Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

Original article

Interobserver variability of corresponding anatomical landmark placement in pelvic CT and MRI scans

Georgios Andreadis [a],*, Wendy Groot [a], Stephanie M. de Boer [a], Peter A.N. Bosman [b], Tanja Alderliesten [a],*

[a] *Dept. of Radiation Oncology, Leiden University Medical Center, P.O. Box 9600, Leiden, 2300 RC, Zuid-Holland, The Netherlands*
[b] *Evolutionary Intelligence Group, Centrum Wiskunde & Informatica (CWI), P.O. Box 94079, Amsterdam, 1090 GB, Noord-Holland, The Netherlands*

ARTICLE INFO

Keywords:
Anatomical landmarks
Interobserver variability
Deformable image registration
Cervical cancer

ABSTRACT

Background and purpose: Anatomical landmarks are often used to assess the quality of a deformable image registration (DIR) between two scans. However, such landmarks are manually placed on both scans, which is prone to observer variability. We analyzed the interobserver variability of the placement of corresponding landmarks on pelvic scans, to provide context for DIR validation studies that use landmarks as a reference.
Material and Methods: Pelvic CT and MRI scans of nine cervical cancer patients were distributed to 17 observers. Three annotation settings were considered, each including scan pairs of five patients: CT-CT (13 observers), MRI-CT (eight observers), and MRI-MRI (eight observers). The observer group consisted of 15 RTTs professionally trained in working with scans of the given modality, and two fourth-year Ph.D. students in the domain. During annotation, observers received the same reference scan of each patient with 23 anatomically relevant, pre-annotated landmarks, and were asked to annotate the corresponding location of each reference landmark on a second scan of the same patient. To quantify the interobserver variability between different landmark placements on the same patient scan, their geometric median was used to approximate the true corresponding landmark location.
Results: Placements of landmarks on soft tissue for all patients deviated from their geometric median by a median 3D Euclidean distance of 3.0 mm (CT-CT), 5.6 mm (MRI-CT), and 2.6 mm (MRI-MRI). On bony anatomy, variability was significantly lower. Overall, variability was positively correlated with the deformation magnitude in the region.
Conclusions: There is large interobserver variability in anatomical landmark placements on pelvic CT and MRI scans. Variability frequently exceeds voxel size, challenging the AAPM guideline recommending landmark-based DIR quality assessment.

Introduction

Modern radiation treatment involves several image acquisitions before, during, and after treatment. For these acquisitions, different medical imaging modalities are used, such as CT and MRI. Several promising applications in radiation treatment have emerged which require the transfer of information between two acquired imaging scans of the patient, such as contour propagation, dose accumulation, and treatment plan adaptation [1–3]. These applications all require deformable image registration (DIR) to provide a spatial correspondence between the two scans. In many applications, these scans may have been acquired using differing imaging modalities, requiring a multi-modal DIR [3].

Finding a transformation which defines a spatial correspondence between scans is considered especially challenging in the case of large, local deformations and content mismatch [4,5]. Furthermore, validation and quality assurance of DIR transformations is hampered by the lack of a known ground truth transformation, which the DIR should approximate. A common quality measure for DIRs involves the manual annotation of anatomical landmarks [3,6,7]. Such landmarks are placed at anatomically relevant sites on both scans to measure the registration accuracy at these locations. However, manual placement of landmarks is prone to interobserver variability, which could impact their validity as a quality measure.

Interobserver variability has been studied for annotation tasks of various structures, such as target volumes [8–10], blood vessels [11], and organs at risk (OARs) [12,13]. Although variability of landmark annotations has also been investigated, existing studies either are limited to anatomical sites with small deformations [14,15], only report sec-

* Corresponding author.
E-mail addresses: G.Andreadis@lumc.nl (G. Andreadis), T.Alderliesten@lumc.nl (T. Alderliesten).

https://doi.org/10.1016/j.radonc.2026.111430
Received 22 September 2025; Received in revised form 17 December 2025; Accepted 29 January 2026
Available online 2 February 2026

Table 1
Listing of all anatomical landmarks included in this study. For each landmark, we indicate whether it is located on bony anatomy (or on soft tissue), and on which reference modality it is defined. Those on CT are used in CT-CT annotation, those on MRI are used both for MRI-CT and MRI-MRI annotation.

ID	Description	Bony	Ref. modality	
		anatomy	CT	MRI
L01	os coccygis	✓	✓	✓
L02	medial tip of trochanter minor right	✓	✓	
L03	medial tip of trochanter minor left	✓	✓	
L04	most caudal and dorsal point of lumbar vertebra 3	✓	✓	
L05	most caudal and ventral point of lumbar vertebra 3	✓	✓	
L06	most caudal and dorsal point of lumbar vertebra 5	✓	✓	
L07	most caudal and ventral point of lumbar vertebra 5	✓	✓	
L08	most ventral intersection of S1 and S2	✓		✓
L09	most ventral intersection of S2 and S3	✓		✓
L10	most ventral intersection of S3 and S4	✓		✓
L11	anterior superior border symphysis	✓		✓
L12	posterior inferior border symphysis	✓		✓
L13	femur head right	✓		✓
L14	femur head left	✓		✓
L15	acetabulum right	✓		✓
L16	acetabulum left	✓		✓
L17	caudal tip of the right kidney		✓	
L18	caudal tip of the left kidney		✓	
L19	umbilicus		✓	
L20	bifurcation of the aorta		✓	
L21	bifurcation of the vena iliaca communis right		✓	
L22	bifurcation of the vena iliaca communis left		✓	
L23	bifurcation of the artery iliaca communis right		✓	
L24	bifurcation of the artery iliaca communis left		✓	
L25	external anal sphincter		✓	
L26	internal anal sphincter		✓	✓
L27	internal urethral ostium		✓	✓
L28	external urethral ostium		✓	✓
L29	ureteral ostium right		✓	✓
L30	ureteral ostium left		✓	✓
L31	uterus top		✓	✓
L32	cervical ostium externum		✓	✓
L33	uterine isthmus			✓
L34	intrauterine canal top			✓
L35	ligament rotundum right			✓
L36	ligament rotundum left			✓
L37	right entrance of uterine artery in cervix			✓
L38	left entrance of uterine artery in cervix			✓

