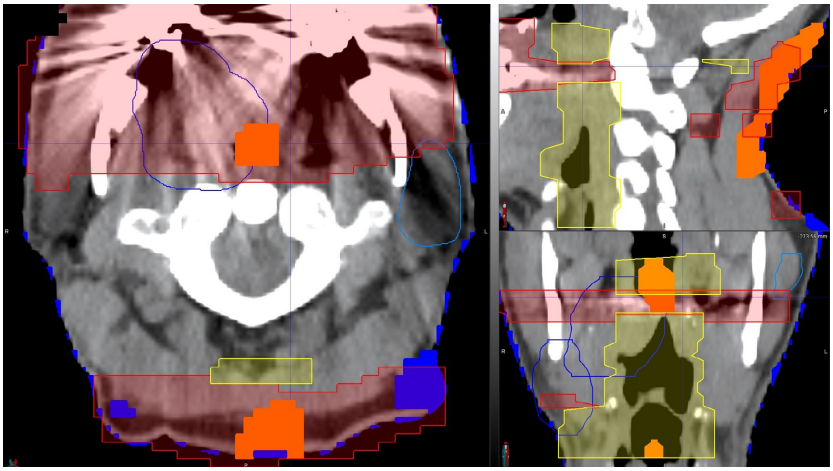

Figure 4. Spatial overlap of segmented subregions of the Jacobian map with the different traffic light tool ratings. Transverse (left), sagittal (top right) and coronal (bottom right) slices of a deformed patient image showing the segmentation of < 0.5 (blue) and> 1.5 (orange) Jacobian thresholds overlapping the fair (yellow) and bad (red) ratings

Table 6. Krippendorff's alpha reliability coefficients[9] for the five RTs and the consensus ratings for patients 31–35.

Metric	Structure	KALPHA for good rating	KALPHA for fair rating	KALPHA for bad rating
HD	Left parotid	0.66	0.52	-0.16
	Right parotid	0.66	0.17	0.53
	Mandible	0.24	0.24	0.14
MDA	Left parotid	0.91	0.32	0.19
	Right parotid	0.74	0.76	0.92
	Mandible	0.37	0.10	0.67
DSC	Left parotid	0.93	0.19	-0.15
	Right parotid	0.85	-0.15	0.35
	Mandible	0.50	0.09	0.12
Jacobian determinant	Mean	0.15	0.02	0.34
	Median	0.09	-0.13	0.87

KALPHA values < 0.50 indicate poor reliability, 0.50–0.75 indicates moderate reliability. 0.75–0.90 indicates good reliability and values> 0.9 indicate excellent reliability.[10]
Abbreviations: KALPHA = Krippendorff's alpha reliability coefficient, HD = Hausdorff Distance, MDA = Mean Distance to Agreement, DSC = Dice Similarity Coefficient.

by the fair rating. If this qualitative approach were to be implemented in clinical practice, RTs would need specific training to assess the difference between good, fair and bad registrations. Another approach to avoid the uncertainty of the three ratings would be to only highlight bad registrations. This would still be very clinically relevant as it would show the ROs and RTs where to be cautious when outlining tumour and OAR volumes. Because the current study demonstrates common agreement amongst operators in discriminating regions of bad registrations, this alternative single rating approach would more easily be implemented than the three-level traffic light tool. Although other studies[13,14] do not comment on the time to rate a DIR, the time to perform a rating for this study also further emphasises that utilising one rating level would be suffice. This would enhance the ease and effectiveness of practically implementing this method in clinical practice.

Another factor that negatively contributed to the tool's efficacy in this study was that ratings could split through an anatomical structure and hence create small segmented volumes for evaluation. These small isolated volumes created contributed to the large variations in standard deviations of HD and MDA. Therefore, it is recommended that a single rating is applied to a whole structure rather than subregions. This would eliminate the influence of small isolated subregions on HD and MDA values and have the added benefit of reducing the time taken to perform DIR ratings.

Regions of dental artefact were associated with higher levels of inter-operator variation as the mandible consistently obtained the lowest reliability across all metrics. This is similar to findings in a study by Bhatnagar et al.[15] where they noted issues with operator variation when delineating PTVs around dental artefact. Because of this, the consensus RTs agreed that artefact must always be considered as a bad registration for all future ratings.

