A Review of Estimating Image Registration Error and Confidence with an Application in Image-Guided Neurosurgery

Joshua Lewis Bierbrier

Biological and Biomedical Engineering
McGill University, Montreal
May 2022

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Engineering

Table of Contents

Abstract

Image registration is widely used in the processing and analysis of clinical and research medical imaging tasks. However, validating and verifying nonlinear image registrations remains a challenging problem and limits its use in clinical settings. The field of registration error and confidence estimation provides a means to verify the result of image registration on a per-case basis. This field was objectively reviewed following PRISMA guidelines. Out of 570 potential records, twenty studies met our inclusion criteria. A taxonomy was developed to structure the field and enable comparison among its diverse algorithms. Trends, suggestions for best practices and opportunities for future research are discussed. Based on the results of the review, a method was implemented for an application with significant potential: estimating errors from MRI to ultrasound registrations in the context of Image-Guided Neurosurgery. Such an algorithm could pave the way for the clinical use of nonlinear registration algorithms in neuronavigation systems that are used to correct for the intra-operative phenomenon of brain shift, ultimately yielding safer surgeries. Ultrasound images were simulated from twelve pre-operative MRI images and deformed artificially to have the ground truth mis-registration. A 3D convolutional neural network was trained on augmented versions of this data to predict error on a voxel-wise basis. Experiments were performed to determine optimal training parameters. Deformations of varying complexity on two held-out test subjects were used to evaluate the model. With known mis-registrations up to 15 mm, the trained model achieved a mean absolute error of 0.849 mm and a Pearson correlation of 0.838 between the known and estimated mis-registrations. Future work includes evaluating the model on real ultrasound data and implementing it on the neuronavigation systems developed by our group. Overall, this work impacts on a more global level, by providing guidance to a growing field, and on a more local level, by providing the groundwork for registration error estimation in Image-Guided Neurosurgery.

Résumé

Le recalage d'images est très utilisé dans le traitement et l'analyse d'imagerie médicale aussi bien en clinique qu'en recherche. Cependant, la validation et la vérification du recalage non linéaire des images reste un problème difficile, ce qui limite son utilisation en milieu clinique. Le domaine de l'estimation de l'erreur ainsi que l'incertitude de recalage fournit un moyen de vérifier le résultat du recalage d'images au cas par cas. Ce domaine a été objectivement examiné conformément aux directives PRISMA. Sur 570 articles potentiels, vingt études répondaient à nos critères d'inclusion. Une taxonomie a été développée pour structurer le domaine d'étude et permettre la comparaison entre ses divers algorithmes. Les tendances, les suggestions de meilleures pratiques et les possibilités de recherches futures sont discutées. Sur la base des résultats de cette revue de littérature, une méthode a été mise en place pour une application à fort potentiel : l'estimation des erreurs de recalage des images IRM et échographiques dans le cadre de la neurochirurgie guidée par l'image. Un tel algorithme pourrait ouvrir la voie à l'utilisation clinique d'algorithmes de recalage non linéaire dans les systèmes de neuronavigation. Ces derniers sont utilisés pour corriger le phénomène préopératoire de déplacement du cerveau, aboutissant finalement à des chirurgies plus sûres. Les images échographiques ont été simulées à partir de douze images IRM préopératoires et déformées artificiellement pour obtenir la déformation initiale considérée comme vérité terrain. Un réseau neuronal convolutif 3D a été entrainé sur des versions augmentées de ces données pour prédire l'erreur au niveau de chaque voxel. Des expériences ont été réalisées pour déterminer les paramètres d'entraînement optimaux. Des déformations de complexité variable sur deux sujets de test retenus ont été utilisées pour évaluer le modèle. Avec des erreurs de recalage préalablement connues allant jusqu'à 15 mm, le modèle pré-entrainé a atteint une erreur absolue moyenne de 0,849 mm et une corrélation de Pearson de 0,838 entre les erreurs de recalage connues et estimées. Les travaux futurs incluent l'évaluation du modèle sur des données échographiques réelles et son implémentation sur les systèmes de neuronavigation développés par notre groupe. Au niveau global, ce travail fournit des lignes directrices et des conseils pour un domaine en pleine croissance, et à un niveau plus local, il fournit les bases pour l'estimation de l'erreur de recalage en neurochirurgie guidée par l'image.

Acknowledgements

First and foremost, I want to acknowledge the support and encouragement of Dr. Louis Collins. More than a supervisor, he was also a mentor and role model.

Thank you to Daniel Di Giovanni and Mohammadreza Eskandari for keeping me company in the lab and for the hours of stimulating conversation, and to the rest of the NIST Lab for the camaraderie. I would also like to acknowledge Dr. Vladimir Fonov and Dr. Houssem Eddine Gueziri for the technical support and guidance.

Thank you to the current and past BBMESS council for providing the BBME program and me a community during these challenging years.

I would like to acknowledge my former supervisor and professor, Dr. Michael Noseworthy. Without his continual mentorship, support and guidance, I would not be where I am today.

I am very grateful for Johnny Der Hovagimian – I could not have asked for a better friend throughout the past two years.

Finally, I want to thank my family – my parents, for their unconditional love and support, my brothers, for constantly impressing and inspiring me, and, in particular, my sister, for always being right around the corner.

List of Figures

Chapter 2

Chapter 3

List of Tables

Chapter 2

Chapter 3

List of Abbreviations

MRI: Magnetic Resonance Imaging

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

CNN: Convolutional Neural Network

CT: Computed Tomography

mm: Millimetres

Chapter 1: Introduction

Image registration is fundamental to medical imaging tasks that require images to be spatially aligned. Such tasks are ubiquitous in the clinical and research domains of medical imaging. They range from intra-subject registrations to fuse multimodal images (e.g., positron emission tomography images to magnetic resonance imaging (MRI) images) to inter-subject registrations for the comparison of patient populations (Holden, 2008; Knowlton, 2009; Rueckert and Schnabel, 2011). Image registration spatially aligns corresponding points in two or more images (Crum et al., 2004; Maintz and Viergever, 1998; Maurer and Fitzpatrick, 1993; Sotiras et al., 2013).

Compared to nonlinear (or deformable, curved or elastic) image registration, linear (i.e., rigid and affine) image registration is a relatively simple task and has achieved clinical acceptance in numerous applications (Viergever et al., 2016). Nonlinear registration, on the other hand, has yet to achieve widespread clinical acceptance primarily due to the difficulty of the task and the challenges inherent in validating such algorithms. For these reasons, in an address following up on Maintz and Viergever's 1998 publication *A survey of medical image registration* (Maintz and Viergever, 1998), Viergever et al. "*strongly advocate a shift of attention to the aspects of validation and clinical acceptance*" (Viergever et al., 2016).

Validation, in this context, refers to ensuring an algorithm can robustly produce accurate image registrations for its intended use (Brock et al., 2017; Jannin et al., 2006). Validating an algorithm does not guarantee it will perform without error in future registrations; registration errors can be expected. Verifying the performance of an algorithm (i.e., ensuring the accuracy of a specific registration (Brock et al., 2017)) is therefore an incredibly important task. Traditional metrics used to characterize the result of a registration, in validation or verification, have shortcomings. The field of estimating registration error and confidence emerged as a result. The goal of error and confidence estimating algorithms is to quantitatively estimate the quality of a registration result on a per-case basis. These algorithms confer numerous benefits to the field of image registration, ranging from increasing the clinical acceptance of deformable image registration to informing downstream processing and analysis tasks of possible errors or bias.

