Patient-specific validation of deformable image registration in radiation therapy: Overview and caveats

Chiara Paganelli,[a] and Giorgia Meschini
Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milano 20133, Italy

Silvia Molinelli
Centro Nazionale di Adroterapia Oncologica, Pavia 27100, Italy

Marco Riboldi
Department of Medical Physics, Ludwig-Maximilians-Universitat Munchen, Munich 80539, Germany

Guido Baroni
Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milano 20133, Italy
Centro Nazionale di Adroterapia Oncologica, Pavia 27100, Italy

Over the last few decades, deformable image registration (DIR) has gained popularity in image-guided radiation therapy for a number of applications, such as contour propagation, dose warping, and accu-mulation. Although this raises promising perspectives for the improvement of treatment outcomes and quality of radiotherapy clinical practice, the variety of proposed DIR algorithms, combined with the lack of an effective quantitative quality control metric of the registration, is slowing the transfer of DIR into the clinical routine. Recently, a task group (AAPM TG132) report was published outlining the essential aspects of DIR for image guidance in radiotherapy. However, an accurate and efficient patient-specific validation is not yet defined, and appropriate metrics should be identified to achieve the definition of both geometric and dosimetric accuracy. In this respect, the use of a dense set of anatomical landmarks, along with additional evaluations on contours or deformation field analysis, are likely to drive patient-specific DIR validation in clinical image-guided radiotherapy applications to account for geometric inaccuracies. Automatic and efficient strategies able to provide spatial information of DIR uncertainties and to evaluate monomodal and multimodal image registration, as well as to describe homogenous and un-contrasted regions are believed to represent the future direction in DIR validation. But especially in the case of DIR applications for dose mapping and accumulation, the need of accurate patient-specific validation is not only limited to the evaluation of geometric accu-racy. In fact, the need to account for dosimetric inaccuracies due to DIR represents another important area in the field of adaptive treatments. Different approaches are currently being investigated to quan-tify the effect of DIR error on dose analysis, mainly relying on clinically relevant dose metrics, or on the study of deformation field properties for a voxel-by-voxel evaluation. However, novel research is required for the definition of dedicated and personalized measures capable to relate the geometric and dosimetric inaccuracies, thus bearing useful information for a safe use of DIR by clinical end users. In this paper we provide insights on DIR results evaluation on a patient-specific basis, facing the issues of both geometric and dosimetric paradigms. Challenges on DIR validation are overviewed and dis-cussed, in order to push preliminary clinical guidelines forward on this fundamental topic and boost the implementation of more robust and reliable patient-specific evaluation metrics.

Key words: DIR, DIR assessment, DIR in radiotherapy, DIR validation

1. INTRODUCTION

Image guidance has been at the forefront of advances in external beam radiotherapy over the past decades. Thanks to the integration of imaging information in treatment design and workflow, the targeting of tumor and sparing of surrounding healthy tissues has increasingly become more accurate and effective in conventional X-rays and particle therapy.[1–4] The use of image-guided radiation therapy (IGRT) techniques provides a method to support clinicians in the definition of a personalized treatment, as well as to quantify inter- and intra-fractional anatomo-pathological changes occurring during radiotherapy, thus performing adaptive treatments.

The current clinical workflow focuses on the use of X-ray imaging, from single projections to volumetric and time-resolved (4D) imaging, able to capture organ motion. Computed Tomography (CT) represents the standard in radiation therapy for treatment planning, along with the support of on-board intra-modality imaging capable of capturing daily variations by means of cone beam CT (CBCT), in-room CT on

rails, kilovoltage (kV) or megavoltage (MV) imaging, often in combination with implanted fiducial markers.[2] This image modality can be also augmented by the acquisition of other image modalities which are increasingly gaining importance in radiotherapy. Among these, we find ultrasound,[5] Positron Emission Tomography (PET),[6] and Magnetic Resonance Imaging (MRI),[7,8] with integrated novel systems being one of the novel area of IGRT, such as PET/CT, PET/MRI[6] as well as in-room MRI units.[9]

In this scenario, modeling the transformation due to inter- and intra-fractional motion via image registration is fundamental to map, overlap, and integrate the information coming from different images. In particular, Deformable Image Registration (DIR) plays a key role to account for nonrigid changes which typically occur during a radiotherapy treatment.[10–13] Different algorithms have been proposed in the literature (e.g., intensity-based approaches such B-spline and demons, landmark-based thin-plate spline, or biophysical and finite element modeling-based registration) and DIR has extensively been proposed to analyze target motion and monitor tumor changes, including 4D motion modeling,[14–16] contour propagation,[17–22] and treatment adaptation by means of the so-called "virtual CT".[23–27] Other applications have been proposed also for multi-modal PET/CT,[28–31] PET/MRI,[32] and MRI/CT.[15,33–35]

An additional advantage of the use of DIR in a radiotherapy clinical workflow relies on the possibility of quantifying the actual distribution of radiation dose absorbed over the course of the treatment, by mapping the dose back to a common reference anatomy.[13,36–39] Warping the dose grid to the reference anatomy according to the obtained deformation vector field (DVF), represents a widely used approach. This solution has been adopted for dose accumulation,[13,36,37] "dose of the day" from virtual CT[24–26,35,40,41,42] and finally for 4D optimization strategies.[38,39] For this latter, an alternative solution to dose warping consists in directly including the changes in anatomy in the delivered fluence,[41,43–47] although this approach is not widely available for most commercial planning systems and still incorporate the use of a DVF.[39]

Even if several algorithms have been developed for DIR and several applications rely on its use, the lack of gold standard and quantitative control metrics are slowing the transfer of DIR into the clinical workflow. A recent survey by Viergever et al.[12] highlights that scientific research reached a milestone with DIR developments over the last 20 yr; however, several issues are not yet solved, especially for what concerns validation (Fig. 1). DIR is an ill-posed problem, since multiple solutions to the matching process may be found. The resulting DVFs are associated with a level of uncertainty that depends on several factors, such as the deformable algorithm, the image content, the presence of homogeneous regions, tissue changes, or physical fidelity of the deformation field itself.[48–50] Moreover, DIR algorithms and their validation are strongly case-specific and cannot be applied without a proper assessment for each clinical scenario. This recently motivated the publication of a task group AAPM TG132,[51] outlining the essential aspects of DIR for image guidance in radiotherapy. In this report the need to quantify the registration quality is highlighted, in order to be reliably applied in the clinical practice. Emphasis is especially posed on the use of physical and digital phantoms as tools for DIR geometric validation and commissioning. The need of a patient-specific evaluation is also reported. The aim is to provide a systematic patient-specific assessment of the registration quality, thus allowing the implementation of improved personalized treatments. However, the definition of a patient-specific gold standard still remains an open issue, with a qualitative analysis supported by quantitative contour and landmarks measures being the suggested strategy. Further concerns exist also for multimodal imaging and dose warping. For example, in case of functional image comparison at the voxel level for sequential PET/CT images, DIR is applied between CTs and the DVF used to warp PET volumes. However, DIR requires not only accurate tumor boundary registration, but also accurate registration of the interior tumor structure for a safe application of the DVF. The influence of the registration algorithm performance has been in fact demonstrated and different deformable algorithms can result in large voxel-by-voxel PET differences.[52] Similarly for dose warping, accuracy relies on DIR performance and errors can be introduced especially in regions of high dose gradients,[53–57] as for particle therapy where the treatment accuracy is significantly sensitive to geometrical and associated density variations. An accurate and efficient patient-specific assessment of DIR performances in terms of both geometric and dosimetric accuracy remains therefore elusive. The implementation of an appropriate workflow for patient-specific registration assessment would significantly increase the confidence in applying DIR to clinical practice.

This contribution aims at providing insights on the current state of the art of DIR validation strategies for its application in image-guided radiation therapy. More specifically, we will review the literature with the purpose to provide a prospective discussion on patient-specific DIR validation strategies. Challenges on DIR validation and caveats for accurate evaluation in the treatment workflow in terms of geometric and dosimetric accuracy will be overviewed and discussed. This aims at providing the groundwork for (a) boosting the implementation of novel automatic and robust patient-specific metrics and (b) providing preliminary guidelines for their use in application of DIR in adaptive image-guided radiotherapy.

Articles searching was performed with Scopus investigating terms "deformable image registration," "deformable image registration validation," "deformable image registration evaluation," "deformable image registration radiotherapy," "deformable image registration particle therapy," with a careful selection of patient-specific applications. We refined searches for peculiar issues with terms such as "dose warping," "dose accumulation," "virtual CT," "adaptive radiation therapy," "4D motion modeling" and combination of these. Only papers published in English between January 1997 and June 2018 were included.

2. THE PARADIGMS OF DIR VALIDATION

The development of DIR was motivated by the need to quantify organ motion in the framework of adaptive treatment strategies. However, from a practical perspective, the need to assure the quality of the registration is of primary importance for a safe and conscious use of DIR in the clinical practice. Similarly to the structure presented in Jaffray et al.[2] for IGRT, the great amount of DIR algorithms present in the literature (Fig. 1) can be considered as a divergence in DIR practice; however, all of these need to reconverge to enable clinical application, by means of the definition of common guidelines for DIR validation, especially on a patient-specific basis (Fig. 2). The ability of DIR to account for geometric changes has to be quantified accurately (the geometric accuracy paradigm) in both mono and multi-modal imaging, as well as the effects of DIR on dose mapping (the dosimetric accuracy paradigm). This can be achieved by available datasets and multi-institutional studies for benchmark definition, phantoms and tools derived from patient data, as well as metrics able to provide patient-specific accuracy measures. In this review, we will focus on patient-specific solutions. For additional details on current clinical guidelines, readers are referred to AAPM TG132.[51]

3. THE GEOMETRIC ACCURACY PARADIGM

Validation of DIR is known to be a challenging task because of the lack of a ground truth, which subsequently does not allow to define standardized means of evaluating the results of a DIR method. In a clinical setting, validation has indeed proven to be notoriously cumbersome, since anatomical and especially pathological variations are not readily included in a validation protocol.[12]

The lack of the known deformation for DIR assessment can be overcome with the use of physical or virtual phantoms, which have been deeply investigated in the AAPM TG132[51] for commissioning. However, even if phantoms or

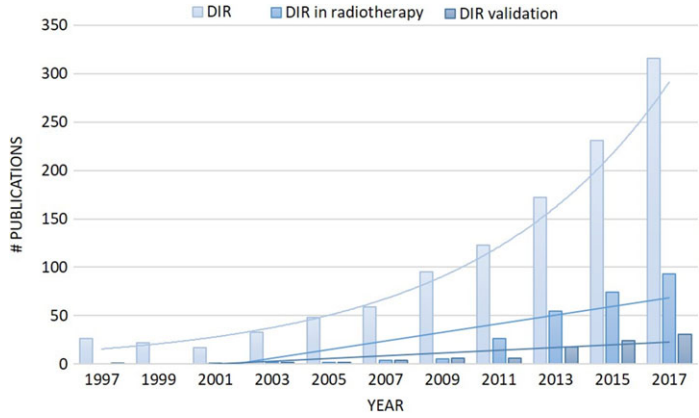

FIG. 1. Trend of publications for the search terms "DIR," "DIR in radiotherapy," and more specifically "DIR validation", from 1997 up to now.