ondary variability outcomes derived from landmark placements, such as the accuracy of landmark-based rigid registration [6,16–19], only investigate intraobserver variability over time [20,21], or only include two observers and therefore lack statistical power [22,23]. Moreover, the impact of different imaging modalities on landmark variability has not yet been investigated. Consequently, validation studies which use anatomical landmarks to assess DIR quality, lack a reliable quantitative variability estimate. Although there are guidelines for quality control of clinical DIR applications [3], the quality impact of interobserver landmark variability has not been sufficiently explored. In this study, we therefore aim to quantify and analyze this variability on pelvic CT and MRI scans that feature large, local deformations and content mismatch, in both single- and multi-modality settings.

Material and methods

Patients

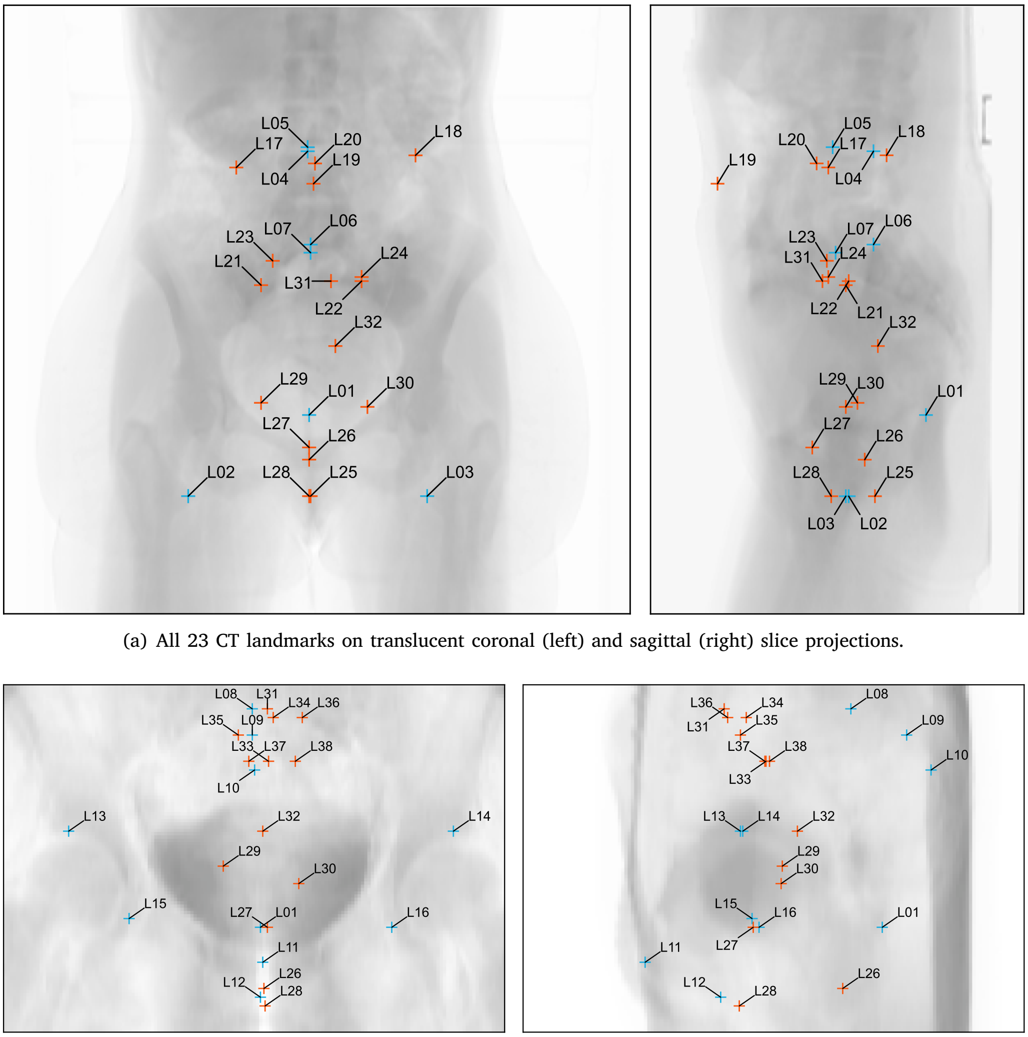
CT and MRI scans of nine cervical cancer patients previously treated at the department of Radiation Oncology at the Leiden University Medical Center (NL) were included. The responsible Medical Research Ethics Committee determined that the Medical Research Involving Human Subjects Act (WMO) does not apply to this retrospective study. Patients were treated between 2019 and 2020 according to EMBRACE-II criteria and aged 51 years on average at the time of treatment, ranging between 31 and 74 years. Further details are provided in Appendix A.