A hypothetical yet relevant and realistic application of such a method lays in Image-Guided Neurosurgery for tumour resections. A pre-operative MRI image is obtained from a patient with a brain tumour. A neuronavigation system is used during surgery to guide surgeons, displaying their surgical tools in relation to the patient's pre-operative MRI image (Drouin et al., 2017; Gerard et al., 2017; Grimson and Kikinis, 2009). The pre-operative MRI image loses fidelity throughout the surgery due to the phenomenon of 'brain shift' – deformation the brain undergoes during surgery (Dickhaus et al., 1997; Gerard et al., 2021, 2017; Hartkens et al., 2003; Hastreiter et al., 2004; Kelly et al., 1986; Nabavi et al., 2001). To mitigate the effects of brain shift, surgeons can acquire intra-operative ultrasound images that show the 'shifted' state of the brain. To correct for brain shift, the pre-operative MRI image can be registered to the intra-operative ultrasound, resulting in an updated pre-operative MRI image that reflects the anatomical/physical changes (i.e., brain shift) that have occurred since the beginning of surgery (Fedorov et al., 2014; Gerard et al., 2017; Sastry et al., 2017). The registration enables surgeons to continue using the neuronavigation platform. One can imagine how registration errors could have devastating effects on the brain tumour resection surgery: the surgeon could be misled and damage eloquent regions or miss resecting tumour tissue. A registration error estimating algorithm could highlight anatomical regions where the registration process performed well, as well as those that the surgeon should be weary of.

The field of registration error and confidence estimation is relatively new. A challenge for researchers new to the field, and even those already in it, is the diversity of the publications it includes. The key feature that unites methods in this field is that they aim to provide error or confidence estimates on a per-registration basis to indicate the quality of a registration. However, these methods range from Bayesian registration algorithms that produce confidence estimates to complex neural network architectures that estimate error. An encompassing and concise resource to unify the field with a structure to link seemingly disparate methods is lacking. To this end, we wrote a compendious systematic review of the field of registration error and confidence estimation. We introduced a novel taxonomy to classify registration error and confidence estimation methods, which provides structure to the field and allows researchers to compare and contrast competing algorithms. Drawing on insights from the review process, we provided suggestions for best practices and fruitful areas for future research. This review was submitted to the international journal Medical Image Analysis on October 29, 2021 (MEDIA-D-21-01149).

The manuscript is presently under second review and composes Chapter 2 of this thesis. Our review will help orient new and current researchers, and guide the growth and development of the field of registration error and confidence estimation.

The application of an error estimating algorithm in Image-Guided Neurosurgery, outlined above, can yield immediate and significant benefits. It could easily be integrated in a neuronavigation system (e.g., Intraoperative Brain Imaging System (IBIS), the neuronavigation system developed in our research group by Drouin et al. (Drouin et al., 2017)). It could then warn surgeons of regions in a registration that did not perform well and instill confidence in nonlinear registration algorithms that can better model brain shift (Archip et al., 2007; Fedorov et al., 2014; Sastry et al., 2017). To the best of our knowledge, an error estimating algorithm for the purpose of ultrasound-MRI registrations in the context of Image-Guided Neurosurgery has yet to be proposed. Therefore, I expand the scope of the field of registration error and confidence estimation to include ultrasound-MRI registrations in the context of Image-Guided Neurosurgery. In Chapter 3, I implemented an error estimation method discovered in the systematic review of Chapter 2 to ultrasound-MRI registrations. The algorithm is based on a sliding-window convolutional neural network that estimates the residual mis-registration error on a voxel-wise basis. To create training data where the true registration error is known exactly, ultrasound images were simulated from pre-operative MRI images and artificially deformed. Experiments were performed to determine optimal training parameters for the algorithm. This exploratory analysis provides the groundwork for such a method to be incorporated in neuronavigation systems.

The thesis ends with Chapter 4, which serves as a conclusion and outlines promising areas of future research.

The contributions of this work include:

- Chapter 2:
    - A taxonomy to classify methods that densely estimate registration error and confidence.
    - A detailed review of the literature on methods that estimate registration error and confidence guided by the PRISMA guidelines (Page et al., 2021). The review is conducted according to the proposed taxonomy.
    - An analysis of the trends and advantages and disadvantages of methods in the field of registration error and confidence estimation. Additionally, suggestions for best practices and directions for future research are discussed.
    - The taxonomy, review and discussion are detailed in: "*Estimating Medical Image Registration Error and Confidence: A Taxonomy and Systematic Review*" by Joshua Bierbrier, Houssem-Eddine Gueziri and D. Louis Collins, submitted to Medical Image Analysis (MEDIA-D-21-01149).

      Joshua Bierbrier: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing - original draft/review & editing.
      Houssem-Eddine Gueziri: Conceptualization, Methodology, Writing - review & editing.
      D. Louis Collins: Conceptualization, Funding acquisition, Methodology, Supervision, Writing - review & editing.
- Chapter 3:
    - The implementation of a registration error estimation algorithm for MRI-ultrasound registrations. To the best of our knowledge, this implementation is novel in the following ways:
        1. The first dense registration error estimating algorithm applied to multi-modal registrations; in particular MRI-ultrasound registrations.
        2. The first dense registration error estimating algorithm using simulated images to obtain reference data for training and validation.
        3. The first dense registration error estimating algorithm applied in the context of Image-Guided Neurosurgery.

- Validation of a registration error estimation algorithm on artificially deformed simulated ultrasound images and (real) pre-operative MRI images.

Joshua Bierbrier: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing - original draft/review & editing.

D. Louis Collins: Conceptualization, Funding acquisition, Methodology, Resources, Supervision, Writing - review & editing.

- Chapter 4:
  - A comprehensive action plan for improving and continuing the development of the algorithm, including enhancing the deep learning model, training data and validation procedure, as well as noteworthy future research directions.

Joshua Bierbrier: Conceptualization, Formal analysis, Investigation, Writing - original draft/review & editing.

D. Louis Collins: Conceptualization, Supervision, Writing - review & editing.

Chapter 2: Estimating Medical Image Registration Error and Confidence: A Taxonomy and Systematic Review

Joshua Bierbrier[a,b], Houssem-Eddine Gueziri[b], D. Louis Collins[a,b,c]

[a] Department of Biomedical Engineering, McGill University, Montreal (QC), Canada

[b] McConnell Brain Imaging Center, Montreal Neurological Institute and Hospital, Montreal (QC), Canada

[c] Department of Neurology and Neurosurgery, McGill University, Montreal (QC), Canada

joshua.bierbrier@mail.mcgill.ca

houssem.gueziri@mcgill.ca

louis.collins@mcgill.ca

Corresponding author

Joshua Bierbrier

joshua.bierbrier@mail.mcgill.ca

3801 University Street

Montreal, Quebec, Canada

H3A 2B4

Abstract

Given that image registration is a fundamental and ubiquitous task in both clinical and research domains of the medical field, errors in registration can have serious consequences. Since such errors can mislead clinicians during image-guided therapies or bias the results of a downstream analysis, methods to estimate registration error are becoming more popular. To give structure to this new heterogenous field we developed a taxonomy and performed a systematic review of methods that quantitatively and automatically provide a dense estimation of registration error. The taxonomy breaks down error estimation methods into Approach (Image- or Transformation-based), Framework (Machine Learning or Direct) and Measurement (error or confidence) components. Following the PRISMA guidelines, the 570 records found were reduced to twenty studies that met inclusion criteria, which were then reviewed according to the proposed taxonomy. Trends in the field, advantages and disadvantages of the methods, and potential sources of bias are also discussed. We provide suggestions for best practices and identify areas of future research.

Abbreviations

PRISMA: Preferred Reporting Items for Systematic reviews and Meta-Analysis
MIND: Modality Independent Neighbourhood Descriptor
CT: Computed Tomography
MRI: Magnetic Resonance Imaging
MCMC: Markov chain Monte Carlo
RF: Random Forest
ML: Machine Learning
mm: Millimetres

1. Introduction

Image registration is the process of spatially aligning homologous points in two or more images (W R Crum et al., 2004; Maintz and Viergever, 1998; Maurer and Fitzpatrick, 1993; Sotiras et al., 2013). The product of a registration is useful for numerous downstream clinical and research tasks including motion correction, motion determination, cross modality image fusion, change detection, distortion correction, atlas construction, atlas registration and segmentation (Holden, 2008). While linear (i.e., rigid and affine) registration was once the focus of most research, nonlinear (or deformable, curved or elastic) image registration currently receives much more attention (Viergever et al., 2016).