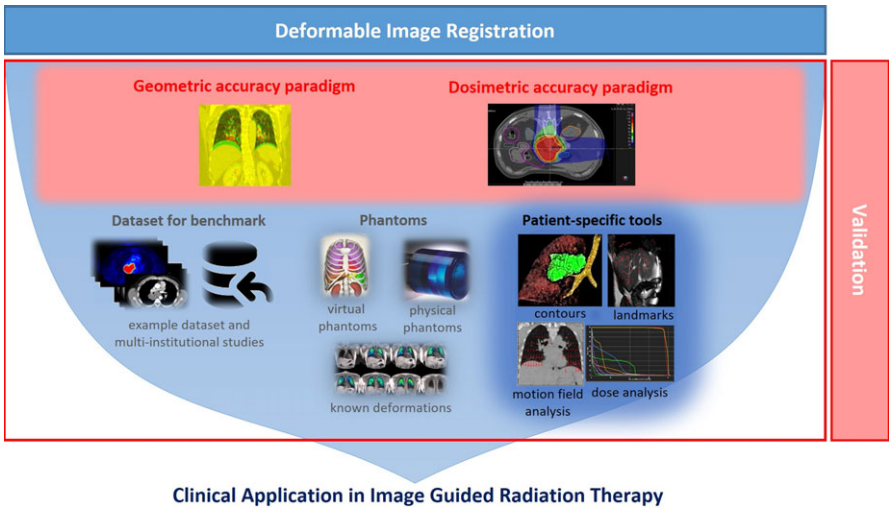

FIG. 2. The paradigms of DIR validation: geometric and dosimetric accuracy. Patient-specific tools for validation represent a fundamental aspect to boost the use of DIR in the clinical practice.

known deformations applied to patient data provide the ground truth transformation, this latter is typically an approximation of the patient-specific anatomo-pathological situation. To date, multi-institutional studies which rely on example dataset are a well-established approach to define benchmarks in DIR for different anatomical sites.[58–62] In these studies, different commercial and/or open-source registration algorithms are typically compared in a controlled framework, relying on both image-based or DVF-based metrics (as reported in the following sections). These comparative studies could be exploited for DIR evaluation and to provide tolerances for different anatomical sites. However, they require different registration algorithms, which are not often available in a clinical environment, and their selection could be affected by a systematic failure of the registration. Various DIR types and especially the class of biophysical and finite element modeling-based registration approaches should be in fact included in the comparison, since they could outperform purely intensity-based DIR, as shown for liver registration.[58]

In this Section we overview different methods proposed in the literature which could provide a patient-specific quantitative evaluation of DIR geometric accuracy.

3.A. Operator-dependent strategies: Image-based

The simplest patient-specific approach to evaluate DIR relies on the experience of clinicians, which visually assess the performance of the registration by means of overlay, checkboards,[28,51] or visualization features for exploration of candidate regions.[58] This, however, is far beyond a quantitative analysis for the definition of clinical guidelines. A possible way to quantitatively evaluate DIR is the manual selection of image surrogates, to be compared with an appropriate distance/similarity metric.

The most widespread standard metrics are typically established on manually contoured planning structures. Classical metrics based on contours rely on overlap and distance measures. Among these we can find the Dice Coefficient (DSC, percentage of contours' overlap), Mean Distance to Agreement (MDA, average distance between contours), Center of Mass distance (COM, distance between contours' centers of mass), and Hausdorff distance (maximum distance between contour points). A detailed definition of some of these metrics is present in Ref. [51], and their values need to be in the order of 0.80–0.90 and 2–3 mm (voxel dimension) for overlap and distance measures, respectively. Contour propagation and its validation with respect to ground truth contours have been largely adopted in both mono-modal and multi-modal imaging of different anatomical sites, with prevalence in head & neck,[22,32,59–64] thoraco-abdominal regions[14,17–19,31,34,65–69] and pelvis.[21,66,67,70,71] Specifically for multi-modal imaging, contour-based metrics are the most used methods for DIR evaluation, since they can provide an easy solution for the definition of corresponding anatomical (and functional) information. It has also been shown that in PET/CT registration, the SUV (Standardized Uptake Value) assessment in delineated contours, can provide further information on image alignment.[65]

Although contour propagation techniques seem to provide an efficient way of validation, they often do not confirm that the volume within the contour has been properly registered.[72] The volume dependence of DSC has been documented deeply.[22,51,73] As well, Rohlfing[74] showed that surrogate measures based on contour overlap are only weakly related to registration accuracy. In his study, a method was implemented with the aim to perform well on certain surrogate measures of registration performance, but completely disregarding any actual mapping of corresponding anatomical points. Of the tested criteria, only overlap of sufficiently small and localized labeled regions survives as a reliable discriminator between good and bad registrations, suggesting the need of a dense set of landmarks to gain a more complete and global understanding of registration accuracy.

Landmarks indeed represent another common way to evaluate DIR accuracy.[75] In order to avoid the invasiveness of implanted surrogates,[71] corresponding anatomical landmarks are usually extracted manually[15,27,34,76–82] and then used to compute their distance (or the so-called Target Registration Error, if landmarks are within the tumor).[51] Landmark techniques are, however, limited with poor image quality and contrast. Indeed, in homogeneous regions such as the abdomen for CT data, robust identification of landmark is challenging due to missing image structure.[83–85] A recent study, on a phantom demonstrated that if DIR performance is solely assessed with the contrast rich features present in clinical anatomy, the results may not be reflective of the true DIR performance in uniform low contrast anatomy.[86]

Even if all these strategies are the most widespread in the literature, they are, however, time-consuming and operator-dependent. Several studies investigated the inter- and intra-observer variability in manual contouring, highlighting the impact on the registration evaluation, and the subsequent need to account for it.[20,84,87] Automatic and efficient strategies are therefore effective alternatives to overcome operator-dependent variability and to decrease the workload in the clinical procedure.

3.B. Automatic strategies: Image-based

A simple automatic evaluation can be performed by means of global image similarity metrics based on image intensity, such as cross-correlation, mean squared error or mutual information between the reference volume and the deformed one.[64,74,88] Evaluations of the similarity measure (or its variants) in sub-regions was also exploited to provide a local measure of the registration error in both mono-modal[89] and multi-modal imaging.[90] In these studies, random variations of the DVF[89] or forms of interpolation[90] have been applied to add spatial information to the similarity metric, by relating similarity values to the local physical error. An alternative was proposed by Neylon et al.,[91] which relied on the construction of a neural network to translate the cost function values to an actual physical distance measure. A nonlinear

relationship between the similarity measure and the registration error was modeled by training the network on patient-specific model-generated data. Determining where and how to apply such networks should be an intense area of research, as their applications are wide-ranging and largely unexplored.

Semiautomatic and automatic contouring have been also investigated in the literature[92–94] and could be potentially an efficient alternative to manual delineation. However, it should be noticed that most of them rely on the accuracy of DIR in the contour generation (e.g., via atlases). More recent machine learning algorithms do not, but they lack patient-specific validation, and their generalization capabilities beyond the training dataset are uncertain.[93,94] Therefore the use of automatic contouring to validate DIR on a patient-specific basis is quite questionable and is often not able to match the accuracy of the expert clinician, which remains the universally acknowledged gold standard.[92] Moreover, it would still suffer of the same limitations of the above-mentioned contour-based metrics.[74]

Automatic landmarks extraction methods have been also proposed as an alternative to manual clicking, in order to provide a patient-specific quantification, to improve the clinical routine and to increase the reproducibility of results.[87,95] Murphy et al.[68,96] proposed a semiautomatic method for landmark extraction. A distinctiveness measure of each point with respect to its neighbors was implemented and corresponding matches were defined in a semiautomatic way. The same technique was adopted from Muenzing et al.,[97,98] which provided a supervised learning of local registration uncertainties, captured by statistical image features at distinctive landmark points. This technique, which require a training set for each new application and data set, was used to assess registration accuracy in longitudinal CT images of the lungs.

A fully automated method able to deal with scale changes has been also investigated in the literature, known as scale invariant feature transform (SIFT).[99,100] Preliminary studies analyzed the SIFT applicability and related benefits in medical imaging for DIR contour propagation.[101–103] Advantages of SIFT rely on the availability of invariance properties to different transformations and the definition of a feature descriptor for accurate landmark correspondence, which are two crucial aspects in medical imaging. This applies specifically to those cases where images coming from different temporal series of the same patient with anatomo-pathological changes have to be analyzed. SIFT was proposed as a method for DIR validation in head & neck CT for both radiotherapy[104] and particle therapy.[26] Recently, the applicability of the algorithm also to MRI data, where high contrasted anatomical structures are visible, has been also demonstrated.[105, 106]

As for manual landmarks, image quality, and tissue characteristics play a relevant role for robust landmark identification. For what concerns automatic landmarks, Paganelli et al.[107] provided an extension of SIFT with a local adaptive implementation in order to increase feature identification in tissues with low contrast. The extended SIFT was tested on lung 4DCT[107] (Fig. 3, panel A) and applied to abdominal 4DCT data treated with carbon ions,[108] quantifying DIR errors below the voxel resolution. Similarly, a large number of landmarks were also identified for a dense representation in head & neck CT, lung 4DCT, and pelvic MRI by means of improved landmark matching.[109] Limitations of these approaches are present in terms of computational cost, whereas advantages rely on using the dense set of landmarks to drive the DIR with subsequently improved results in the registration accuracy.[77,107]

Automatic landmark extraction has been successfully applied in intra-modal imaging, as for CT/CBCT,[26,104] whereas peculiar feature descriptors for landmark correspondence identification require to be defined and evaluated for more complex multi-modal imaging. In anatomical CT/MRI data, the modality independent descriptor[110] was proposed as similarity measure in the registration optimization process. It was based on the construction of an image descriptor by means of a self-similarity measure, which could be adopted for the definition of corresponding feature points. An extension of SIFT was also investigated for CT/MRI[111]; the proof-of-concept study was, however, limited to a 2-D implementation of the descriptor. To our knowledge, automatic landmark identification between anatomic and functional imaging has not been deeply investigated in the literature. The anatomical component coming from novel combined scanner such as PET/CT or PET/MRI could be therefore involved to drive DIR and to subsequently perform the validation. As suggested by Hwang et al.,[112] the registration of a PET/CT volume with treatment planning CT is preferable to a stand-alone PET.