Annotation task

A retrospective observer study was initiated to measure interobserver variability of anatomical landmark placements. The task was defined as follows: given a reference scan with up to 23 anatomically relevant landmarks already visibly annotated (see Table 1 and Fig. 1), identify the corresponding location of each reference landmark on a second scan of the same patient. The reference set of anatomical landmarks was annotated by an experienced radiation therapy technologist (RTT, co-author W.G.) and verified by a radiation oncologist specialized in gynaecological tumours (co-author S.M.B.). Observers were provided with the annotation protocol used for the reference landmarks (see Appendix B). During annotation, observers also recorded their confidence for each placement on a three-level scale: certain, small uncertainty, and large uncertainty. After annotation, placements were reviewed for protocol conformity. Annotation was performed using the RayStation v10B treatment planning system (RaySearch Laboratories, Stockholm, Sweden).

Observers

A group of 17 observers, consisting of 15 RTTs and two researchers, participated in this study. All RTTs are professionally trained in working with scans of the given modality, with at least three years of experience. The two participating researchers (one of which is author



(a) All 23 CT landmarks on translucent coronal (left) and sagittal (right) slice projections.

(b) All 23 MRI landmarks on translucent coronal (left) and sagittal (right) slice projections.

Fig. 1. Locations of all defined landmarks on the reference scan of a patient, for both tested modalities (see (a) for CT and (b) for MRI). The color of each landmark indicates whether it is on bony anatomy *(blue)* or soft tissue structures *(red)*. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

G.A.) are not formally trained, but are fourth-year Ph.D. students in the domain. Additional details are listed in Appendix A.

Modalities

The influence of imaging modality on interobserver variability was examined by dividing the study into two groups. In one group, observers were provided with a reference CT scan and tasked with annotating another CT scan of the same patient. This modality pair is referred to as “CT-CT”. In the other group, observers were provided with a reference MRI scan and tasked with annotating both an MRI scan (“MRI-MRI”) and a CT scan (“MRI-CT”) of the same patient. For each study group, anatomical landmarks were defined which are clearly identifiable on the reference modality, based on existing specifications by a radiation oncologist of anatomical locations relevant to cervical cancer radiation treatment [24,25].

CT as reference modality

In the study group with CT as the reference modality, five patients were included. Two pelvic CT scans, acquired for external beam radiation treatment (EBRT) planning purposes, were used per patient. A CT scan with a full bladder was used as the reference, while a scan of the same patient with an empty bladder, acquired shortly after the reference scan, was annotated by the observers. The patients were selected to represent diverse bladder volumes and volume-differentials between scan pairs (see Appendix A.1). A reference set of 23 anatomically relevant landmarks was defined, with seven landmarks located on bony anatomy and 16 on soft tissue structures (see Fig. 1(a) and Table 1). Of the 17 observers, 13 participated in this group, of which 11 are RTTs and two are researchers.

MRI as reference modality

In the study group with MRI as the reference modality, five patients were included, one of which (P05) had also been included in the CT group. Two pelvic scans were used per patient MRI1 and MRI2, each acquired for treatment planning after implantation of a brachytherapy applicator and interstitial needles (if applicable). MRI1 was used for planning fractions 1 and 2, while MRI2 was used for planning fractions 3 and 4. Patients were randomly selected from a set of recently treated patients, all treated with the same applicator type. A reference set of 23 anatomically relevant landmarks was defined for these MRI scans, with 10 landmarks on bony anatomy and 13 on soft tissue structures (see Fig. 1(b) and Table 1). Eight of these landmarks were also present in the CT landmark set. Of the 17 observers, eight participated in this group, of which six are RTTs and two are researchers. Two of the RTTs only partially participated: one only annotated P06 and P07, and one only annotated P09.

Imaging

The CT scans were acquired with a Philips Brilliance Big Bore 120 kV scanner, with a slice thickness of 3 mm and in-slice resolutions between $(0.86, 0.86)$ mm and $(1.08, 1.08)$ mm. The MRI scans were acquired with a Philips Ingenia 1.5 Tesla T2-weighted TSE scanner, with a slice thickness of 4 mm and in-slice resolutions of either $(0.42, 0.42)$ mm or $(0.53, 0.53)$ mm (only the axial MRI sequence was used). Each scan pair was rigidly registered before annotation. Further details on imaging procedure are provided in Appendix A.4.

Quantification of interobserver variability

To quantify the interobserver variability of a landmark, we consider different placements of that landmark on one scan of a patient. As the “true” corresponding landmark location is not known, we use the geometric median of placements as an approximation. This is the placement that minimizes the sum of Euclidean distances to other placements, making the measure more robust to outliers than the coordinate-wise median of placements. We compute the interobserver variability of a particular landmark by measuring all placement distances to the geometric median, both in 3D and in separate orthogonal dimensions. The 3D distance of a placement to its geometric median can then be interpreted as its “accuracy”.