Despite such increased attention, deformable image registrations are not immune to error. Registration errors occur whenever the registration fails to correctly align corresponding points in the images (Crum et al., 2003; Rohlfing, 2012). There are several causes for registration error. It may be caused by the quality of the images, which can be noisy, suffer from movement artefacts or have distortions (Brock et al., 2017; Heinrich et al., 2016; Janoos et al., 2012a; Risholm et al., 2013; Saygili, 2018; Zhong et al., 2007). The images may be missing correspondences due to pathology in one of the images but not the other. Misregistration can also be due to differences in contrast between the images, or to the inherent challenges of inter-subject registrations (Brock et al., 2017; Heinrich et al., 2016; Risholm et al., 2010a, 2013; Saygili, 2018; Schultz et al., 2018, 2019), to incorrect modelling assumptions used for the registration (Brock et al., 2017; Janoos et al., 2012a; Kierkels et al., 2018; Nix et al., 2017; Schultz et al., 2019; Zhong et al., 2007), or incorrect optimization or interpolation parameters (Brock et al., 2017), or to the lack of features driving the local deformations in homogeneous regions (Hub et al., 2009; Hub and Karger, 2013; Li et al., 2013).

Deformable image registration algorithms are, consequently, thoroughly validated prior to their general acceptance and use. Recently, the American Association of Physicists in Medicine commissioned Task Group 132 to review the field of image registration in radiotherapy and provide recommendations on quality assurance and control (Brock et al., 2017). They presented measures for assessing the accuracy of image registration algorithms during validation, including target registration error of corresponding landmarks, distance and overlap measures for contoured areas, and properties of the deformation vector field.

While a number of metrics exist to characterize registration results (Brock et al., 2017), each have their drawbacks. For example, *anatomical landmarks* do not yield a dense estimate of the error, are lacking in homogeneous regions, often restricted to highly selective features, difficult to find in large numbers, require expert knowledge of the anatomy and are subject to inter- and intra-observer variability (Bender and Tomé, 2009; R. Castillo et al., 2009; Eppenhof and Pluim, 2018; Hub et al., 2009; Hub and Karger, 2013; Kierkels et al., 2018; Li et al., 2013; Lotfi et al., 2013; Neylon et al., 2017; Obeidat et al., 2016; Ribeiro et al., 2015; Schlachter et al., 2016; Schreibmann et al., 2012; Sokooti et al., 2019b; Zhong et al., 2007). *Contour-based metrics* suffer from similar shortcomings. *Overlap measures* based on contours do not provide error measurements, but are surrogates (Eppenhof and Pluim, 2018; Rohlfing, 2012), and *distance measures* are only sensitive to inaccuracies normal to their surfaces (Obeidat et al., 2016; Rohlfing, 2012). *Consistency and field smoothness measures* may suggest issues with the resulting registration transformation, but do not necessarily infer error (Bender and Tomé, 2009; Hub and Karger, 2013; Ribeiro et al., 2015). *Visualizations* yield qualitative measures (Li et al., 2013; Sokooti et al., 2021) and are open to inter- and intra-observer variability (Sokooti et al., 2021), human fatigue (Sokooti et al., 2021), and are not always feasible for large amounts of data (Nix et al., 2017; Saygili et al., 2016; Sokooti et al., 2021). Finally, metrics that quantify the similarity between images (i.e., *similarity measures/metrics*) can be seriously misleading to assess registration error (R. Castillo et al., 2009; Eppenhof and Pluim, 2018; Obeidat et al., 2016; Rohlfing, 2012).

A thorough validation does not imply perfectly performing algorithms. For example, in 2009, Klein et al. performed an in-depth comparison of leading deformable image registration algorithms (Klein et al., 2009). While their interest lay in determining the best algorithm, Simpson et al. pointed out that none of the algorithms performed without error (Simpson et al., 2011). In addition, it should be noted that the performance of any given algorithm is not uniform throughout the entire registration field.

The idea that the performance of deformable image registration is variable, both within and across subjects, is not new (Brock et al., 2017; Gunay et al., 2018; Kirby et al., 2016; Luo et al., 2020; Muenzing et al., 2012, 2009; Paganelli et al., 2018; Saygili, 2021; Zhong et al., 2007). The results obtained from a validation experiment do not necessarily ensure the success of each subsequent registration with that algorithm. Risholm et al. further point out that if a registration

algorithm is presented with images that contain pathology or artifacts that did not appear in the validation data, one cannot assume it will perform to the same degree of accuracy reached in validation (Risholm et al., 2013).

There are specific applications that necessitate highly robust and accurate registrations, for example in image-guided spine surgeries, where registration accuracies of 1-2 mm are required (Cleary et al., 2000). Registration errors will affect and can bias downstream processing or analysis tasks (Brock et al., 2017; Ribeiro et al., 2015; Saygili, 2020; Sokooti et al., 2019b). This includes clinical decision making in image-guided therapies (Heinrich et al., 2016; Heiselman and Miga, 2021; Kierkels et al., 2018; Luo et al., 2020; Neylon et al., 2017; Nix et al., 2017; Obeidat et al., 2016; Risholm et al., 2013; Saygili, 2020; Schultz et al., 2018, 2019; Sedghi et al., 2019; Sokooti et al., 2019b), and more general research applications, like in atlas construction, multi-modal image alignment, inter-subject registration and segmentation (Holden, 2008). Moreover, the variable performance of deformable image registration impedes its clinical acceptance (R. Castillo et al., 2009; Luo et al., 2020; Neylon et al., 2017; Paganelli et al., 2018; Schlachter et al., 2016; Sedghi et al., 2019). It is therefore no surprise that Viergever et al. "*strongly advocate a shift of attention to the aspects of validation and clinical acceptance*" (Viergever et al., 2016).

Given the ubiquitous use of image registration in the medical field, it seems important that the performance of an algorithm be verified on a per-registration basis. Indeed, there is growing interest in directly estimating the quality of individual registrations (Paganelli et al., 2018; Schultz et al., 2019). This systematic review focuses on such methods. In particular, we review methods that densely estimate the deformable image registration error, or confidence, between two registered images, without any manual intervention. The goal is to be able to automatically and quantitatively assess the results of any given registration immediately after it is performed. This could enable, for example, a surgeon to place trust in a region of a registration with low estimated error (or confidence) during an image-guided (e.g., (Luo et al., 2020; Risholm et al., 2013; Sedghi et al., 2019)) or radiation therapy (e.g., (Paganelli et al., 2018)) procedure. Alternatively, it could be used to inform downstream processing tasks of areas with potentially high error (e.g., (Gil et al., 2021; Simpson et al., 2011)). An error estimation method could further be used to improve the results of a registration (e.g., (Lotfi et al., 2013; Muenzing

et al., 2014)). It can vastly decrease manual assessment time and enable automatic quality control of large-scale image analysis (Sokooti et al., 2019b, 2021).

The field of automatic registration error and confidence estimation is relatively new, with early methods coming from Kybic and Smutek, and from Muenzing et al. (Kybic and Smutek, 2006; Muenzing et al., 2009). Our review is not the first in the field. Paganelli et al. provided an overview of patient-specific validation methods in deformable image registration (Paganelli et al., 2018), focusing on image-guided radiotherapy. We differentiate our review from theirs in several ways. First, our review is more general, rather than focused on radiotherapy. Second, we propose a taxonomy for methods that estimate registration error and confidence, which serves as an objective way for researchers to classify and compare methods. Third, we employ a systematic and objective approach in the review, following PRISMA guidelines (Page et al., 2021) with clear paper inclusion criteria. This provides a succinct yet thorough representation of the field of registration error and confidence estimation that includes recent developments.