3.C. Automatic strategies: DVF-based

Another automatic solution is to work directly on the nonrigid transformation, by analyzing how physically plausible the registration deformation is by means of DVF regularity indexes.[51,67,113] The Jacobian determinant is one of the possible metrics to compute as regularity index.[51] A determinant greater than 1 indicates expansion at that location, whereas a value below one indicates contraction. Negative values and large local changes may represent physically non-plausible deformations, thus indicating potential inaccuracies in the registration.[51] Exceptions exist when sliding at anatomical interfaces is modeled by the DIR algorithm, such as in biomechanical models: in this case, local values at the interface reflect DVF discontinuities due to sliding. Local nonlinear changes in the displacement field (e.g., vortexes) that do not correspond to underlying anatomical changes have been also quantified by the curl operator[114] and the harmonic energy (i.e., derived from the gradient of DVF).[72,115] In addition, transitivity error (TE) (i.e., difference between the composition of different transformations and an identity map) or the inverse consistency error (ICE) (i.e., difference between the composition of the forward and reverse transformations) are metrics that are typically computed.[51,72,116] Similarly, testing the reproducibility of the DVF by performing multiple registrations has been also proposed.[49]

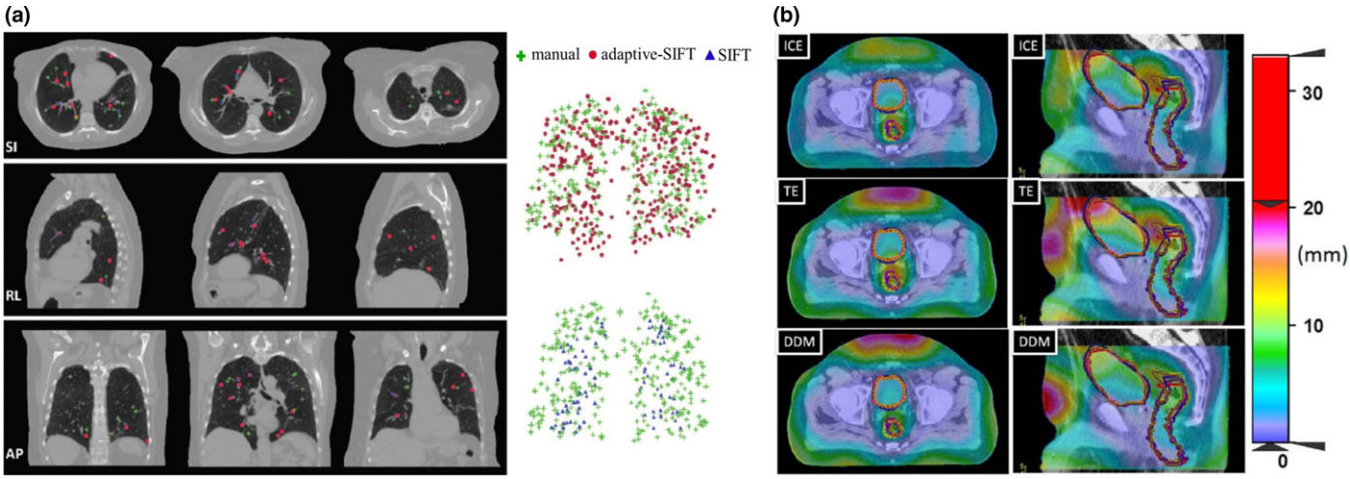

FIG. 3. Geometric accuracy paradigm. (a): image-based metric established on manual and automatic dense landmarks identification with SIFT for thoraco-abdominal sites [reprinted with permission from Ref. [107]]. (b): maps of DVF-based metrics (inverse consistence error (ICE), transitivity error (TE) and distance-discordance metric (DDM)) for the prostate site [reprinted with permission from Ref. [83]].

The main limitation of these approaches relies on the fact that a deformation field may fulfill requirements of physical fidelity, but still differ from the underlying ground truth of the deformation.[49] In addition, DVF indexes can only provide information about tissue expansion and shrinkage without conveying any information on DIR uncertainties.[83] The consistency check is in fact a necessary but not sufficient condition for an accurate deformation method and it has been shown to lack of a relation with the registration error.[56,74] Investigations on consistency error in a simulated framework showed comparable results between registration methods, even if better geometric correspondence was observed for specific algorithms.[72]

Other metrics related to the DVF were therefore proposed as alternative solutions. For prostate, a distance discordance metric (DDM) was recently proposed.[50,83] This resulted more correlated with the absolute registration error with respect to the ICE (Fig. 3, panel B). The DDM is based on the variability in the distance between corresponding voxels from different co-registered images. The method however requires at least four registered images to estimate the uncertainty of the DIR, which are not often available and/or need to be simulated. Alternatively, DVF could be substituted into a finite element-based elastic framework (i.e., biomechanical model) to calculate an energy metric which indicates the quality of the DVF in its neighborhood.[33,117] This method, however, requires the construction of a finite element model which could affect computational time and it is limited to homogeneous regions such as the prostate. Alternative biomechanical models have been proposed in the literature for DIR evaluation with improved efficiency[118] and capability to work on heterogeneous regions.[119] Bayesian and probabilistic methods have been also proposed to estimate registration uncertainty, but they do not provide a dense representation of the uncertainty of the DVF.

In a recent study,[115] some of these metrics have been compared in a controlled framework (i.e., simulated known deformations) to evaluate the most predictive DIR error metrics to identify voxels with a specific DIR error tolerance in head & neck and lung CT: it was shown that DDM and harmonic energy with thresholds of 0.49 mm and 0.014, respectively, can be used to identify voxels with DIR errors >2.0 mm.

An interesting approach has been recently proposed by Ribeiro et al.[120] Here the authors exploit the use of ground truth DVFs extracted from a 4DMRI to be compared to estimated DVFs computed from a 4DCT-MRI (i.e., 3DCT warped with DVFs extracted from the 4DMRI). The availability of a ground truth 4DMRI allowed to compare the geometric accuracy of six DIR algorithms (five commercially available and one research version) in terms of DVFs difference. Geometric differences of up to 1.0 mm for small motion amplitude and 3.2 mm for large motions have been observed in the liver. It should be noted that DIR induced errors are present in the ground truth 4D MRI motion. These are claimed to have a minimal impact on the 4DCT-MRI data set itself and to be limited due to the higher contrast in abdomen MRI than CT images. Also in this case, additional data (i.e., 4DMRI) are required.

4. THE DOSIMETRIC ACCURACY PARADIGM

Unrealistic warping not visible to standard voxel-based solution assessment can produce erroneous results when the DVF is applied on a secondary dataset, such as dose matrix.[114] The effect of DVF uncertainties on dose mapping is complex and depends on the spatial locations of both the DVF errors and the dose gradients.[53,57,121] An evaluation of the effect of DIR on dose warping is therefore required when such a strategy is used in either X-rays or particle therapy treatments, since the grid deformation due to DVF application does not necessarily define a one-by-one correspondence between voxels, thus affecting dose mapping.

As for geometric accuracy, the lack of a ground truth deformation can be overcome by means of physical or virtual phantoms.[42,122,123] Also, the comparison of different DIR techniques has been recently investigated in the literature to support the evidence of DIR uncertainties on dose warping. Indeed, despite similar results in terms of geometrical matching, different DIR techniques resulted in dose differences up 2% of the prescribed dose.[55,124] Comparative studies also highlighted that regions of higher dose gradient and poorer image contrast are more prone to larger variability in warped doses[55,125] and that dose deformation accuracy computed with dose-based metrics (similar to those presented in Section 4.A) are not correlated with DIR geometric accuracy.[120,126] Although these approaches allow to estimate dose uncertainties due to DIR as well as to provide tolerances for a clinical workflow,[123,125] they are not suitable for a patient-specific application.

In this Section, we report methods presented in the literature which could potentially provide information about dosimetric inaccuracies caused by DIR for a patient-specific evaluation and lay the groundwork for further developments.

4.A. Automatic strategies: Dose-based

Differently from the geometric paradigm, in which DIR accuracy can be directly computed with several metrics by comparing the reference and the registered (ideally overlapping) images, the effects of DIR on dose distribution is not trivial and the need of a ground truth represents the most viable solution. This is the case of studies exploiting the use of a virtual CT for adaptive treatment. Here, the "dose of the day" accuracy has been evaluated by comparing the recalculated dose on the deformed CT and the dose calculated on a replanning CT, which acted as ground truth. Dose-volume histograms (DVH), gamma analysis,[127] or dose difference are common metrics to compare the two doses. DVH represents the percentage or absolute volume receiving dose in the corresponding dose bin (differential DVH), or the percentage or absolute volume receiving a dose greater than or equal to the value in the corresponding bin (cumulative DVH). Gamma analysis instead compares two dose distributions determining on a voxel-by-voxel basis their local similarity, given spatial and dose difference acceptance criteria. Literature studies which exploited these metrics, showed that the recalculated dose properly matched the replanning one.[25,40,54] In head & neck CT/CBCT adaptive radiotherapy, Veiga et al.[25] reported a dose difference between the dose calculated on the virtual CT and the dose calculated on the replanning CT smaller than 2% of the prescribed dose on 90% of the patient's volume. The corresponding gamma pass rate (dose difference and distance to agreement criteria of 2%/2 mm) was 95%. Similar results were also found for virtual CT derived from megavoltage CT scans.[40] The evaluation of a water equivalent thickness (WET) uncertainty as a measure of the impact of DIR inaccuracies on dose calculation was also recently proposed for the use of virtual CT in passive scattering proton therapy of lung tumors.[24,54] The WET is the thickness of water needed to cause a proton beam to lose the same amount of energy as in a given thickness of a different medium. When comparing tumor margins of lung treatment plans as defined on the virtual and re-planning CTs, the resulting root mean squared uncertainty in WET was 3.3 ± 1.8 mm.[54] The WET metric is limited to particle therapy applications, but can provide a quantitative evaluation in these kind of treatments.