Statistical analysis

Given the distributions of distances between placements and their respective geometric medians, we perform a set of Mann–Whitney U tests. We test for differences between landmark variability on bony anatomy and on soft tissue, for all modality pairs. Moreover, we perform pairwise tests between the overall landmark variability distributions of all modality pairs. We also test if observer confidence differs between modality pairs. In addition to these tests, we perform a set of Spearman rank-order correlation coefficient tests for each modality pair to examine correlations between placement distances and other study variables. Firstly, we test for correlation with the magnitude of deformation associated with the landmarked region, by considering the distance between the reference location of a landmark on one scan and the geometric median of observer placements on the other. Secondly, we test for correlation with the observer-reported confidence, to assess if observers reliably estimate their own accuracy. Finally, we test for correlation with the formal training of observers, and their years of experience. For all statistical tests, we use an alpha level of 0.05 and Bonferroni correction where applicable. An overview of all tests and the number of data points used in each test is provided in Appendix C.

Results

While reviewing all landmark placements, we excluded outlier MRI-CT placements of L09 and L10, as some observers confused adjacent vertebrae. We also observed potential misunderstandings of the L31 landmark placement. In some patients, the uterus tilted ventrally and caudally between scans due to a change in bladder volume, causing the top of the uterus to become dislocated on the second scan, as shown in Figure D.13 Appendix D. Of the 13 observers in the CT study group, three placed the landmark at the absolute most cranial slice of the uterus regardless of its tilt. When asked for their reasoning, all three reported having misinterpreted the instruction for that landmark. To avoid skewing results, we excluded their CT-CT placements of this landmark across all patients.

Table 2
Overall interobserver variability for different modality pairs, across patients. The variability is measured by the distribution of distances from the geometric median of each landmark, with 23 landmarks and 5 patients defined per reference modality. The median and IQR is reported for each modality pair, landmark type (i.e., a landmark on bony anatomy or on soft tissue), and distance type (i.e., 3D, or one of the three dimensions).

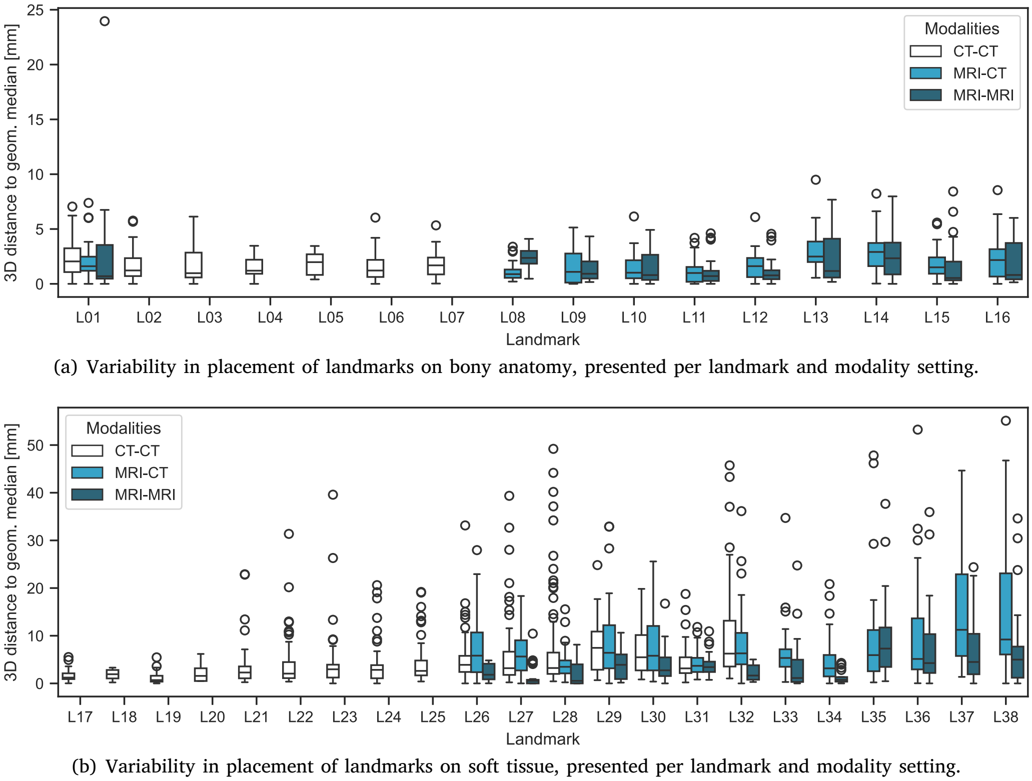
	CT-CT	MRI-CT	MRI-MRI
Variability of landmarks on bony anatomy [mm]			
Number of samples	$n = 372$	$n = 280$	$n = 291$
3D distance	1.4 ([0.8–2.5])	1.5 ([0.7–3.0])	0.9 ([0.5–2.9])
Left–right	0.4 ([0.0–2.0])	0.3 ([0.1–2.6])	0.2 ([0.0–2.0])
Anterior-posterior	0.6 ([0.2–1.1])	0.5 ([0.1–1.0])	0.4 ([0.1–0.9])
Cranial-caudal	0.4 ([0.1–1.0])	0.4 ([0.1–1.0])	0.4 ([0.1–0.7])
Variability of landmarks on soft tissue [mm]			
Number of samples	$n = 891$	$n = 414$	$n = 429$
3D distance	3.0 ([1.4–5.7])	5.6 ([2.8–10.7])	2.6 ([0.7–5.0])
Left–right	1.9 ([0.3–3.4])	2.8 ([0.8–6.5])	0.5 ([0.1–3.9])
Anterior-posterior	1.2 ([0.5–2.9])	2.4 ([0.8–5.5])	0.9 ([0.3–2.3])
Cranial-caudal	0.8 ([0.3–1.6])	1.4 ([0.5–3.4])	0.6 ([0.2–1.9])