The rest of the review is structured as follows. We first describe a taxonomy to characterize different methods and then present the literature review methodology. The results of the systematic review are then presented in terms of the proposed taxonomy. Trends, advantages and disadvantages, and forms of bias are discussed thereafter. Suggestions and best practices are given throughout the Discussion section.

2. Material and methods

2.1. Taxonomy

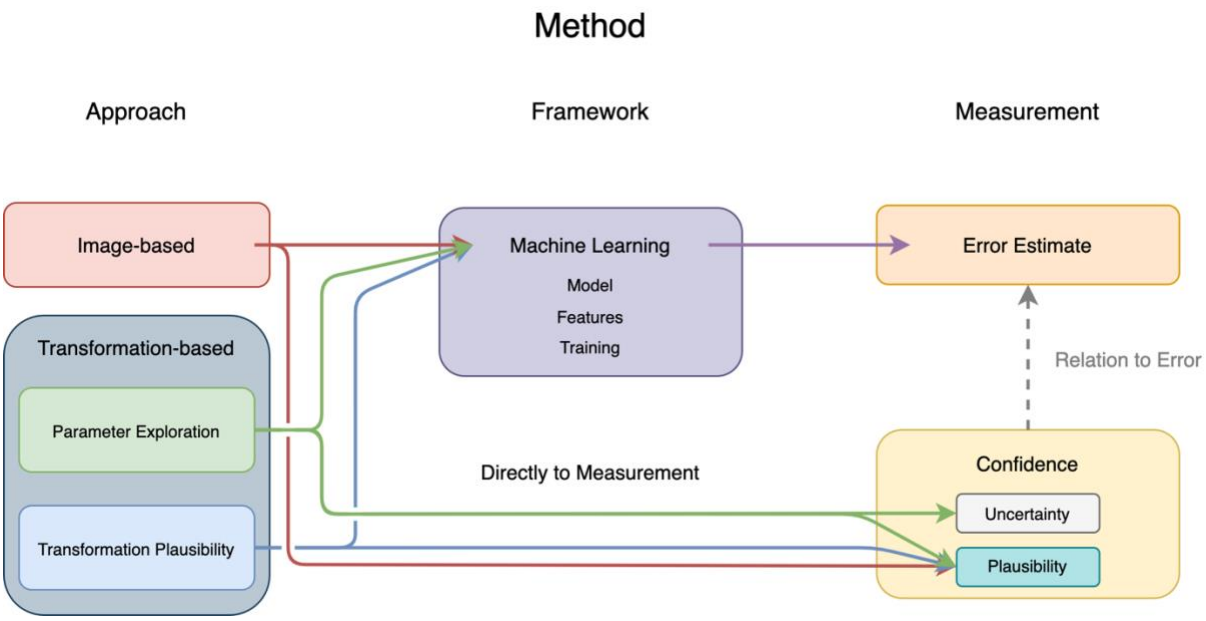

Figure *2.1*. The proposed taxonomy for methods that estimate error and confidence.

We propose a taxonomy to classify methods that automatically and densely predict error and confidence of medical image registrations (Figure 2.1). The taxonomy's format enables different modules to be combined to describe an error and confidence estimation method as required. Note that the method applied to estimate registration error or confidence is usually distinct from the registration method. Image registration transforms a *moving image* to a *fixed image* and yields a *registered (moving) image;* whereas estimating registration error or confidence is generally performed *after* the registration algorithm has been applied.

The taxonomy comprises three components: Approach, Framework and Measurement. The *Approach* is the most fundamental level of a method and describes how the method extracts features from the registered images to estimate *registration error* or *confidence*. The *Framework* specifies whether additional processing is required to transform these features into an estimate, in millimetres (mm) for example, of local misregistration. If such a transformation between the features output from the Approach is not needed, the result of the Approach is used directly as the Measurement. Finally, the *Measurement* specifies the type of output the method produces; this output corresponds to determining the magnitude (and sometimes direction) of the error or

the confidence result. The taxonomy is described in detail below. We begin by describing the types of Measurements, then explain how they are generated by the different Approaches and Frameworks.

2.1.1. Measurement

We classify the output of a method into two types: either *estimated error* or *confidence*. Error and confidence are fundamentally different concepts. Estimated error is a prediction of the registration error, that is, an estimate of the misalignment of corresponding points in the registered images (Crum et al., 2003; Fitzpatrick, 2001; Rohlfing and Avants, 2012). The *accuracy* of a registration relates to the amount of error in the estimated registration transformation. Fundamentally, error estimates must be in distance units (Rohlfing, 2012). Confidence is a direct or indirect measure indicating the level of belief one can have in the correctness of a registration result (Eppenhof and Pluim, 2018; Lotfi Mahyari, 2013). Most Approaches produce a measure of confidence. We identify two categories of confidence Measurement: *uncertainty* and *plausibility*. Uncertainty has been defined as "*a measure of the possible variations within the given model*" (Schultz et al., 2019). It is a measure of precision (or repeatability), not accuracy. The distinction between estimated error and uncertainty is of utmost importance, particularly for end-users. To illustrate, Luo et al. claim that many clinicians believe high (or low) registration uncertainty indicates high (or low) registration error (Luo et al., 2020). In such cases, it is imperative to know how well the uncertainty relates to the error since having little uncertainty (i.e., small variation within a model) does not necessarily imply a lack of registration error. As summarized by Lotfi Mahyari, "*one can be highly certain of an incorrect answer*" (Lotfi Mahyari, 2013). Plausibility is another measure of confidence related to the credibility of the registration results based on *a priori* assumptions. Such assumptions include that high image similarity (quantified through a similarity metric) indicates correct alignment, or that correctly aligned images do not display certain characteristics in their deformation fields (e.g., loss of volume). Even if these assumptions are met, an error-free registration is not guaranteed. Therefore, plausibility Measurements are not synonymous with estimated error Measurements. Finally, note that the Measurement can be a vector (with X, Y and Z components) or a scalar.

2.1.2. Approach

We identified two unique but related Approaches: Image-based and Transformation-based (shown in the first column of Figure 2.1).

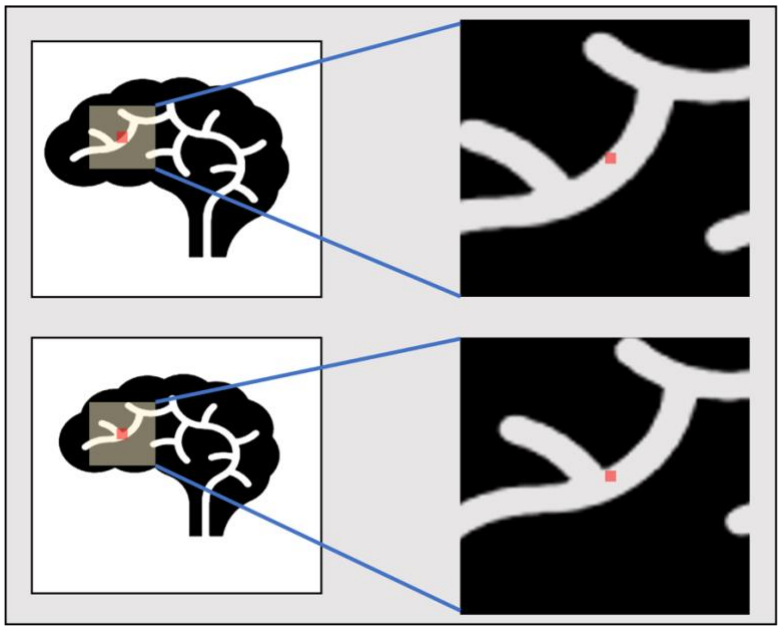

Figure 2.2. A toy example of an Image-based Approach with two slightly different images. The column on the left shows the registered images. The yellow boxes indicate the patches around corresponding (red) voxels assumed to correspond, given the transformation to be evaluated. The column on the right shows the corresponding patches; these are assessed for similarity in Image-based Approaches.