It should be, however, noticed that although replanning CTs could provide a ground truth, they are not always available in a clinical procedure and they represent the goal for the virtual CT approach.

In a similar fashion, the 4DCT-MRI concept[120] exploited the availability of a 4DMRI to derive the ground truth motion for the evaluation of optimized 4D dose distribution in liver proton therapy. Relying on this available data, a comparative framework was then proposed in which DVFs were estimated with different DIR algorithms and corresponding 4D optimized doses compared. Differences with respect to the dose calculation with ground truth DVFs in $V_{95\%}$ (volumes receiving the 95% of the dose) of the clinical target volume (target coverage) were quantified as high as $7.9 \pm 3.4\%$ and $11.3 \pm 12.5\%$ for small (mean displacement of liver points of 7.8 mm) and large (16.8 mm) motions, respectively, and no correlation was observed between geometric and dosimetric accuracies.

Other methods for a patient-and case-specific evaluation have been investigated without the need of a ground truth. A simplified quantification was defined by computing the difference in the D_{mean} (mean dose) within an arbitrary ROI before and after deformation to measure local dose mapping accuracy,[128] assuming that the mass and integral dose within any sub-volume are conserved. This, however, does not necessarily hold in presence of inter-fractional motion. Beyond D_{mean}, Moriya et al.[129] proposed the use of generalized Equivalent Uniform Dose (gEUD)[130,131] as accuracy index for intra-fractional dose accumulation in the lung. The gEUD is defined as the absorbed dose that, if homogeneously delivered to a tissue, causes the same radiobiological effect as the actual nonhomogeneous dose distribution. It is easily computed from the differential DVH considering tissue-specific parameters.

However, it has to be noted that metrics derived from DVH are sensitive to both organ delineation and contour propagation accuracy.[123] In head & neck CT/CBCT, García-Mollá et al.[113] tested dose metric sensitivity to contour propagation, by using the ICE property of DVF. This study demonstrated that any small DIR error greatly impacted on dose metrics of small-size structures, and that DIR accuracy in poor-contrast areas is reduced leading to dose differences delivered over 1 Gy in high dose gradient areas (prescription dose not provided).

Computing dose difference in corresponding anatomical landmarks could be also exploited as an alternative approach.[132,133] In CT of the lung, an average absolute

difference between planned dose values in manually tracked and in DIR warped features of 3.5 Gy (over a prescription dose ≥ 60 Gy) and average landmarks registration error of 5.2 mm were found, bearing the 4% margin of error to the treatment site that was expected during treatment planning.[133] Nevertheless, this solution is limited by the possibility to efficiently define landmarks within the irradiated area, which may encompass homogeneous low-contrasted tissues, as previously mentioned in Section 3.B.

An interesting approach able to link geometric uncertainty to dose mapping accuracy, has been implemented in the distance to dose difference (DTD) metric.[134] Given a dose distribution, the DTD is defined as the distance to observe a given dose difference in the irradiated geometry. Specifically, it indicates how large a DVF error can be before the DVF error could introduce a predetermined maximum tolerable dose mapping error. The DTD (Fig. 4, panel A) is adaptable based on the desired evaluation (e.g., defined for a given percentage of the local dose or maximum dose). With this metric, Saleh-Sayah et al.[134] demonstrated that to guarantee a dose mapping accuracy within 5% of the prescribed dose, DVF accuracy requirements are tight (~1 mm) in high dose gradient regions, whereas up to 10 mm could be tolerated in uniform dose regions. The DTD is therefore aimed at providing the DIR geometric accuracy required to satisfy dose accuracy constraints a priori and independently from any DVF; however, it is not usable for estimating the dosimetric uncertainty resulting from a specific measured geometric error.

4.B. Automatic strategies: DVF-based

Post-processing techniques on DIR motion field to improve ICE and reduce TE errors have been shown to provide a positive effect on the dose accumulation accuracy.[56,135] Also, the inclusion of dose-induced shrinkage of the tumor in the DIR workflow or refinement of DIR results with biomechanical model in homogeneous regions, can substantially improve the registration outcome and reduce dose warping uncertainties.[121,136] A probabilistic registration technique was instead described by Rishom et al.,[137] which permits to estimate both the cumulative radiation dose delivered to tissues and the corresponding dose uncertainty, visualized as error bound in DVH curves. However, these approaches require to be included in the registration algorithms from vendors if directly used in the clinical procedures, or eventually implemented as off-line evaluation tools.

A voxel-by-voxel estimation of dose mapping accuracy was also attempted. In this case several computational steps are often required, thus leading to more complex procedures compared to mono-dimensional dose-based metrics.

Hub et al.[138] introduced random DVF variations and measured dose mapping uncertainty through the maximum deviation of the mapped dose that was found among those random modifications that did not increase the local image similarity metric for DIR. Vickress et al.[139] defined the range of dose uncertainty as the maximum and minimum of the doses within a sphere around a voxel, defining the sphere radius

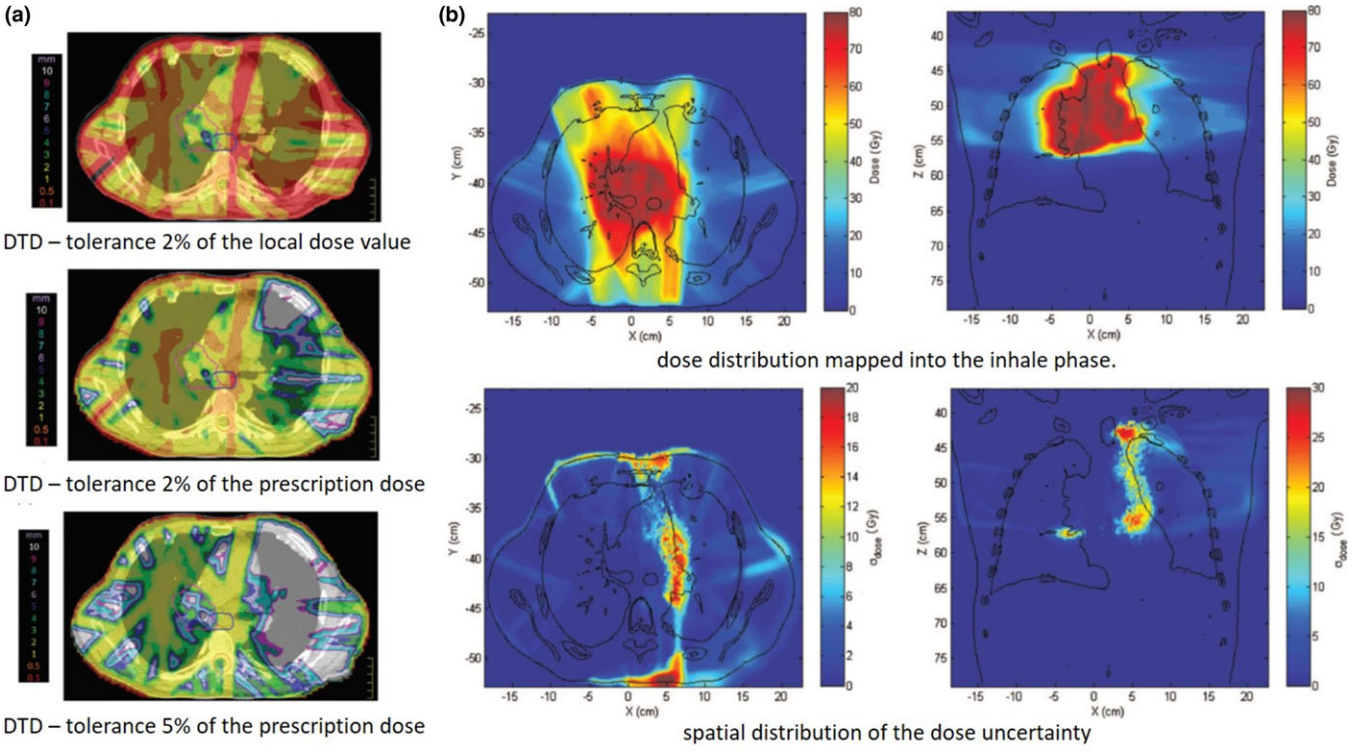

FIG. 4. Dosimetric accuracy paradigm. (a): Dose-based metric derived by means of the distance to dose difference (DTD) metric with tolerances with respect to the local or prescribed dose [reprinted with permission from Ref. [134]]. (b): DVF-based metric showing voxel-by-voxel representation of dosimetric uncertainty [reprinted with permission from Ref. [53]].

equal to a measure of DIR error. They found that the distance discordance metric[50] performs as a better predictor of dose uncertainty with respect to the ICE and the TE.

Several authors explored the idea of assuming a spatial model of DIR error in order to study its correlation with dose mapping accuracy. Spatial uncertainty models for DIR have been derived relying on the variance of the modules of inconsistency vectors or their covariance matrix.[53] These patient-specific models were then used to blur the dose map to obtain the spatial distribution of dose uncertainty (Fig. 4, panel B). Alternatively, DIR error maps were created from a training set of different DVF maps by means of principal component analysis and used to produce a set of mapped dose distributions.[57] The variance of these dose distributions revealed the pattern of dose mapping uncertainty arising from DVF uncertainties. Tilly et al.[140] created a different model, by sampling DIR errors at a sparse grid of control points which was then applied to the denser dose grid by means of 3-D cubic interpolation. They studied the impact of the dose mapping uncertainties on the radiobiological outcome metrics for spot scanned proton therapy of the prostate, concluding that an uncertainty of the dose to the 95% of the tumor volume ($D_{95\%}$) less than 3% and a tumor control probability uncertainty less than 2% would require a mean absolute DIR error better than 2.5 and 3.5 mm, respectively.

A recent approach proposed an automated DIR evaluation confidence tool (AUTODIRECT) for dose warping uncertainty estimation[141,142] which allows to predict voxel-specific dose mapping inaccuracies due to DIR on a patient-by-patient basis. The main limitation of the method rely on the availability and quality of different DIR algorithms used to model dose warping uncertainties, as AUTODIRECT can only accurately predict DIR errors for deformation scenarios that these generator algorithms can mimic.[141] This could potentially affect its use in a patient-specific clinical application, similarly to comparative studies.