The overall interobserver variability for each scan modality pair is shown in Table 2. We find that landmarks on bony anatomy are subject to significantly lower variability than landmarks on soft tissue, for all modality pairs ($p < 0.001$). For the full set of landmarks, median variability on MRI-MRI is significantly lower than on both other modality pairs ($p < 0.001$). Considering the results in Table 2, we observe the highest median variability of soft tissue placements in the MRI-CT setting. When the three dimensions are considered separately for the entire landmark set, we do not find a consistent relation between a dimension and the variability in it.

In Fig. 2, we aggregate the variability by landmark and modality pair. We give two examples in Fig. 3: in one, we observed a small variability (median of 1.7 mm, $n = 13$), in the other, a large one (median of 9.9 mm, $n = 13$). For all modality pairs, variability tends to significantly increase ($p < 0.001$) when the 3D Euclidean distance between a landmark’s geometric median on one scan and its reference point on the other increases, i.e., the magnitude of deformation in the region. We provide a histogram-like overview of the placement variability of each landmark per modality pair across patients, in Figure D.12 Appendix D.

In Fig. 4, we aggregate the interobserver variability of landmarks by patient and modality pair. Variability strongly differs between patients, indicating some may be more difficult to annotate than others. For example, on CT-CT, there is a median variability across landmarks of 3.1 mm on P02 ($n = 218$), but just 2.1 mm on P05 ($n = 230$). There also appear large differences between landmarks: While some are consistently placed within a small margin of variability (e.g., L08), other landmarks have outlier placements at a distance of over 50 mm from the geometric median, mainly in the MRI-CT annotation setting (e.g., L38). These extreme outliers only occur in soft tissue landmark placements, with the less severe outlier of bone landmark L01 attributable to poor bone-tissue contrast. On the landmarks shared across both CT and MRI modalities (L01 and L[26-32]), variability is consistently lowest in the MRI-MRI setting.

We find that an observer’s confidence about a landmark placement is negatively correlated with the distance of their placement from the respective geometric median ($p < 0.001$), indicating that observers often correctly assess their own placement accuracy. Correlation coefficients are −0.42, −0.52, and −0.36 for CT-CT, MRI-CT, and MRI-MRI, respectively. In general, we find that observers are significantly more confident on MRI-MRI placements, compared to other modality pairs ($p < 0.001$). For all three modality pairs, we do not find a significant correlation between whether an observer was formally trained and their accuracy. Similarly, no significant correlation was found between



(a) Variability in placement of landmarks on bony anatomy, presented per landmark and modality setting.

(b) Variability in placement of landmarks on soft tissue, presented per landmark and modality setting.

Fig. 2. Interobserver variability of anatomical landmark placement *grouped by landmarks*, as placed on different modality pairs. Variability is captured by the absolute three-dimensional (3D) distances between observer placements and their geometric median. Distribution boxes extend from quartiles Q1 to Q3, with the median indicated by a line within the box. Whiskers show the furthest data points still within 1.5 IQR from either Q1 or Q3, and the remaining points are indicated as outliers.

the number of years of experience of an observer and their accuracy for MRI-CT and MRI-MRI annotations, and only a weak correlation (coefficient 0.08, $p = 0.003$) for CT-CT annotations.

Discussion

Implications for DIR

We find large observer variability across landmarks, patients, and modalities. Several factors contribute to increased placement variability of certain landmarks. For example, landmarks in more heavily deformed regions are often more variable between observers. We also observe that the annotation task involving different modalities (MRI-CT) is especially susceptible to variability. In this task, the complementary tissue contrast of these modalities, as well as the content mismatch induced by the appearing applicator, complicates the matching of corresponding structures.

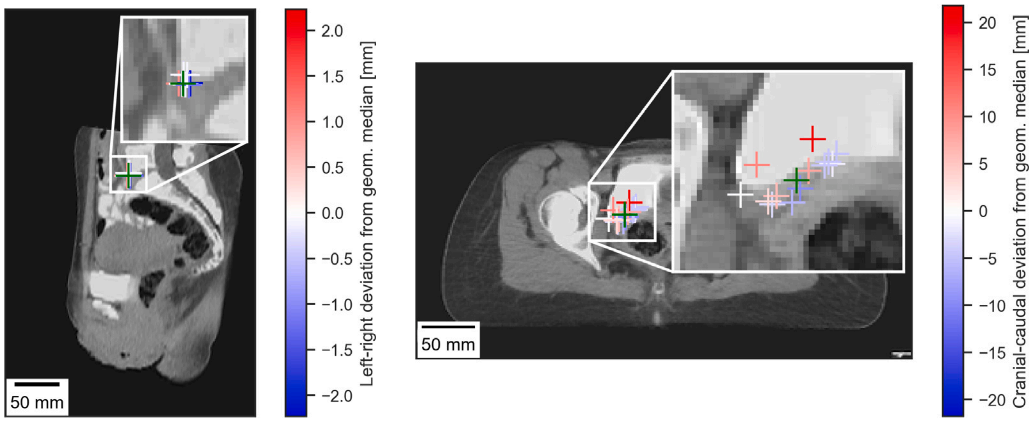
The large variability we find, exposes limitations of current image-based DIR validation practices, since these often rely on landmarks as a reference. Precisely those landmarks that are most relevant to the validation of challenging DIR tasks, such as those on heavily deformed regions, multi-modal scans, and mismatched content, are often the most variable. Such regions are already inherently more susceptible to registration errors [6,26–28]. For clinical applications such as dose accumulation, voxel-level accuracy can be critical. However, the accuracy of a registration can only be measured to this level if the landmark accuracy is known and sufficient. Existing guidelines claim the uncertainty of a landmark is “typically less than the voxel size” [3], but our findings challenge this claim. On each tested modality pair, the mean variability of several soft tissue landmarks exceeds the scan voxel size (see Figure D.12), bearing consequence for the use of landmarks as DIR reference. In Appendix E, we compare our findings to the interobserver variability reported in related work.