Image-based Approaches use local measures to assess the similarity between corresponding regions of registered images. Such regions are often defined as the immediate neighbourhood (or 'patch') around corresponding voxels of the registered images (as illustrated in Figure 2.2). The neighbourhood similarity can be estimated from the voxel intensities or from other features extracted from the images. These Approaches aim to assess if the regions match, which is *assumed* to indicate a lack of registration error. In many Machine Learning Frameworks, the image intensities are directly used as features. With the exception of neural network methods that consider the input images in one shot, most Image-based Approaches iterate through the voxels of the images to get dense Measurements.

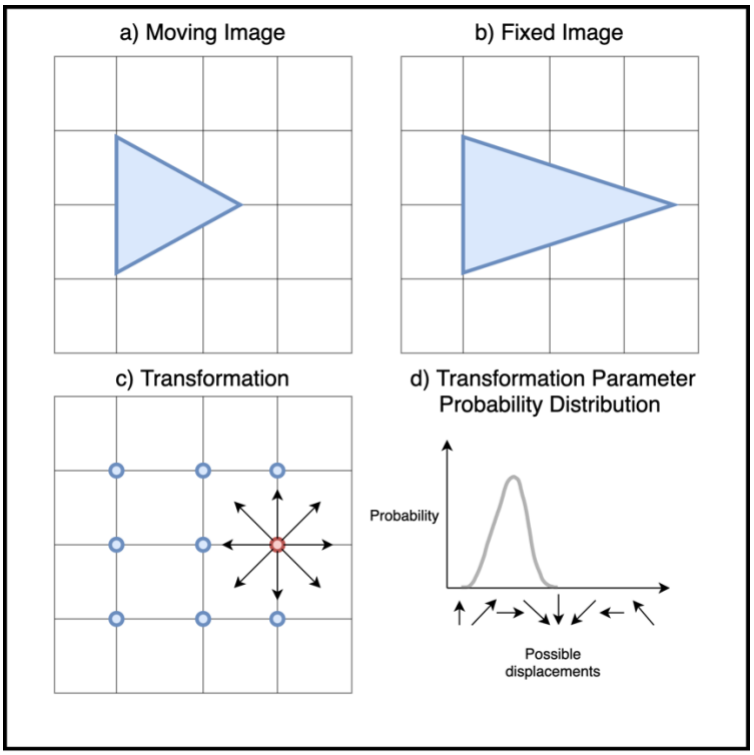

Figure 2.3. A toy example of a Parameter Exploration Approach. The transformation control point under consideration (in red) has different possible options for its displacement. The distribution on deformations, shown in d), illustrates how likely each possible displacement is after the different possibilities have been explored.

There are two types of Transformation-based Approaches: Parameter Exploration and Transformation Plausibility. The first type of Transformation-based Approach is Parameter Exploration. Parameter Exploration Approaches explore the parameter space of a registration transformation. In this process, they can make use of local neighbourhood similarity metrics used in Image-based Approaches. The explored parameters can be those from the registration, for example, the node displacements that drive a registration. The parameters may be explored during the registration itself, which is the case in probabilistic image registration methods. A distinction can be made between Parameter Exploration Approaches that seek to explore the transformation space to determine the distributions of the most likely parameters and those that seek to find the specific parameters that yield a better registration.

As an example of the former, in Figure 2.3, the possible displacements of a node that drives the registration are compared. If the method is more certain about a limited set of deformations, with a peaked distribution as in the case of Figure 2.3d, the confidence is higher. Probabilistic image registration algorithms fit into this category, since they provide a probability distribution for the parameters of the registrations (also referred to as a distribution over

deformations or displacements (Risholm et al., 2013)), shown in Figure 2.3d. Such Transformation-based Approaches typically summarize the obtained distributions by a simple statistic, like variance. Higher dimensional summaries of such distributions are usually possible but lend themselves to intuitive visualization with difficulty.

Transformation-based Approaches that seek to find parameters that yield a better registration may consider the displacement vectors of a deformation field as the transformation parameters to be explored. These methods then search for specific parameters (e.g., the voxel displacement vectors) that yield a better alignment. The output of such an Approach is typically the displacement that purports to improve the alignment.

The second type of Transformation-based Approach is Transformation Plausibility. These Approaches analyze the deformation vector field obtained from the registration to assess if it is physically legitimate by evaluating the physiological realism of the transformation or its numerical consistency (Kierkels et al., 2018). Examples include the Jacobian determinant of the deformation vector field or the inverse consistency error.

2.1.3. Frameworks

The Framework takes as input the results of an Approach and outputs a Measurement of estimated error or confidence.

2.1.3.1. Machine Learning

The Machine Learning Framework can take input from any of the Approaches and is divided into three dimensions. The first dimension refers to the type of machine learning method used – either a classical machine learning model or a deep learning model. The second dimension describes the features that the model uses. These features can come from any of, or a combination of, the Approaches. Finally, the model requires data that indicates the known registration error. This comes from a Reference; either artificially deformed data, where the underlying deformation, and hence the registration error, is known, or from annotated (corresponding) landmarks in both images. In the Machine Learning Framework, a model that learns the relationship between features and registration error is trained. (Note that a Machine Learning Framework could be based on simple regression between the output of an Approach and the known error.) These models output the estimated error.

2.1.3.2. Directly to Measurement

The result of the Approaches can be directly used as Measurements of confidence. In this case, the output of the Approach is not processed through a Machine Learning Framework.

Transformation Plausibility Approaches yield Measurements of plausibility. These Approaches indicate where and why a transformation may or may not be physiologically or physically realistic, based on the registration deformation field. Numerically consistent transformations as well as physiologically or physically realistic registrations do not necessarily ensure a lack of error (Bender and Tomé, 2009; Hub and Karger, 2013; Ribeiro et al., 2015); it is assumed, however, that they make the transformation more likely.

Image-based Approaches also yield Measurements of plausibility. True corresponding points do not necessarily always look the same through the lens of a similarity metric (Pluim et al., 2016; Rohlfing, 2012; Rohlfing and Avants, 2012). Even if their surrounding patches match, it is not guaranteed that they contain corresponding points that are correctly aligned (R. Castillo et al., 2009; Crum et al., 2003; Rohlfing, 2012; Rohlfing and Avants, 2012). Therefore, we designate the result of such Approaches as Measurements of plausibility. Because two patches match (in terms of high similarity), it is more plausible that they are in correct alignment. The same follows for Parameter Exploration Approaches that use Image-based Approaches that seek better alignments.

Parameter Exploration Approaches that summarize the distribution of several possible parameter configurations yield Measurements of uncertainty. In probabilistic image registration settings, for example, the uncertainty measure summarizes the posterior distribution on the deformation or transformation parameters (Janoos et al., 2012b; Lotfi et al., 2013; Luo et al., 2020; Risholm et al., 2013; Schultz et al., 2019), as seen in Figure 2.3.