5. RECOMMENDATIONS FOR ACCURATE DIR VALIDATION

In this Section we review the current clinical guidelines on DIR validation metrics, and we provide a focused analysis on the peculiarities that a metric has to satisfy for an accurate and efficient patient-specific validation. The roadmap for the definition of a DIR validation will be also outlined, thus providing the groundwork for the definition and implementation of novel robust and reliable patient-specific evaluation metrics to solve the geometric and dosimetric paradigms.

5.A. Current guidelines for DIR validation

As stated in AAPM TG132,[51] for initial commissioning of an image registration system, quantitative validation is required. A relevant role during commissioning is played by physical phantoms, which are needed for end-to-end tests and for validation of data transfer integrity as they can be physically imaged on various imaging devices and then images of these phantoms can be transferred through the image registration process, simulating the actual steps that are used for patient data processing. The major advantage of phantoms is that they can provide a ground truth which is not available when dealing with DIR in patients. Multi-modality (if necessary) phantoms with unambiguous internal landmarks and orientation should be imaged and used for validation. Digital phantoms also represent an additional step for use in commissioning and quality assurance programs for DIR accuracy tests. The TG132 reported specific digital phantom parameters with detailed tests and suggested tolerances, which could be used by the clinical end users as guidelines to test a DIR algorithm. Additional quantification should be also performed on example clinical datasets, designed on the clinical protocols, by means of qualitative and quantitative metrics (contours, landmarks and DVF). The magnitude of the registration uncertainties (geometric error) resulting from these tests should be incorporated in the definition of clinical margins.

All these solutions are, however, far beyond a patient-specific analysis. Indeed, for patient-specific evaluation of image registration, quantitative verification is not always possible due to limited time and resources and difficulty in determining the ground truth. In the TG132, it is therefore reported that qualitative evaluation of the image registration should always be performed in the routine clinical practice to ensure acceptability of the registration. In addition to this, quantitative metrics could be computed to complement qualitative evaluation. This could also support quality check as well as the definition of patient-related documentation and report. Among the different annotations that the TG132 suggests, we can find (a) images and techniques used to perform the registration, (b) uncertainties in the final registration for local regions of importance and anatomical landmarks, (c) verification of acceptable tolerances, which should be within the maximum voxel dimension (in case of DSC and Jacobian, 0.8/0.9 and non-negative values, respectively). However, the need to establish a patient-specific quality assurance practice is also reported as a clinical recommendation, putting forward the demand for robust an efficient metrics for the evaluation of image registration results to address both geometric and dosimetric paradigms.

5.B. Toward efficient patient-specific guidelines

5.B.1. Criteria to be satisfied by patient-specific metrics

Patient-specific validation tools should cover the geometric and dosimetric paradigms to provide effective metrics that could be useful in the clinical procedures. Relevant criteria that should be met are as follows:

(i) *Spatial information*. The most widespread methods for image registration assessment relies on manual contours, clicked landmarks, or analysis of the DVF, which can

provide a support for clinical applications. Manual contours and landmarks allow a direct measure of DIR errors, whereas DVF-based metrics do not provide a comprehensive estimation of the spatial accuracy of DIR. This steers future research toward the development of more sophisticated measures (e.g., DDM or biomechanical approaches), which, however, have to satisfy the demand for an efficient clinical use [criterion (ii)]. Additionally, it has been shown that contours suffer of a weakly correlation with registration accuracy. This suggests the use of a wide and dense set of landmarks as a promising approach for DIR validation. As for the dosimetric impact of DIR uncertainty, it should be noted that greater accuracy is needed in correspondence to dose gradients, as a small geometric error might have a severe impact on warped doses. The distance to dose difference (DTD)[134] could be a potential tool for an a priori evaluation of the needed DIR accuracy to guarantee limited dosimetric uncertainty, whereas DVF-based methods could be exploited to define maps of dose uncertainty associated to a certain DIR result. The availability of a 4DMRI to generate a ground truth motion field[120] could provide a dense image feature set for geometric and dosimetric evaluations and needs to be further investigated.

(ii) *Automatic and efficient implementation*. Manual contour and landmarks resulted to be time-consuming and operator-dependent. Automatic methods have been implemented in the literature and used for geometric accuracy evaluation. In the case of automatic landmarks identification, stable and robust detection of anatomical features is required to ensure an effective utility. Dense landmarks sets can be automatically detected at the cost of computational efficiency. DVF-based metrics such as DDM can provide a better evaluation of the registration error, however, they require multiple images to estimate the error for an individual registration. Similarly, DVF-based frameworks for the definition of dose uncertainty maps are expected to require a greater effort for efficient implementation compared to simple dose-based metrics. When patient data are available and can act as ground truth, it is suggested to exploit and include them in the patient-specific evaluation. However, attention is required for the use/implementation of patient-specific metrics which are independent from the available data, automatic, easy, and efficient for the integration in a clinical environment.

(iii) *Description of un-contrasted or homogeneous regions, which suffer from poor anatomical details*. In this scenario, the application of image preprocessing or tissue enhancement could allow a better identification of internal structures. Implementation of dedicated feature extraction methods could be also considered, as well as the use of DVF-based metrics. Integration of information coming from different image modalities could be taken into account, as shown for 4DMRI, which could lead to better motion estimations in regions encompassing soft tissue such as the abdomen with respect to 4DCT. However, special care should be used when applying a registration result to different images, as reported in TG132.[51] As a good geometrical matching is a necessary but not sufficient condition for dose warping accuracy, it is important to carefully evaluate DIR in these regions, especially when they encompass dose gradients. The use of techniques for improving DIR outcome, like regularization methods based on landmarks and DVF properties or the integration of biomechanical methods, are encouraged and they could provide a positive effect on mapped dose accuracy. However, maps of dose uncertainty should be able to highlight the risk of erroneous warping more than general metrics, such as D_{mean} and gEUD.

(iv) *Multi-modality*. In this case, the extension of automatic landmark extraction methods for multi-modal imaging is, however, still a challenge and particular attention should be paid when different information is involved such as anatomical vs. functional data (e.g., CT with PET or CT with functional MRI). Contours for these cases could provide a preliminary assessment and analysis on DVF could play an important function in the evaluation procedure. Moreover, it should be noted that any error in Hounsfield unit mapping from the reference CT to the evaluation intra- or multi-modal imaging would result in dose misestimation, which would increase the uncertainty in accumulated dose on the reference geometry. In the context of virtual CT generation, authors reported techniques validation through the use of re-planning CTs for ground truth dose estimation; nevertheless, this approach is often unfeasible due to the lack of data. Similarly, the use of a 4DCT-MRI could lead to ground truth motion estimation for geometric and dosimetric evaluation, despite being still affected by DIR errors.

(v) *Capability to relate geometric with dosimetric inaccuracies*. In this scenario, the quantified geometric DIR errors due to a failure of the registration should be distinguished by DIR uncertainties due to limited performance of the registration [e.g., criterion (iii)], when assessing the dosimetric impact. However, the procedure to achieve such a distinction is not yet well-defined in the literature. The comparison of different DIR techniques is an established method for geometric and dosimetric evaluation. To date, however, this approach did not find a strict correlation of the sensitivity to DIR performance of clinically relevant metrics, such as DVH parameters, to geometric error/uncertainty. Moreover, it should not be neglected that this path is limited for a patient-specific application and potentially affected by systematic errors in the selected DIR algorithms. DVF-based solutions are promising methods to provide a more reliable patient-specific dose mapping evaluation, although their estimation depend on the robustness and validity of hypothesized DIR error models. An interesting approach is the DTD tool, which gives a priori indications on DVF accuracy required to gain a certain dose accuracy, however, it is not aimed at estimating the dose uncertainty associated to a computed geometric uncertainty. Therefore, to solve this criterion, the investigation of specific metrics able to provide more precise quantitative assessment of the relationship between geometric and dosimetric inaccuracies is still required.

(vi) *Definition of acceptable tolerances*. It should be verified that tolerances defined for both geometric and dosimetric

metrics are valid, in order to guarantee a negligible effect in the patient-specific evaluation of treatment outcome. As defined in TG132,[51] tolerances on the geometric accuracy should be considered (e.g., maximum voxel dimension for landmark distance and acceptable contours overlap). However, these tolerances should be evaluated on a patient-specific basis according to the clinical scenario and the anatomical site, with a dense spatial evaluation required on the whole imaging volume [as reported in criteria (i) and (iii)]. Tolerances should be also reflected in the dosimetric accuracy by defining clinically relevant thresholds, in relation to dose prescription reported in ICRU guidelines[143] for tumor and organs at risk. However, due to a limited number of references, the definition of clinical tolerances on the dosimetric paradigm is still challenging. Further investigations are therefore needed, and well-designed multi-institutional and comparative studies encouraged to provide a groundwork for the definition of appropriate guidelines. Clinical studies are also recommended, in which DIR and dose warping are validated according to specific tolerance limits and then used in parallel to the current clinical workflow. This would allow one to assess the difference between the accumulated dose and the planned dose, as well as the relative clinical outcome.[125]

5.B.2. Roadmap of guidelines for patient-specific DIR validation

Research and developments in DIR for IGRT during the past decades provided useful solutions, especially for what concerns organ motion quantification, mono- and multi-modal image fusion, as well as adaptive radiotherapy. Several algorithms have already been proposed in the literature, thus providing a strong background for DIR application in the clinical workflow. All the necessary elements are available, as demonstrated by the great amount of open-source software, as well as the inclusion of DIR in current commercially available clinical workstations. However, the roadmap of guidelines for an accurate and efficient patient-specific DIR validation in a clinical setting (Fig. 5) is not yet well established.