Despite the differing reliability measures obtained, the findings in Tables 3 and 4 demonstrate only minor differences between the five RT ratings and consensus ratings. Koo et al.[10] study performed similar intra-class correlation testing and suggested at least 30 subjects and

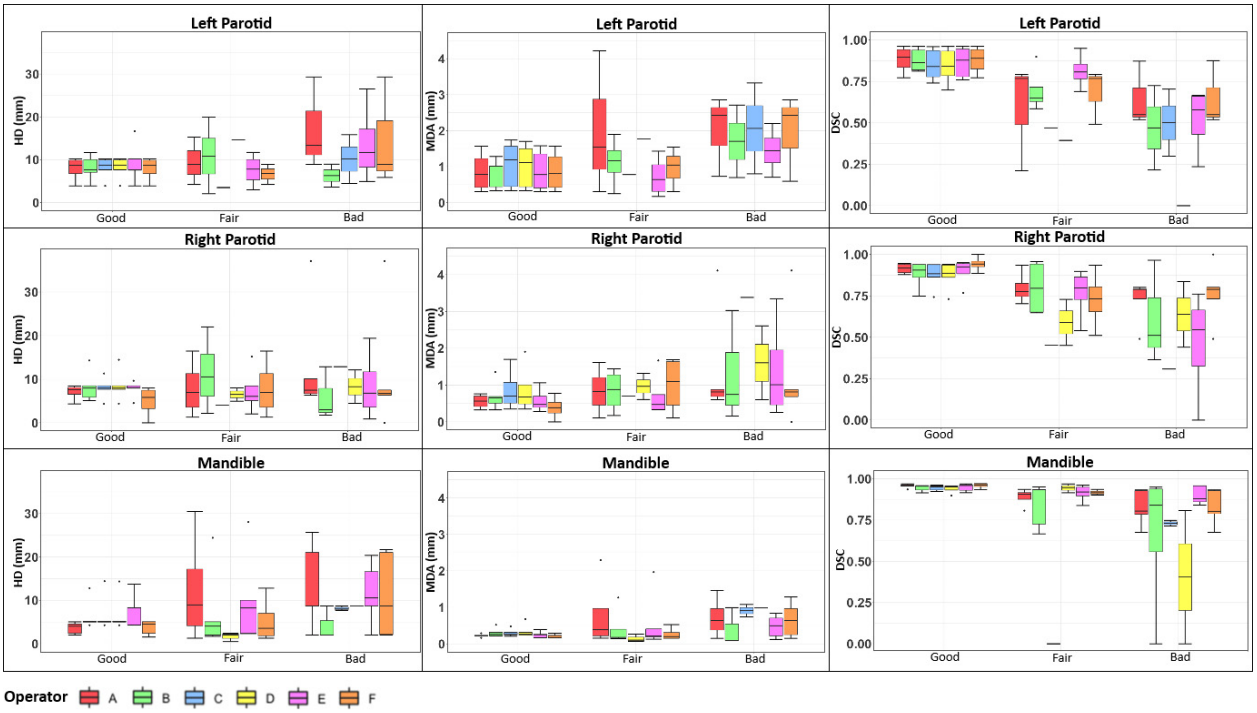

Figure 5. Boxplots of each contour comparison metric for all three structures and the ratings of the five RTs and consensus RTs for patients 31–35 (note F = consensus ratings). Abbreviations: HD = Hausdorff Distance, MDA = Mean Distance to Agreement, DSC = Dice Similarity Coefficient.

Table 7. Volume highlighted for each traffic light tool rating using the MIM6.7.6 QA tool on the deformed images across all patients and operators.

Patient	Operator	Good rating (cm^3)	Fair rating (cm^3)	Bad rating (cm^3)
1–30	A, B, C	2202.28 ± 980.68	290.00 ± 361.45	455.25 ± 573.26
31–35	A-E	1564.79 ± 689.67	299.57 ± 257.96	727.39 ± 1202.56
31–35	Consensus	1156.35 ± 569.78	375.05 ± 286.03	1050.02 ± 1492.79

three operators to be included when interpreting inter-operator reliability. As only five patients were utilised in this study's testing, the low reliability estimates obtained may be indicative of the small sample size used to test inter-operator reliability.

The main limitations of this study are that DIR was performed on a single modality imaging technique and of the head and neck region only. To improve volume delineation for head and neck cancers, future research should examine DIR between multimodality images of diagnostic MRI to planning CT.[16] Compared with other studies that evaluated several organs in the head and neck region, this study only evaluated three organs. Despite this, several quantitative metrics, qualitative ratings and operators were included in this study providing an accurate assessment of the DIR. By also comparing the individual operator ratings to the consensus ratings rather than as a whole, as performed in this study, it would demonstrate the specific inter-operator agreement between RTs with varying levels of experience when using this tool. This would again assist in developing clinical guidelines for this QA tool. As this study focussed on a combined RIR and DIR workflow, it cannot be assumed that RIR alone would provide similar results when using this tool. Despite this, DIR is becoming widely implemented and is a more appropriate method of image registration for head and neck CT imaging.[5] Because this study was limited to evaluating DIR accuracy to assist with decision-making regarding appropriate margins, when incorporating multimodality imaging information to delineate tumour and OAR volumes during treatment planning, the findings may not be relevant for DIR applications of dose accumulation and image guidance. Although this method has not yet been implemented, the

results of this study will be used to develop clinical training and user guidelines with the opportunity to assess their efficacy via a future inter-rater reliability study.