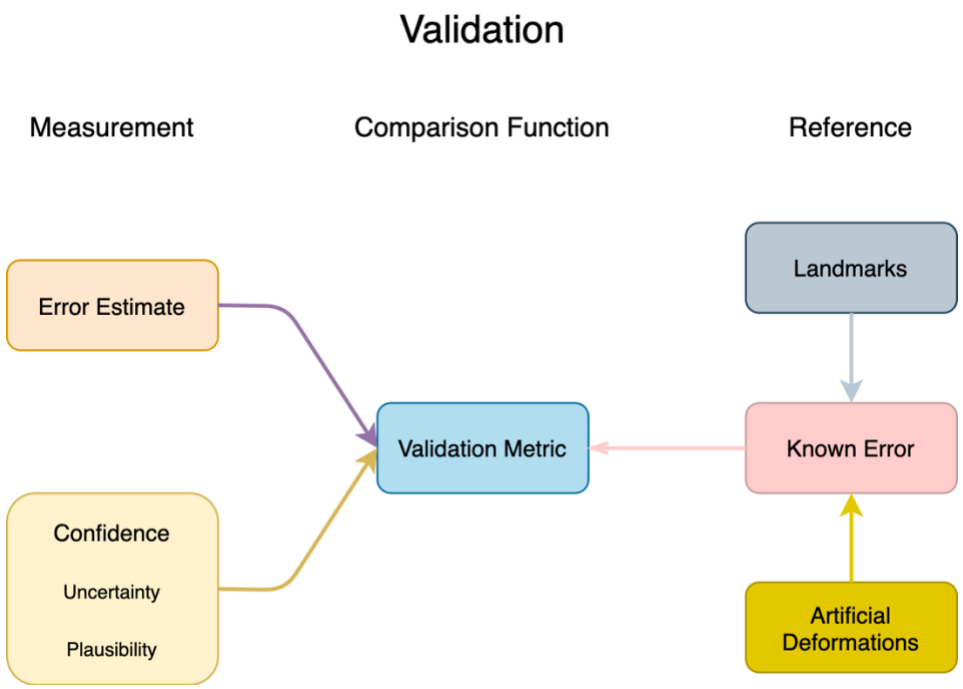

Figure *2.4*. The Reference and Comparison Function are notable aspects of validation procedures. Note that the Measurement is the output of a method, as shown in Figure *2.1*.

2.1.4. Validation

The validation of any method has two main components (see Figure 2.4): the Comparison Function and the Reference. Here, we borrow terminology from Jannin et al. (Jannin et al., 2006). It is important to note that although the estimated error is a direct prediction of the actual registration error, it will likely not be exact. Thus, we distinguish between the estimated (or predicted) error and the known (or actual, true) error obtained from a Reference (Lotfi Mahyari, 2013). We identified two types of References: landmarks and artificial deformations. Note that although these References yield the known error, they are not perfect. In particular, landmarks are subject to inter- and intra-rater variability. While artificial deformations do not have this limitation, they can struggle to match the complexity of real registrations (Fitzpatrick, 2001). Note that labelled anatomical structures can have utility in assessing the quality of a registration as overlap or distance metrics (e.g., the Dice coefficient). However, these metrics are not suitable References for error and confidence estimation given they do not yield the known error for individual voxels.

The Validation Metric is a Comparison Function that relates the Reference, the source of known error, to the estimated error or confidence. Comparison Functions include error metrics and correlation metrics. Error metrics, such as root mean square difference or mean absolute

error, compare the estimated error from a Machine Learning Framework to the known error from a Reference. These quantify the error in estimating the error. A correlation metric can relate confidence or error estimates to the known error.

2.2. Literature search

This review followed the Preferred Reporting Items for Systematic reviews and Meta-Analysis (PRISMA) 2020 guidelines (Page et al., 2021) without prior publication of a review protocol. Keywords were collected from a preliminary literature review and were used to formulate an advanced database search. A campus librarian was consulted when constructing the advanced database search. The following search (last updated September 13th, 2021) was entered as a query on the Scopus database:

> (TITLE-ABS-KEY ((imag* regist*) AND ((uncertain* OR error OR confidence OR accuracy OR quality) W/5 (estimat* OR predict* OR eval* OR quant*))) AND TITLE ((regist*) AND (uncertain* OR error OR confidence OR accuracy OR quality)))

Essentially, it finds records with relevant terms similar to ‘image registration,’ ‘estimation’ and ‘error’ in the Title, Abstract and Keywords fields. In addition, the records collected from the preliminary literature review, as well as appropriate references from the Scopus-returned records, were considered for inclusion. Only English records were considered. The following seven inclusion criteria were considered:

1. Error or confidence must be explicitly related to known error (e.g., through correlation).
2. The method must provide a dense estimation.
3. 3D-3D nonlinear image registration.
4. During application, the method only requires two (registered) images (e.g., no further imaging or manually annotated data) (note that this does not exclude methods that require training data).
5. The method is automatic.
6. Focus is on quantitative estimation of error or confidence.
7. Focus is on biomedical image registration.

These criteria are partially inspired by the "*qualities of a yet to be developed ideal evaluation method for measuring neuroimage registration accuracy*" suggested by Garlapati et al. (Garlapati et al., 2015). For the estimated error or confidence to have physical significance, it should be related to known error (e.g., measured in mm). A dense estimation, meaning that for every registered voxel there is an accompanying Measurement, ensures that the entire field of the 3D-3D deformable registration can be assessed. The method should only require the two registered images – no manual segmentations or landmark identification, or additional image data. The remaining criteria restricted our review to publications on biomedical image registration that have a focus on estimating error or confidence.

One reviewer considered each record relative to the inclusion criteria to determine its fit. When screening records, only the title and abstract were considered. When assessing reports for eligibility, the title, abstract and full publication were considered as necessary. In cases of doubt, a second reviewer was consulted. The first reviewer also extracted results from the included papers, which are presented qualitatively. Only results relevant to the validation of error and confidence estimation methods were sought. The Systematic Review Accelerator's deduplicating tool was used to automatically identify and (manually) remove duplicate records (Rathbone et al., 2015). The systematic review software from Rayyan Systems Inc. was used to structure and organize the paper selection process (Ouzzani et al., 2016).

3. Results

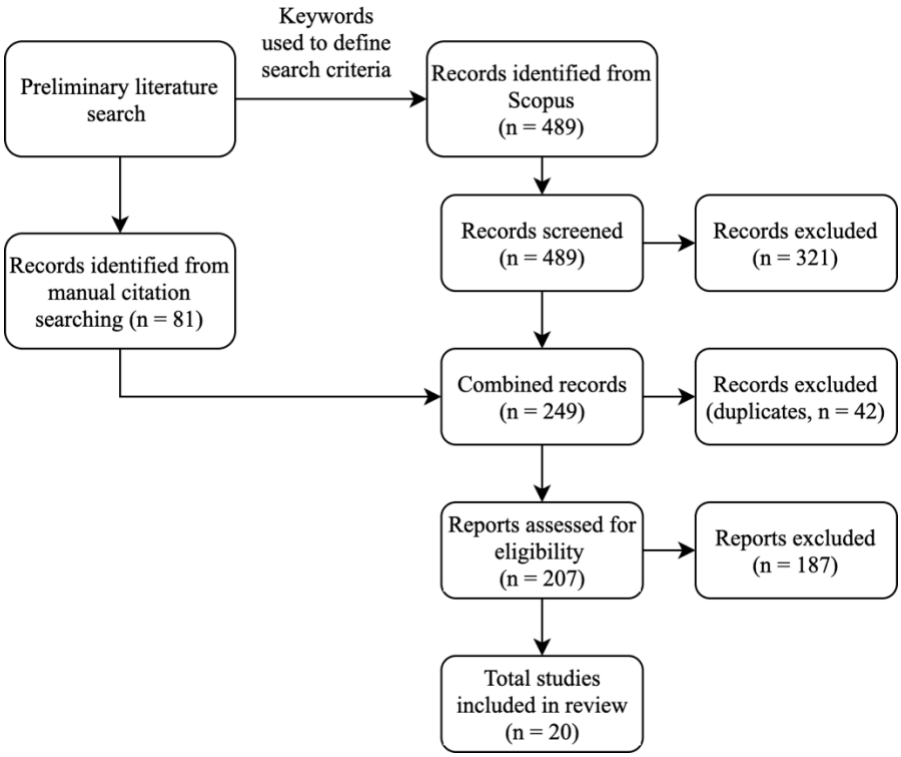

Figure *3.1*. PRISMA flow diagram highlighting the screening and selection process.

The results of the literature search and selection process are given in the PRISMA flow diagram of Figure 3.1. A total of 570 records were considered. After removing duplicates and screening the records, twenty papers were found to match the inclusion criteria and are reviewed here in terms of the taxonomy introduced above (Figure 2.1). The results are summarized in Table 3.1.