From the above considerations, a dense set of anatomical landmarks could be recommended for the solution of the geometric accuracy paradigm, supporting both mono-modal and multi-modal image registration. This should be complemented by additional evaluations on contours or deformation field analysis, when landmarks lack of a proper identification or anatomical description, like in low-contrast sites. In the clinical context, automatic and efficient approaches should be considered to avoid operator-dependent and time-consuming manual solutions. Solving the dosimetric accuracy paradigm seems instead to be a more challenging task, with a preliminary recommended guideline based on the estimation of the DIR-related uncertainty of clinically relevant metrics. Current dose-based solutions and DVF-based methods relying on models of DIR error could lay the groundwork for the development of more reliable metrics. As reported by the TG132, available datasets, physical, or digital phantoms tools are required for commissioning and play a key role to complement and overcome limitations of the above-mentioned approaches. However, the overall effects of deformation vary

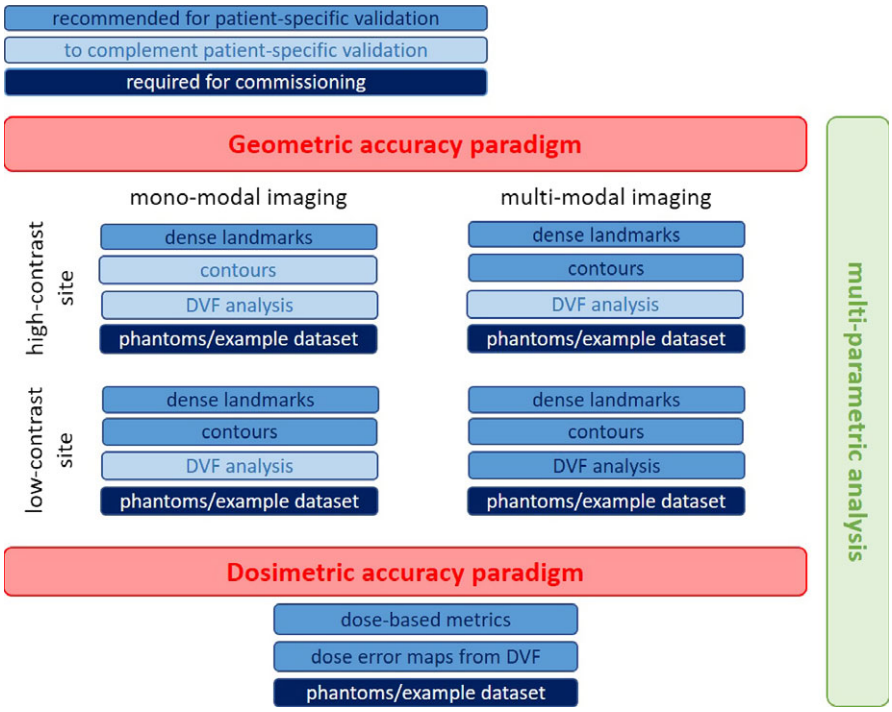

FIG. 5. The DIR validation roadmap. Recommended and complementing analysis to solve patient-specific geometric and dosimetric accuracy paradigms, integrated with required phantoms and example dataset for DIR commissioning.

among patients, depending on tumor location, field size, volume expansion, tissue heterogeneity, and direction of tumor displacement with respect to the beam. This suggests a preferential study toward patient-specific metrics able to derive DIR accuracy. Further research is also needed to investigate the benefits of increasing the redundancy of validation metrics for a multi-parametric analysis able to cover the whole aspects of DIR outcome.

CONFLICT OF INTEREST

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

[a)]Author to whom correspondence should be addressed. Electronic mail: chiara.paganelli@polimi.it; Telephone: +39 02 2399 9021; website: www.ca rtcas.polimi.it.

REFERENCES

1. Dawson LA, Sharpe MB. Image-guided radiotherapy: rationale, benefits, and limitations. *Lancet Oncol*. 2006;7:848–858.
2. Jaffray DA. Image-guided radiotherapy: from current concept to future perspectives. *Nat Rev Clin Oncol*. 2012;9:688.
3. Baumann M, Krause M, Overgaard J, et al. Radiation oncology in the era of precision medicine. *Nat Rev Cancer*. 2016;16:234.
4. Linz U. *Ion Beam Therapy: Fundamentals, Technology, Clinical Applications*. Berlin: Springer Science & Business Media; 2011.
5. O'Shea T, Bamber J, Fontanarosa D, et al. Review of ultrasound image guidance in external beam radiotherapy part II: intra-fraction motion management and novel applications. *Phys Med Biol*. 2016;61:R90.
6. Parodi K. Vision 20/20: positron emission tomography in radiation therapy planning, delivery, and monitoring. *Med Phys*. 2015;42:7153–7168.
7. Schmidt MA, Payne GS. Radiotherapy planning using MRI. *Phys Med Biol*. 2015;60:R323.
8. Bainbridge H, Salem A, Tijssen RHN, et al. Magnetic resonance imaging in precision radiation therapy for lung cancer. *Transl Lung Cancer Res*. 2017;6:689.
9. Raaymakers B, Jürgenliemk-Schulz IM, Bol GH, et al. First patients treated with a 1.5 T MRI-Linac: clinical proof of concept of a high-precision, high-field MRI guided radiotherapy treatment. *Phys Med Biol*. 2017;62:L41.
10. Maintz JA, Viergever MA. A survey of medical image registration. *Med Image Anal*. 1998;2:1–36.
11. Sarrut D. Deformable registration for image-guided radiation therapy. *Zeitschrift für medizinische Physik*. 2006;16:285–297.
12. Viergever MA, Maintz JBA, Klein S, Murphy K, Staring M, Pluim JPW. A survey of medical image registration–under review. *Med Image Anal*. 2016;33:140–144.
13. Oh S, Kim S. Deformable image registration in radiation therapy. *Radiat Oncol J*. 2017;35:101.
14. Balik S, Weiss E, Jan N, et al. Evaluation of 4-dimensional computed tomography to 4-dimensional cone-beam computed tomography deformable image registration for lung cancer adaptive radiation therapy. *Int J Radiat Oncol Biol Phys*. 2013;86:372–379.
15. Yang Y, Teo SK, Van Reeth E, et al. A hybrid approach for fusing 4D-MRI temporal information with 3D-CT for the study of lung and lung tumor motion. *Med Phys*. 2015;42:4484–4496.
16. McClelland JR, Hawkes DJ, Schaeffter T, King AP. Respiratory motion models: a review. *Med Image Anal*. 2013;17:19–42.
17. Liu Y, Jin R, Chen M, et al. Contour propagation using non-uniform cubic B-splines for lung tumor delineation in 4D-CT. *Int J Comput Assit Radiol Surg*. 2016;11:2139–2151.
18. Peroni M, Spadea MF, Riboldi M, et al. Validation of automatic contour propagation for 4D treatment planning using multiple metrics. *Technol Cancer Res Treat*. 2013;12:501–510.
19. Hardcastle N, van Elmpt W, De Ruyscher D, Bzdusek K, Tomé WA. Accuracy of deformable image registration for contour propagation in adaptive lung radiotherapy. *Radiat Oncol*. 2013;8:243.
20. Riegel AC, Antone JG, Zhang H, et al. Deformable image registration and interobserver variation in contour propagation for radiation therapy planning. *J Appl Clin Med Phys*. 2016;17:347–357.
21. Gardner SJ, Wen N, Kim J, et al. Contouring variability of human-and deformable-generated contours in radiotherapy for prostate cancer. *Phys Med Biol*. 2015;60:4429.
22. Kumarasiri A, Siddiqui F, Liu C, et al. Deformable image registration based automatic CT-to-CT contour propagation for head and neck adaptive radiotherapy in the routine clinical setting. *Med Phys*. 2014;41:121712.
23. Peroni M, Ciardo D, Spadea MF, et al. Automatic segmentation and online virtualCT in head-and-neck adaptive radiation therapy. *Int J Radiat Oncol Biol Phys*. 2012;84:e427–e433.
24. Veiga C, Janssens G, Teng CL, et al. First clinical investigation of cone beam computed tomography and deformable registration for adaptive proton therapy for lung cancer. *Int J Radiat Oncol Biol Phys*. 2016;95:549–559.
25. Veiga C, McClelland J, Moinuddin S, et al. Toward adaptive radiotherapy for head and neck patients: feasibility study on using CT-to-CBCT deformable registration for "dose of the day" calculations. *Med Phys*. 2014;41:031703.
26. Landry G, Nijhuis R, Dedes G, et al. Investigating CT to CBCT image registration for head and neck proton therapy as a tool for daily dose recalculation. *Med Phys*. 2015;42:1354–1366.
27. Yang D, Chaudhari SR, Goddu SM, et al. Deformable registration of abdominal kilovoltage treatment planning CT and tomotherapy daily megavoltage CT for treatment adaptation. *Med Phys*. 2009;36:329–338.
28. Mattes D, Haynor DR, Vesselle H, Lewwllen TK, Eubank W. PET-CT image registration in the chest using free-form deformations. *IEEE Trans Med Imaging*. 2003;22:120–128.
29. Kovalchuk N, Jalisi S, Subramaniam RM, Truong MT. Deformable registration of preoperative PET/CT with postoperative radiation therapy planning CT in head and neck cancer. *Radiographics*. 2012;32:1329–1341.
30. Gianoli C, Riboldi M, Fontana G, et al. Optimized PET imaging for 4D treatment planning in radiotherapy: the virtual 4D PET strategy. *Technol Cancer Res Treat*. 2015;14:99–110.
31. Hardcastle N, Hofman MS, Hicks RJ, et al. Accuracy and utility of deformable image registration in 68 Ga 4D PET/CT assessment of pulmonary perfusion changes during and after lung radiation therapy. *Int J Radiat Oncol Biol Phys*. 2015;93:196–204.
32. 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;52:1353–1359.
33. Zhong H, Wen N, Gordon JJ, Elshaikh MA, Movsas B, Chetty IJ. An adaptive MR-CT registration method for MRI-guided prostate cancer radiotherapy. *Phys Med Biol*. 2015;60:2837.
34. Fernandez-de-Manuel L, Wollny G, Kybic J, et al. Organ-focused mutual information for nonrigid multimodal registration of liver CT and Gd–EOB–DTPA-enhanced MRI. *Med Image Anal*. 2014;18:22–35.
35. Boye D, Lomax T, Knopf A. Mapping motion from 4D-MRI to 3D-CT for use in 4D dose calculations: a technical feasibility study. *Med Phys*. 2013;40:061702.
36. Jaffray DA, Lindsay PE, Brock KK, Deasy JO, Tomé WA. Accurate accumulation of dose for improved understanding of radiation effects in normal tissue. *Int J Radiat Oncol Biol Phys*. 2010;76:S135–S139.
37. Velec M, Moseley JL, Eccles CL, et al. Effect of breathing motion on radiotherapy dose accumulation in the abdomen using deformable registration. *Int J Radiat Oncol Biol Phys*. 2011;80:265–272.
38. Rosu M, Hugo GD. Advances in 4D radiation therapy for managing respiration: part II–4D treatment planning. *Zeitschrift für Medizinische Physik*. 2012;22:272–280.
39. Chang JY, Zhang X, Knopf A, et al. Consensus guidelines for implementing pencil-beam scanning proton therapy for thoracic