Potential limitations

There are potential limitations to this study. Firstly, the task definition may impact the outcome: observers were given an annotated reference scan and only annotated the second scan of a patient. In practice, observers annotate both scans simultaneously, possibly affecting interobserver variability. However, having observers annotate both scans would not allow for direct measurement of variability, as the reference locations of landmarks would differ. Furthermore, emphasis was placed in the protocol on finding the *corresponding location* of a reference location, which could differ from the observer’s own interpretation of the landmark on both scans. Additionally, the task definition used in this study limits our ability to compare our variability results with those reported in studies using different methodologies.

Secondly, the anatomical region and specification of landmarks may impact the measured variability. It is unclear whether the variability in this anatomical region generalizes to other regions. Furthermore, the landmark set specifications and observer groups may impact the measured variability comparison between different modalities, as both differ between modalities. However, we argue that the landmark sets are representative for their modality, both in terms of deformation magnitude and tissue structures, and that the number of observers in both study groups is sufficiently large to support a comparison.

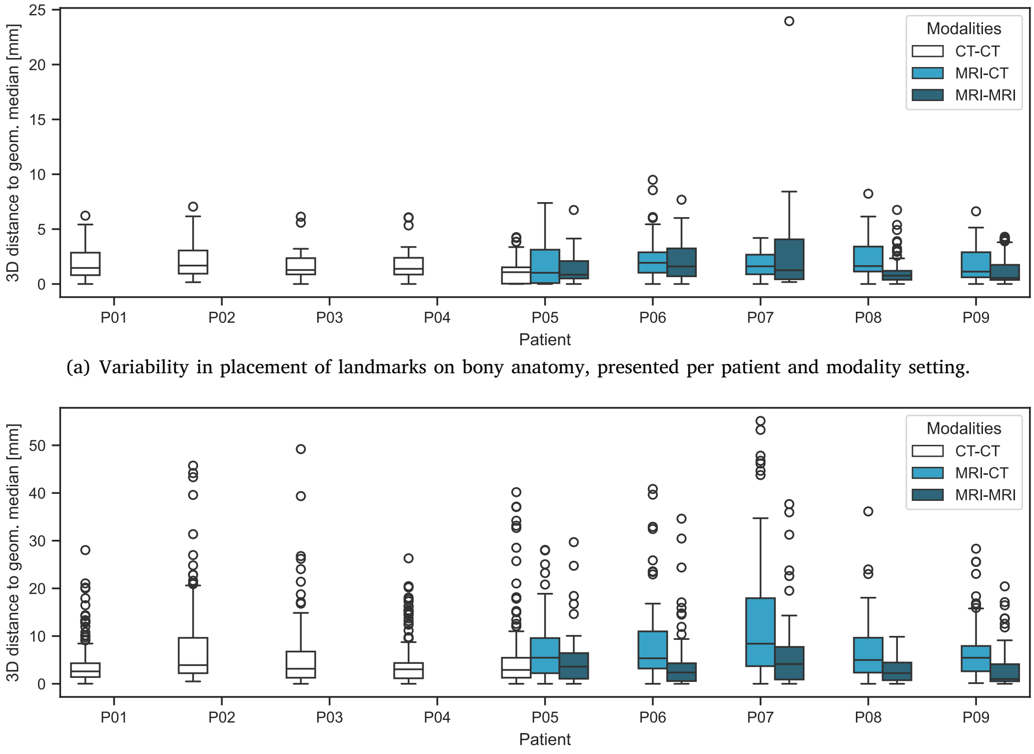
Thirdly, scan characteristics may impact the outcome. For example, local motion artifacts may complicate placement in certain instances. Moreover, only one out of three orthogonal MRI sequences was provided to observers. Additional sequences may have helped further reduce variability, but were not retrospectively available. Furthermore, the use of contrast agent was tailored to each patient, which may impact inter-patient variability of landmarks placed close to areas with enhanced contrast.


50 mm
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
Left-right deviation from geom. median [mm]
50 mm
20
15
10
5
0
−5
−10
−15
−20
Cranial-caudal deviation from geom. median [mm]

(a) A landmark where low interobserver variability (median of 1.7 mm, $n = 13$) was observed: L07, the most caudal and ventral point of lumbar vertebra 5 (sagittal view).