Author	**Approach**	**Framework**	**Measure-ment**	**Reference**	**Data**	**Results**
(Bender and Tomé, 2009)	Transformation Plausibility	Directly to Measurement	Plausibility	Landmarks	1. CT of physical phantom containing identifiable landmarks, physically deformed to different levels	Average Spearman's rank order correlation coefficient for two types of registrations: 0.342, 0.661
(Li et al., 2013)	Transformation Plausibility	Directly to Measurement	Plausibility	Landmarks Artificial deformations	1. POPI-model: 1 thorax 4DCT dataset (40 landmarks) (Vandemeulebroucke et al., 2007) 2. CREATIS: 6 thorax 4DCT of cancer patients (100 landmarks per case) (Vandemeulebroucke et al., 2011)	Pearson's correlation coefficient: 0.50 (artificial deformations), 0.64 (landmarks)

(Kierkels et al., 2018)	Transformation Plausibility	Directly to Measurement	Plausibility	Artificial deformations	1. CT of 26 head and neck cancer patients 2. 4DCT of 12 lung cancer patients	Pearson's correlation for Harmonic Energy: 0.44 (Head and Neck region), 0.49 (Lung region)
(Schlachter et al., 2016)	Image-based	Directly to Measurement	Plausibility	Artificial deformations	1. DIR-Lab-4DCT: 10 thoracic CT, half from patients with thoracic malignancies (>300 landmarks per image pair) (E. Castillo et al., 2009; R. Castillo et al., 2009)	Correlation coefficient for Histogram Intersection (plausibility): 0.713 (for one case of artificial deformation; only at landmark locations)
(Saygili et al., 2016)	Parameter Exploration	Directly to Measurement	Plausibility	Artificial deformations Landmarks	1. SPREAD: follow-up chest CT of 21 emphysema patients (~100 landmarks per case, chosen semi-automatically (Staring et al., 2014)) (Stolk et al., 2007) 2. HAMMERS: 30 healthy brain MRI (Hammers et al., 2003) 3. RIRE: T1 and T2 brain MRI (West et al., 1997)	No single value given (only graphs); qualitative correlations shown
(Nix et al., 2017)	Parameter Exploration	Directly to Measurement	Plausibility	Artificial deformations	1. MRI-CT of 14 head and neck cancer patients	Versions of the method detected errors of 1.60 ± 0.67 mm and 1.44 ± 0.92 mm for known in-plane translations of 1.5 mm
(Sokooti et al., 2016)	Parameter Exploration Image-based Transformation Plausibility	ML (RF; Landmarks + surrounding region)	Error	Landmarks	1. SPREAD	Mean absolute error: 0.72 ± 0.96 mm
(Sokooti et al., 2019b)	Parameter Exploration Image-based Transformation Plausibility	ML (RF; Landmarks + surrounding region)	Error	Landmarks	1. SPREAD 2. DIR-Lab-4DCT 3. DIR-Lab-COPDgene: 10 thoracic CT of patients with severe breathing disorders (>300 landmarks per case) (Castillo et al., 2013)	Mean absolute error: 1.07 ± 1.86 mm (intra-database; SPREAD data), 1.76 ± 2.59mm (inter-database)
(Eppenhof and Pluim, 2018)	Image-based	ML (CNN; Artificial deformations)	Error	Landmarks Artificial deformations	1. DIR-Lab-4DCT 2. DIR-Lab-COPDgene 3. POPI-model 4. CREATIS	Root mean square difference: 0.51 mm (Artificial deformations), 0.66 mm (Landmarks)
(Saygili, 2018)	Parameter Exploration	ML (RF; Landmarks + surrounding region)	Error	Landmarks	1. DIR-Lab-4DCT	Mean absolute error: 1.64 ± 1.81 mm
(Saygili, 2020)	Parameter Exploration	ML (RF; Landmarks + surrounding region)	Error	Landmarks	1. DIR-Lab-4DCT 2. CREATIS	Mean absolute error: 2.00 mm (Three Orthogonal Planes approach with RF) $R^2 = 0.74$
(Saygili, 2021)	Parameter Exploration	ML (RF; Landmarks + surrounding region)	Error	Landmarks Artificial deformations	1. DIR-Lab-4DCT 2. CREATIS 3. HAMMERS	R^2 correlations: 0.63756, 0.58005 and 0.68825 (POPI data (x,y,z)) R^2 correlations: 0.65163, 0.56063, 0.77355 (DIRLab data (x,y,z)) Mean absolute error: 1.05 ± 1.28, 0.81 ± 0.87, 1.71 ± 2.91, 1.75 ± 2.47 mm (x, y, z, magnitude)
(Sokooti et al., 2021)	Image-based	ML (Encoder-Decoder; Artificial deformations)	Error	Landmarks	1. DIR-Lab-COPDgene 2. DIR-Lab-4DCT 3. SPREAD	Classification accuracy: 87.1% Average F1 score: 66.4%
(Hub et al., 2009)	Parameter Exploration	Directly to Measurement	Plausibility	Artificial deformations	1. 5 lung 4DCT datasets	No single value given (only graphs); qualitative correlations shown
(Hub and Karger, 2013)	Parameter Exploration	Directly to Measurement	Uncertainty	Artificial deformations	1. 2 toy images (one with translation, other with B-spline deformation) 2. 5 lung 4DCT datasets	Pearson's correlation: 0.19-0.47 (range for all lung cases and directions)
(Lotfi et al., 2013)	Parameter Exploration Image-based Transformation Plausibility	Directly to Measurement ML (RF; Artificial deformations)	Uncertainty Error	Artificial deformations	1. Synthetic texture image 2. BRATS 2012: 6 brain tumour MRI ("MICCAI BRATS 2012," n.d.) 3. LBPA40: 40 brain MRI (Shattuck et al., 2008)	Correlation coefficient: 0.730 (uncertainty; synthetic data), 0.215 (est. error; without uncertainty as a feature), 0.542 (est. error; with uncertainty as a feature)

(Sedghi et al., 2019)	Image-based Parameter Exploration	Directly to Measurement ML (CNN; Artificial deformations)	Uncertainty	Artificial deformations	1. IXI: T1 and T2 brain MRI ("IXI Dataset – Brain Development," n.d.)	Correlation coefficient: 0.94
(Luo et al., 2020)	Parameter Exploration	Directly to Measurement	Uncertainty	Landmarks	1. RESECT: 17 pre-operative and intra-operative ultrasound of brain tumour patients (landmarks) (Xiao et al., 2017) 2. MIBS: 6 pre-operative and intra-operative ultrasound of brain tumour patients (Machado et al., 2018)	Spearman's rank order correlation coefficient: 0.2899 (RESECT), 0.4014 (MIBS)
(Risholm et al., 2013)	Parameter Exploration	Directly to Measurement	Uncertainty	Artificial deformations	1. 1 pre-operative T1 MRI 2. 1 pre-operative and intra-operative (post resection) T2 MRI of a brain tumour patient	Inter-quartile range (uncertainty) contained the ground truth deformation 92% of the time
(Heinrich et al., 2016)	Parameter Exploration	Directly to Measurement	Uncertainty	Landmarks	1. DIR-Lab-4DCT	R^2 correlation coefficient: 0.58 (two registration settings; one subject)

Table 3.1. Summary of the methods by their taxonomy classification, validation, datasets used and results. (CNN: Convolutional Neural Network, ML: Machine Learning, RF: Random Forest)