malignancies on behalf of the PTCOG thoracic and lymphoma subcommittee. *Int J Radiat Oncol Biol Phys*. 2017;99:41–50.
40. Branchini M, Fiorino C, Dell'Oca I, et al. Validation of a method for "dose of the day" calculation in head-neck tomotherapy by using planning ct-to-MVCT deformable image registration. *Phys Med: Eur J Med Phys*. 2017;39:73–79.
41. Bernatowicz K, Peroni M, Perrin R, Weber DC, Lomax A. Four-dimensional dose reconstruction for scanned proton therapy using liver 4DCT-MRI. *Int J Radiat Oncol Biol Phys*. 2016;95:216–223.
42. Moteabbed M, Sharp GC, Wang Y, Trofimov A, Efstathiou JA, Lu HM. Validation of a deformable image registration technique for cone beam CT-based dose verification. *Med Phys*. 2015;42:196–205.
43. Heath E, Seco J, Wu Z, Sharp GC, Paganetti H, Seuntjens J. *A comparison of dose warping methods for 4D Monte Carlo dose calculations in lung*. in Journal of Physics: Conference Series; 2008. IOP Publishing.
44. Heath E, Seuntjens J. A direct voxel tracking method for four-dimensional Monte Carlo dose calculations in deforming anatomy. *Med Phys*. 2006;33:434–445.
45. Siebers JV, Zhong H. An energy transfer method for 4D Monte Carlo dose calculation. *Med Phys*. 2008;35:4096–4105.
46. Zhong H, Siebers JV. Monte Carlo dose mapping on deforming anatomy. *Phys Med Biol*. 2009;54:5815.
47. Zhang Y, Huth I, Wegner M, Weber DC, Lomax AJ. An evaluation of rescanning technique for liver tumour treatments using a commercial PBS proton therapy system. *Radiother Oncol*. 2016;121:281–287.
48. Sotiras A, Davatzikos C, Paragios N. Deformable medical image registration: a survey. *IEEE Trans Med Imaging*. 2013;32:1153–1190.
49. Hub M, Karger C. Estimation of the uncertainty of elastic image registration with the demons algorithm. *Phys Med Biol*. 2013;58:3023.
50. Saleh ZH, Apte AP, Sharp GC, et al. The distance discordance metric —a novel approach to quantifying spatial uncertainties in intra-and inter-patient deformable image registration. *Phys Med Biol*. 2014;59:733.
51. 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:e43–e76.
52. Yip S, Chen AB, Aerts HJ, Berbeco R. Sensitivity study of voxel-based PET image comparison to image registration algorithms. *Med Phys*. 2014;41:111714.
53. Salguero FJ, Saleh-Sayah NK, Yan C, Siebers JV. Estimation of three-dimensional intrinsic dosimetric uncertainties resulting from using deformable image registration for dose mapping. *Med Phys*. 2011;38:343–353.
54. Veiga C, Veiga C, Janssens G, et al. A comprehensive evaluation of the accuracy of CBCT and deformable registration based dose calculation in lung proton therapy. *Biomed Phys Eng Expr*. 2017;3:015003.
55. Veiga C, Lourenço AM, Mouinuddin S, et al. Toward adaptive radiotherapy for head and neck patients: uncertainties in dose warping due to the choice of deformable registration algorithm. *Med Phys*. 2015;42:760–769.
56. Bender ET, Hardcastle N, Tomé WA. On the dosimetric effect and reduction of inverse consistency and transitivity errors in deformable image registration for dose accumulation. *Med Phys*. 2012;39:272–280.
57. Murphy MJ, Salguero FJ, Seibers JV, Staub D, Vaman C. A method to estimate the effect of deformable image registration uncertainties on daily dose mapping. *Med Phys*. 2012;39:573–580.
58. Schlachter M, Fechter T, Jurisic M, et al. Visualization of deformable image registration quality using local image dissimilarity. *IEEE Trans Med Imaging*. 2016;35:2319–2328.
59. Broggi S, Scalco E, Belli ML, et al. A comparative evaluation of 3 different free-form deformable image registration and contour propagation methods for head and neck MRI: the case of parotid changes during radiotherapy. *Technol Cancer Res Treat*. 2017;16:373–381.
60. Castadot P, Lee JA, Parraga A, Geets X, Macq B, Grégoire V. Comparison of 12 deformable registration strategies in adaptive radiation therapy for the treatment of head and neck tumors. *Radiother Oncol*. 2008;89:1–12.
61. Fortin D, Basran PS, Berrang T, Peterson D, Wai ES. Deformable versus rigid registration of PET/CT images for radiation treatment planning of head and neck and lung cancer patients: a retrospective dosimetric comparison. *Radiat Oncol*. 2014;9:50.
62. Hardcastle N, Tomé WA, Cannon DM, et al. A multi-institution evaluation of deformable image registration algorithms for automatic organ delineation in adaptive head and neck radiotherapy. *Radiat Oncol*. 2012;7:90.
63. Zukauskaite R, Brink S, Hansen CR, et al. Open source deformable image registration system for treatment planning and recurrence CT scans: Validation in the head and neck region. *Strahlenther Onkol*. 2016;192:545–551.
64. Nobnop W, Neamin H, Chitapanarux I, Wanwilairat S, Lorvidhaya V, Sanghangthum T. Accuracy of eight deformable image registration (DIR) methods for tomotherapy megavoltage computed tomography (MVCT) images. *J Med Radiat Sci*. 2017;64:290–298.
65. Kerner GS, Fischer A, Koole MJ, Pruim J, Groen HJ. Evaluation of elastix-based propagated align algorithm for VOI-and voxel-based analysis of longitudinal 18 F-FDG PET/CT data from patients with non-small cell lung cancer (NSCLC). *EJNMMI Res*. 2015;5:15.
66. Velec M, Moseley JL, Svensson S, et al. Validation of biomechanical deformable image registration in the abdomen, thorax and pelvis in a commercial radiotherapy treatment planning system. *Med Phys*. 2017;44:3407–3417.
67. Brock KK. Results of a multi-institution deformable registration accuracy study (MIDRAS). *Int J Radiat Oncol Biol Phys*. 2010;76:583–596.
68. Murphy K, van Ginneken B, Reinhardt JM, et al. Evaluation of registration methods on thoracic CT: the EMPIRE10 challenge. *IEEE Trans Med Imaging*. 2011;30:1901–1920.
69. Wijesooriya K, Weiss J, Dill V, et al. Quantifying the accuracy of automated structure segmentation in 4D CT images using a deformable image registration algorithm. *Med Phys*. 2008;35:1251–1260.
70. Fallone BG, Rivest DR, Riauka TA, Murtha AD. Assessment of a commercially available automatic deformable registration system. *J Appl Clin Med Phys*. 2010;11:101–123.
71. Kim J, Kumar S, Liu C, et al. A novel approach for establishing benchmark CBCT/CT deformable image registrations in prostate cancer radiotherapy. *Phys Med Biol*. 2013;58:8077.
72. 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.
73. Deeley M, Chen A, Datteri R, et al. Comparison of manual and automatic segmentation methods for brain structures in the presence of space-occupying lesions: a multi-expert study. *Phys Med Biol*. 2011;56:4557.
74. Rohlfing T. Image similarity and tissue overlaps as surrogates for image registration accuracy: widely used but unreliable. *IEEE Trans Med Imaging*. 2012;31:153–163.
75. Hoffmann L, Holt MI, Knap MM, Khalil AA, Møller DS. Anatomical landmarks accurately determine interfractional lymph node shifts during radiotherapy of lung cancer patients. *Radiother Oncol*. 2015;116:64–69.
76. Sarrut D, Delhay B, Villard PF, Boldea V, Beuve M, Clarysse P. A comparison framework for breathing motion estimation methods from 4-D imaging. *IEEE Trans Med Imaging*. 2007;26:1636–1648.
77. Shusharina N, Sharp G. Analytic regularization for landmark-based image registration. *Phys Med Biol*. 2012;57:1477.
78. Castillo R, Castillo E, Guerra R, et al. A framework for evaluation of deformable image registration spatial accuracy using large landmark point sets. *Phys Med Biol*. 2009;54:1849.
79. 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;35:775–781.
80. Akbarzadeh A, Gutierrez D, Baskin A, et al. Evaluation of whole-body MR to CT deformable image registration. *J Appl Clin Med Phys*. 2013;14:238–253.
81. Castillo E, Castillo R, Martinez J, et al. Four-dimensional deformable image registration using trajectory modeling. *Phys Med Biol*. 2009;55:305.
82. Cazoulat G, Owen D, MAtuszak MM, Balter JM, Brock KK. Biomechanical deformable image registration of longitudinal lung CT images using vessel information. *Phys Med Biol*. 2016;61:4826.