(b) A landmark where high interobserver variability (median of 9.9 mm, $n = 13$) was observed: L29, the right ureteral ostium (axial view).

Fig. 3. Two examples of landmarks placed by observers, during CT-CT annotation of patient P02. The geometric median is indicated in green, while the remaining landmarks are colored according to their deviation from the geometric median in the dimension orthogonal to the slice plane. The slice shown is the slice in which the geometric median is located. A zoomed-in view is provided, with three-fold magnification.



(a) Variability in placement of landmarks on bony anatomy, presented per patient and modality setting.

(b) Variability in placement of landmarks on soft tissue, presented per patient and modality setting.

Fig. 4. Interobserver variability of anatomical landmark placement *grouped by patients*, as placed on different modality pairs. Variability is captured by the absolute three-dimensional (3D) distances between observer placements and their geometric median. Distribution boxes extend from quartiles Q1 to Q3, with the median indicated by a line within the box. Whiskers show the furthest data points still within 1.5 IQR from either Q1 or Q3, and the remaining points are indicated as outliers.

Lastly, the fractions of trained and untrained observers differ between the study groups. The MRI group had a larger untrained fraction than the CT group. However, as observer expertise has proven uncorrelated with accuracy, this likely does not limit validity.

Future directions

Our findings provide important context to landmark-based DIR validation efforts. The reported interobserver variability could provide a lower bound for a meaningfully achievable level of accuracy, enhancing the interpretation of validation outcomes. Furthermore, the "crowd wisdom" emerging from this study could help in establishing a more reliable set of corresponding landmarks for a given context. We have therefore published a dataset of the imaging data used in this study together with the landmark placements.[1]

Future research could investigate other anatomical sites and observer groups. Moreover, one could explore whether algorithms for corresponding landmark generation [24] can be adapted to tailor their location-agnostic landmarks to specific anatomical sites, potentially enabling a study of human and algorithmic observer variability. Another promising research line is the development of virtual deformable phantoms [3,29], forming a controllable virtual environment. While landmarks can inherently only provide point-wise ground truths, virtual phantoms have the potential to address this shortcoming by providing a fully known deformation.

Conclusion

We conclude that there is large interobserver variability in anatomical landmark placement on pelvic CT and MRI scans. This holds especially for landmarks in heavily deformed regions, which are also the most important for DIR validation. In many cases, the observed variability far exceeds the voxel size, across all scan modality combinations tested. Our findings thereby challenge the prevailing assumptions about the reliability of anatomical landmarks for DIR validation. This can bear consequences for past and future validation efforts, as validation outcomes can only be meaningfully interpreted when the landmark variability is both known and acceptably low for the imaging modalities and anatomical sites involved.

CRediT authorship contribution statement

Georgios Andreadis: Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Wendy Groot:** Writing – review & editing, Investigation, Data curation. **Stephanie M. de Boer:** Writing – review & editing, Supervision, Methodology, Data curation. **Peter A.N. Bosman:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization. **Tanja Alderliesten:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

This research is part of the Open Technology Programme with project number 15586, financed by the Dutch Research Council (NWO), Elekta, and Xomnia. Further, this work is co-funded by the public–private partnership allowance for top consortia for knowledge and innovation (TKIs) from the Ministry of Economic Affairs.

Acknowledgments

The authors would like to thank all participants of the study for their time.

Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.radonc.2026.111430.

[1] The dataset is publicly available at the following DOI: https://doi.org/10.4121/b49aadc1-e435-4d67-a7d2-460d9ed02558.

References

[1] Eppenhof KAJ, Maspero M, Savenije MHF, de Boer JC, van der Voort van Zyp JRN, Raaymakers BW, Raaijmakers AJE, Veta M, van den Berg CAT, Pluim JPW. Fast contour propagation for MR-guided prostate radiotherapy using convolutional neural networks. Med Phys 2020;47(3):1238–48.

[2] Murr M, Brock KK, Fusella M, Hardcastle N, Hussein M, Jameson MG, Wahlstedt I, Yuen J, McClelland JR, Vasquez Osorio E. Applicability and usage of dose mapping/accumulation in radiotherapy. Radiother Oncol 2023;182:109527.

[3] 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):e43–76.

[4] Rigaud B, Simon A, Castelli J, Lafond C, Acosta O, Haigron P, Cazoulat G, de Crevoisier R. Deformable image registration for radiation therapy: principle, methods, applications and evaluation. Acta Oncol 2019;58(9):1225–37.

[5] Andreadis G, Bosman PAN, Alderliesten T. MOREA: a GPU-accelerated evolutionary algorithm for multi-objective deformable registration of 3D medical images. In: Proceedings of the 2023 genetic and evolutionary computation conference. 2023, p. 1294–302.

[6] Castillo R, Castillo E, Guerra R, Johnson VE, McPhail T, Garg AK, Guerrero T. A framework for evaluation of deformable image registration spatial accuracy using large landmark point sets. Phys Med Biol 2009;54(7):1849–70.

[7] Wu Z, Rietzel E, Boldea V, Sarrut D, Sharp GC. Evaluation of deformable registration of patient lung 4DCT with subanatomical region segmentations. Med Phys 2008;35(2):775–81.

[8] Eminowicz G, McCormack M. Variability of clinical target volume delineation for definitive radiotherapy in cervix cancer. Radiother Oncol 2015;117(3):542–7.