Conclusion

When assessing the accuracy of DIR for tumour and OAR delineation, various recommendations were determined for the clinical implementation of the new QA tool in MIM6.7.6. Although data from quantitative methods for evaluating DIR accuracy, except the HD metric, were meaningful as per AAPM TG132[7] recommendations, the rating tool evaluated in this study was found to be an effective and practical way of communicating DIR accuracy that aligned with expected quantitative information associated with the different rating levels. Furthermore, along with the need to create clinical consensus guidelines and training for RTs and ROs, it is also recommended to only utilise a qualitative DIR assessment approach and only rate areas of bad DIR accuracy when using this tool.

Acknowledgements

The authors acknowledge the support of MIM Software, Ohio, USA, for this project by providing in-kind access to the software.

Conflict of Interest

The authors declare no conflicts of interest.

References

1. Chuter R, Prestwich R, Bird D, et al. The use of deformable image registration to integrate diagnostic MRI into the radiotherapy planning pathway for head and neck cancer. *Radiother Oncol* 2017; **122**: 229–35.
2. Batumalai V, Holloway LC, Kumar S, et al. Survey of image-guided radiotherapy use in Australia. *J Med Imaging Radiat Oncol* 2017; **61**: 394–401.
3. Leibfarth S, Mönnich D, Welz S, et al. A strategy for multimodal deformable image registration to integrate PET/MR into radiotherapy treatment planning. *Acta Oncol* 2013; **52**: 1353–59.
4. Latifi K, Caudell J, Zhang G, Hunt D, Moros E, Feygelman V. Practical quantification of image registration accuracy following the AAPM TG-132 report framework. *J Appl Clin Med Phys* 2018; **19**: 125–133.
5. Mohamed AS, Ruangskul MN, Awan MJ, et al. Quality assurance assessment of diagnostic and radiation therapy-simulation CT image registration for head and neck radiation therapy. *Radiology* 2015; **274**: 752–63.
6. Mencarelli A, van Kranen S, Hamming-Vrieze O, et al. 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* 2014; **90**: 680–87.
7. Brock KK, Mutic S, McNutt TR, Li H, Kessler ML. Use of image registration and fusion algorithms and techniques in radiotherapy: Report of the AAPM Radiation Therapy Committee Task Group No. 132. *Med Phys* 2017; **44**(7): e43–e76.
8. Woerner AJ, Choi M, Harkenrider MM, Roeske JC, Surucu M. Evaluation of deformable image registration-based contour propagation from planning CT to cone-beam CT. *Technol Cancer Res Treat.* 2017; **16**: 801–10.
9. Krippendorff K. Computing Krippendorff's Alpha-Reliability, 2011. Available from: https://repository.upenn.edu/asc_papers/43
10. Koo T, Li M. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med.* 2016; **15**: 155–63.
11. Hvid CA, Elstrøm UV, Jensen K, Alber M, Grau C. Accuracy of software-assisted contour propagation from planning CT to cone beam CT in head and neck radiotherapy. *Acta Oncol.* 2016; **55**: 1324–30.
12. Varadhan R, Karangelis G, Krishnan K, Hui S. A framework for deformable image registration validation in radiotherapy clinical applications. *J Appl Clin Med Phys.* 2013; **14**: 192–213.
13. Guy CL, Weiss E, Che S, Jan N, Zhao S, Rosu-Bubulac M. Evaluation of Image Registration Accuracy for Tumor and Organs at Risk in the Thorax for Compliance With TG 132 Recommendations. *Adv Radiat Oncol.* 2018; **4**: 177–85.
14. Hardcastle N, van Elmpt W, De Ruysscher D, Bzdusek K, Tomé W. Accuracy of deformable image registration for contour propagation in adaptive lung radiotherapy. *Radiat Oncol.* 2013; **8**: 1–8.
15. Bhatnagar P, Subesinghe M, Patel C, Prestwich R, Scarsbrook AF. Functional imaging for radiation treatment planning, response assessment, and adaptive therapy in head and neck cancer. *Radiographics* 2013; **33**: 1909–29.
16. Taylor A, Sen M, Prestwich RJ. Assessment of the impact of deformable registration of diagnostic MRI to planning CT on GTV delineation for radiotherapy for oropharyngeal carcinoma in routine clinical practice. *Healthcare.* 2018; **6**: 1–11.