83. Saleh Z, Thor M, Apte AP, et al. A multiple-image-based method to evaluate the performance of deformable image registration in the pelvis. *Phys Med Biol*. 2016;61:6172.
84. Mencarelli A, van Kranen SR, 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;90:680–687.
85. Liu F, Hu Y, Zhang Q, Kincaid R, Goodman KA, Mageras GS. Evaluation of deformable image registration and a motion model in CT images with limited features. *Phys Med Biol*. 2012;57:2539.
86. Varadhan R, Magome T, Hui S. Characterization of deformation and physical force in uniform low contrast anatomy and its impact on accuracy of deformable image registration. *Med Phys*. 2016;43:52–61.
87. 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;39:6879–6884.
88. Heath E, Collins DL, Keall PJ, Dong L, Seuntjens J. Quantification of accuracy of the automated nonlinear image matching and anatomical labeling (ANIMAL) nonlinear registration algorithm for 4D CT images of lung. *Med Phys*. 2007;34:4409–4421.
89. Hub M, Kessler ML, Karger CP. A stochastic approach to estimate the uncertaintyinvolved in b-spline image registration. *IEEE Trans Med Imaging*. 2009;28:1708–1716.
90. Nix MG, Prestwich RJ, Speight R. Automated, reference-free local error assessment of multimodal deformable image registration for radiotherapy in the head and neck. *Radiother Oncol*. 2017;125:478–484.
91. Neylon J, Min Y, Low DA, Santhanam A. A neural network approach for fast, automated quantification of DIR performance. *Med Phys*. 2017;44:4126–4138.
92. Whitfield GA, Price P, Price GJ, Moore CJ. Automated delineation of radiotherapy volumes: are we going in the right direction? *Br J Radiol*. 2013;86:20110718–20110718.
93. Lustberg T, van Soest J, Gooding M, et al. Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer. *Radiother Oncol*. 2018;126:312–317.
94. Sharp G, Fritscher KD, Pekar V, et al. Vision 20/20: perspectives on automated image segmentation for radiotherapy. *Med Phys*. 2014;41:050902.
95. Frantz S, Rohr K, Stiehl HS. Development and validation of a multistep approach to improved detection of 3D point landmarks in tomographic images. *Image Vis Comput*. 2005;23:956–971.
96. Murphy K, van Ginneken b, Klein S, et al. Semi-automatic construction of reference standards for evaluation of image registration. *Med Image Anal*. 2011;15:71–84.
97. Muenzing SE, van Ginneken B, Murphy K, Pluim JP. Supervised quality assessment of medical image registration: application to intra-patient CT lung registration. *Med Image Anal*. 2012;16:1521–1531.
98. Muenzing SE, van Ginneken B, Viergever MA, Pluim JP. DIRBoost–an algorithm for boosting deformable image registration: application to lung CT intra-subject registration. *Med Image Anal*. 2014;18:449–459.
99. Lowe DG. Distinctive image features from scale-invariant keypoints. *Int J Comput Vis*. 2004;60:91–110.
100. Cheung W, Hamarneh G. N-SIFT: N-dimensional scale invariant feature transform. *IEEE Trans Image Process*. 2009;18:2012–2021.
101. Xie Y, Chao M, Lee P, Xing L. Feature-based rectal contour propagation from planning CT to cone beam CT. *Med Phys*. 2008;35:4450–4459.
102. Xie Y, Chao M, Xiong G. Deformable image registration of liver with consideration of lung sliding motion. *Med Phys*. 2011;38:5351–5361.
103. Allaire S, Kim JJ, Breen SL, Jaffray DA, Pekar V. *Full orientation invariance and improved feature selectivity of 3D SIFT with application to medical image analysis*. in IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08; 2008. IEEE.
104. Paganelli C, Peroni M ,Riboldi M, et al. Scale invariant feature transform in adaptive radiation therapy: a tool for deformable image registration assessment and re-planning indication. *Phys Med Biol*. 2012;58:287.
105. Paganelli C, Seregni M, Fattori G, et al. Magnetic resonance imaging–guided versus surrogate-based motion tracking in liver radiation therapy: a prospective comparative study. *Int J Radiat Oncol Biol Phys*. 2015;91:840–848.
106. Mazur TR, Fischer-Valuck BW, Wang Y, Yang D, Mutic S, Li HH. SIFT-based dense pixel tracking on 0.35 T cine-MR images acquired during image-guided radiation therapy with application to gating optimization. *Med Phys*. 2016;43:279–293.
107. Paganelli C, Peroni M, Baroni G, Riboldi M. Quantification of organ motion based on an adaptive image-based scale invariant feature method. *Med Phys*. 2013;40:111701.
108. Meschini G, Seregni M, Pella A, et al. Evaluation of residual abdominal tumour motion in carbon ion gated treatments through respiratory motion modelling. *Phys Med: Eur J Med Phys*. 2017;34:28–37.
109. Yang D, Zhang M, Chang X, et al. A method to detect landmark pairs accurately between intra-patient volumetric medical images. *Med Phys*. 2017;44:5859–5872.
110. Heinrich MP, Jenkinson M, Bhushan M, et al. MIND: modality independent neighbourhood descriptor for multi-modal deformable registration. *Med Image Anal*. 2012;16:1423–1435.
111. Hauler F, Furtado H, Jurisic M, et al. Automatic quantification of multi-modal rigid registration accuracy using feature detectors. *Phys Med Biol*. 2016;61:5198.
112. Hwang AB, Bacharach SL, Yom SS, et al. Can positron emission tomography (PET) or PET/computed tomography (CT) acquired in a nontreatment position be accurately registered to a head-and-neck radiotherapy planning CT? *Int J Radiat Oncol Biol Phys*. 2009;73:578–584.
113. García-Mollá R, de Marco-Blancas N ,Bonaque J, et al. Validation of a deformable image registration produced by a commercial treatment planning system in head and neck. *Phys Med: Eur J Med Phys*. 2015;31:219–223.
114. Schreibmann E, Pantalone P, Waller A ,Fox T. A measure to evaluate deformable registration fields in clinical settings. *J Appl Clin Med Phys*. 2012;13:126–139.
115. Kierkels R, den Otter LA, Korevaar EW, et al. An automated, quantitative, and case-specific evaluation of deformable image registration in computed tomography images. *Phys Med Biol*. 2018;63:045026.
116. Bender ET, Tomé WA. The utilization of consistency metrics for error analysis in deformable image registration. *Phys Med Biol*. 2009;54:5561.
117. Zhong H, Peters T, Siebers JV. FEM-based evaluation of deformable image registration for radiation therapy. *Phys Med Biol*. 2007;52:4721.
118. Neylon J, Qi X ,Sheng K, et al. A GPU based high-resolution multilevel biomechanical head and neck model for validating deformable image registration. *Med Phys*. 2015;42:232–243.
119. Teske H, Bartelheimer K, Meis J, et al. Construction of a biomechanical head and neck motion model as a guide to evaluation of deformable image registration. *Phys Med Biol*. 2017;62:N271.
120. Ribeiro CO, Knopf A, Langendijk JA, Weber DC, Lomax AJ, Zhang Y. Assessment of dosimetric errors induced by deformable image registration methods in 4D pencil beam scanned proton treatment planning for liver tumours. *Radiother Oncol*. 2018;128:174–181.
121. Qin A, Liang J, Han X, O'Connell N, Yan D. The impact of deformable image registration methods on dose warping. *Med Phys*. 2018;45:1287–1294.
122. Roussakis Y, Dehghani H, Green S, Webster GJ. Validation of a dose warping algorithm using clinically realistic scenarios. *Br J Radiol*. 2015;88:20140691.
123. Nassef M, Simon A, Cazoulat G, et al. Quantification of dose uncertainties in cumulated dose estimation compared to planned dose in prostate IMRT. *Radiother Oncol*. 2016;119:129–136.
124. Samavati N, Velec M, Brock KK. Effect of deformable registration uncertainty on lung SBRT dose accumulation. *Med Phys*. 2016;43:233–240.
125. Mogadas N, Sothmann T, Knoop T, Gauer T, Peterson C, Werner R. Influence of deformable image registration on 4D dose simulation for extracranial SBRT: a multi-registration framework study. *Radiother Oncol*. 2018;127:225–232.
126. Abe Y, Kadoya N, Arai N, et al. Effect of DIR uncertainty on prostate passive-scattering proton therapy dose accumulation. *Phys Med: Eur J Med Phys*. 2017;39:113–120.
127. Low DA, Dempsey JF. Evaluation of the gamma dose distribution comparison method. *Med Phys*. 2003;30:2455–2464.

128. Yan C, Hugo G, Salguero FJ, et al. A method to evaluate dose errors introduced by dose mapping processes for mass conserving deformations. *Med Phys*. 2012;39:2119–2128.
129. Moriya S, Tachibana H, Kitamura N, Sawant A , Sato M. Dose warping performance in deformable image registration in lung. *Phys Med: Eur J Med Phys*. 2017;37:16–23.
130. Niemierko A. Reporting and analyzing dose distributions: a concept of equivalent uniform dose. *Med Phys*. 1997;24:103–110.
131. Gay HA, Niemierko A. A free program for calculating EUD-based NTCP and TCP in external beam radiotherapy. *Phys Med: Eur J Med Phys*. 2007;23:115–125.
132. Rigaud B, Simon A, Castelli J, et al. Evaluation of deformable image registration methods for dose monitoring in head and neck radiotherapy. *Biomed Res Int*. 2015;2015:726268.
133. Cunliffe AR, Contee C ,Armato 3rd SG, et al. Effect of deformable registration on the dose calculated in radiation therapy planning CT scans of lung cancer patients. *Med Phys*. 2015;42:391–399.
134. Saleh-Sayah NK, Weiss E, Salguero FJ, Siebers JV. A distance to dose difference tool for estimating the required spatial accuracy of a displacement vector field. *Med Phys*. 2011;38:2318–2323.
135. Hardcastle N, Bender ET, Tomé WA. The effect on dose accumulation accuracy of inverse-consistency and transitivity error reduced deformation maps. *Australas Phys Eng Sci Med*. 2014;37:321–326.
136. Al-Mayah A, Moseley J, Hunter S ,Brock K. Radiation dose response simulation for biomechanical-based deformable image registration of head and neck cancer treatment. *Phys Med Biol*. 2015;60:8481.
137. Risholm P, Balter J, Wells WM. *Estimation of delivered dose in radiotherapy: the influence of registration uncertainty.* in International Conference on Medical Image Computing and Computer-Assisted Intervention; 2011. Springer.
138. Hub M, Thieke C, Kessler ML, et al. A stochastic approach to estimate the uncertainty of dose mapping caused by uncertainties in b-spline registration. *Med Phys*. 2012;39:2186–2192.
139. Vickress J, Battista J, Barnett R, Yartsev S. Representing the dosimetric impact of deformable image registration errors. *Phys Med Biol*. 2017;62:N391.
140. Tilly D, Tilly N, Ahnesjö A. Dose mapping sensitivity to deformable registration uncertainties in fractionated radiotherapy–applied to prostate proton treatments. *BMC Med Phys*. 2013;13:2.
141. Kirby N ,Chen J, Kim H ,Morin O, Nie K, Pouliot J. An automated deformable image registration evaluation of confidence tool. *Phys Med Biol*. 2016;61:N203.
142. Kim H, Chen J, Phillips J, Pukala J, Yom SS, Kirby N. Validating dose uncertainty estimates produced by AUTODIRECT: an automated program to evaluate deformable image registration accuracy. *Technol Cancer Res Treat*. 2017;16:885–892.
143. ICRU. *Report 62. prescribing, recording and reporting photon beam therapy (Supplement to ICRU Report 50)*. Bethesda: International Commission of Radiation Units and Measurements; 1999.