[9] van der Veen J, Gulyban A, Nuyts S. Interobserver variability in delineation of target volumes in head and neck cancer. Radiother Oncol 2019;137:9–15.

[10] Hellebust T, Tanderup K, Lervåg C, Fidarova E, Berger D, Malinen E, Pötter R, Petrič P. Dosimetric impact of interobserver variability in MRI-based delineation for cervical cancer brachytherapy. Radiother Oncol 2013;107(1):13–9.

[11] Oshin O, England A, Mcwilliams R, Brennan J, Fisher R, Rao Vallabhaneni S. Intra- and interobserver variability of target vessel measurement for fenestrated endovascular aneurysm repair. J Endovasc Ther 2010;17(3):402–7.

[12] van der Veen J, Gulyban A, Willems S, Maes F, Nuyts S. Interobserver variability in organ at risk delineation in head and neck cancer. Radiat Oncol 2021;16(1):120.

[13] Sanders JW, Mok H, Hanania AN, Venkatesan AM, Tang C, Bruno TL, Thames HD, Kudchadker RJ, Frank SJ. Computer-aided segmentation on MRI for prostate radiotherapy, part I: Quantifying human interobserver variability of the prostate and organs at risk and its impact on radiation dosimetry. Radiother Oncol 2022;169:124–31.

[14] Seim H, Kainmüller D, Heller M, Zachow S, Hege H. Automatic extraction of anatomical landmarks from medical image data: an evaluation of different methods. In: 2009 IEEE international symposium on biomedical imaging: from nano to macro. 2009, p. 538–41.

[15] van der Merwe J, van den Heever D, Erasmus P. Variability, agreement and reliability of MRI knee landmarks. J Biomech 2019;95:109309.

[16] Kiefer L, Fabian J, Lorbeer R, Machann J, Storz C, Kraus M, Wintermeyer E, Schlett C, Roemer F, Nikolau K, Peters A, Bamberg F. Inter- and intra-observer variability of an anatomical landmark-based, manual segmentation method by MRI for the assessment of skeletal muscle fat content and area in subjects from the general population. Br J Radiol 2018;91(1098):20180019.

[17] Frantz S, Rohr K, Stiehl H, Kim S, Weese J. Validating point-based MR/CT registration based on semi-automatic landmark extraction. In: Proceedings of the conference on computer assisted radiology and surgery. 1999, p. 233–7.

[18] Strasters K, Little J, Buurman J, Hill D, Hawkes D. Anatomical landmark image registration: validation and comparison. In: International conference on computer vision, virtual reality, and robotics in medicine. 1997, p. 161–70.

[19] Park A, Nam D, Friedman M, Duncan S, Hillen T, Barrack R. Inter-observer precision and physiologic variability of MRI landmarks used to determine rotational alignment in conventional and patient-specific TKA. J Arthroplast 2015;30(2):290–5.

[20] Brock KK, Sharpe MB, Dawson LA, Kim SM, Jaffray DA. Accuracy of finite element model-based multi-organ deformable image registration. Med Phys 2005;32(6):1647–59.

[21] Coselmon MM, Balter JM, McShan DL, Kessler ML. Mutual information based CT registration of the lung at exhale and inhale breathing states using thin-plate splines. Med Phys 2004;31(11):2942–8.

[22] Mencarelli A, van Beek S, van Kranen S, Rasch C, van Herk M, Sonke JJ. Validation of deformable registration in head and neck cancer using analysis of variance. Med Phys 2012;39(11):6879–84.

[23] Murphy K, van Ginneken B, Klein S, Staring M, de Hoop BJ, Viergever MA, Pluim JP. Semi-automatic construction of reference standards for evaluation of image registration. Med Image Anal 2011;15(1):71–84.

[24] Grewal M, Wiersma J, Westerveld H, Bosman PAN, Alderliesten T. Automatic landmark correspondence detection in medical images with an application to deformable image registration. J Med Imaging 2023;10(01).

[25] Grewal M, Westerveld H, Bosman PAN, Alderliesten T. Multi-objective learning for deformable image registration. In: Medical imaging with deep learning, vol. 178, 2024, p. 1–16.

[26] Rong Y, Rosu-Bubulac M, Benedict SH, Cui Y, Ruo R, Connell T, Kashani R, Latifi K, Chen Q, Geng H, Sohn J, Xiao Y. Rigid and deformable image registration for radiation therapy: a self-study evaluation guide for NRG oncology clinical trial participation. Pr Radiat Oncol 2021;11(4):282–98.

[27] Kadoya N, Fujita Y, Katsuta Y, Dobashi S, Takeda K, Kishi K, Kubozono M, Umezawa R, Sugawara T, Matsushita H, Jingu K. Evaluation of various deformable image registration algorithms for thoracic images. J Radiat Res 2014;55(1):175–82.

[28] Vickress J, Rangel Baltazar MA, Afsharpour H. Evaluation of varian's SmartAdapt for clinical use in radiation therapy for patients with thoracic lesions. J Appl Clin Med Phys 2021;22(3):150–6.

[29] Rodriguez CJ, de Boer SM, Bosman PAN, Alderliesten T. Bi-objective optimization of organ properties for the simulation of intracavitary brachytherapy applicator placement in cervical cancer. In: SPIE medical imaging 2023: image-guided procedures, robotic interventions, and modeling. 2023, p. 114–25.