3.1. Approach

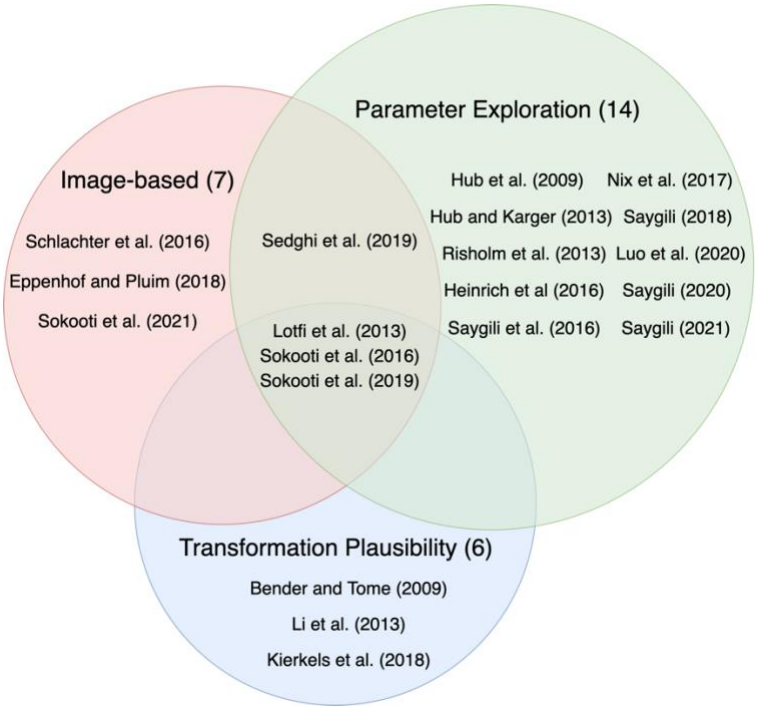

Figure 3.2. Publications classified by their Approach.

3.1.1. Image-based Approaches

Seven of the twenty included publications incorporated Image-based Approaches (Figure 3.2).

Lotfi et al. considered different Image-based Approaches (Lotfi et al., 2013). Three of which were based directly on similarity measures between corresponding patches in the registered images. The difference of the Modality Independent Neighbourhood Descriptor (MIND) features (Heinrich et al., 2012) between corresponding patches was also considered. Similarly, Schlachter et al. proposed computing several similarity metrics patch-wise around each voxel over the registered images (Schlachter et al., 2016). Sokooti et al. used the Euclidean distance between the MIND features (Heinrich et al., 2012) of the fixed and registered moving images (Sokooti et al., 2016). Later, they also included different formulations of the normalized mutual information, as well as other similarity metrics (Sokooti et al., 2019b).

More recently, Sokooti et al. used an Image-based Approach where the intensities of the corresponding patches are extracted as they are, without computing any features, to be used in a Machine Learning Framework (Sokooti et al., 2021). Eppenhof and Pluim, and Sedghi et al., also directly considered the corresponding patches around a voxel in the fixed and registered moving images for a Machine Learning Framework (Eppenhof and Pluim, 2018; Sedghi et al., 2019). However, Sedghi et al. used this Image-based Approach in combination with a Parameter Exploration Approach (Sedghi et al., 2019).

3.1.2. Transformation-based Approaches

3.1.2.1. Parameter Exploration

Fourteen of the included publications employed Parameter Exploration Approaches (Figure 3.2), where an aspect of the transformation parameter space, like the transformation parameters or the vectors of the deformation field, are explicitly explored.

In 2009, Hub et al. proposed a Parameter Exploration Approach for parametric registrations (Hub et al., 2009). They used a B-spline registration, with the vector (i.e., 3D) displacement of each voxel being the parameters explored. Small random perturbations are added to the B-spline coefficients after the registration is performed. The local sum of squared differences similarity metric is computed for each voxel after each perturbation. The vector perturbation that most improves the local similarity metric over what was achieved from the original registration is used as the confidence.

Hub and Karger later proposed a Parameter Exploration Approach for nonparametric registrations (Hub and Karger, 2013). An initial deformation vector field is obtained from the

original Demons-based registration. This deformation vector field is perturbed by adding a fixed offset to each vector component of each voxel displacement. The registration is performed again after each added perturbation, yielding new deformation vector fields. The standard deviation of the resulting deformation vector fields is computed and taken as a measure of uncertainty: if the registration procedure was perfect, it would return to the initial deformation vector field after each perturbation and have a very small standard deviation.

Sokooti et al. used two types of features derived from Parameter Exploration Approaches for a Machine Learning Framework (Sokooti et al., 2016, 2019b). The explored parameters are the displacements of each voxel. The first class of these features was based on the standard deviation of deformation vector fields resulting from several registrations. The registrations can either be randomly initialized each time, or can be perturbed from a base registration, similar to Hub and Karger (Hub and Karger, 2013). Sokooti et al. also computed the difference between the base registration and the mean of the multiple registrations, which they used as a measure of bias. The second Parameter Exploration-based feature used by Sokooti et al. is the coefficient of variation of joint histograms. This feature quantifies the amount the joint histogram of multiple registration varies. If it is high, the registration quality is assumed to be low.

Since deformable image registration can be approximated by local translations (Nix et al., 2017), Nix et al. performed local translation-only re-registrations on corresponding patches of the registered images. Essentially, if a better alignment is found in the local re-registration, based on local similarity measures, it is assumed that the original registration is incorrect. This is a case of a Parameter Exploration Approach that seeks specific parameters (voxel displacements) that yield a better registration.

Saygili et al. also attempted to find better voxel displacement parameters (Saygili et al., 2016). They created a cost space for the voxel under consideration by comparing a dense set of features extracted from the corresponding patches of the fixed and moving images. The local minimum of this cost space is assumed to correspond to the correct alignment. The location of the minimum, as well as the shape of the cost space, are used to quantify the confidence. Saygili further explored this idea by using features derived from the cost space in a Machine Learning Framework (Saygili, 2021, 2018). In 2020, Saygili followed the same approach, but to save computation power, only considered the three orthogonal planes of the neighbourhood around corresponding voxels (Saygili, 2020). The distance to the minimum of each cost space (for each

plane) was extracted. Saygili also applied the pre-trained a stereo matching convolutional neural network (CNN) algorithm (Žbontar and LeCun, 2016) to the three orthogonal planes, and used the resulting features in a Machine Learning Framework (Saygili, 2020).

Five methods use probabilistic image registration. There is an intimate connection between probabilistic image registrations and the Parameter Exploration Approach. These methods inherently explore a range of transformation parameters (see Figure 2.3).

Lotfi et al. used the Random Walker image registration algorithm (Cobzas and Sen, 2011) since it yields a probabilistic output for the displacement of each voxel (Lotfi et al., 2013). A Measurement of uncertainty is computed for each voxel using an improved version of Shannon's entropy on the distribution over displacements. This summarizes the dispersion of a distribution, similar to that shown in Figure 2.3d.

Luo et al. used a Gaussian process probabilistic registration (Luo et al., 2020). The explored parameters are the voxel displacements (similar to Figure 2.3). In Gaussian process registration, features are extracted and matched between the moving and fixed images. The displacement vectors that result from the matched features are interpolated to the remaining voxels using joint gaussian processes. Each voxel is therefore associated with a covariance for its posterior distribution on displacements (similar to Figure 2.3d), which is used as the uncertainty Measurement.

Risholm et al. employed a Parameter Exploration Approach in which the transformation parameters of a linear elastic finite element model are explored (Risholm et al., 2013). Building on previous work (Risholm et al., 2010a, 2010b), Risholm et al. proposed a Bayesian registration method in which the moving image is generated from the fixed image through a deformation (Risholm et al., 2013). Metropolis–Hastings Markov chain Monte Carlo (MCMC) generates samples from the posterior distribution to explore the parameter space. A Measurement of uncertainty is obtained by computing the inter-quartile range of the posterior distribution on deformations for the finite element model vertices.

Heinrich et al. proposed a discrete registration method in which probability distributions over displacements are created for each voxel (Heinrich et al., 2016). Their method consists of dividing the image into several different layers of supervoxels, which are treated as graphs. The registration method is then performed for each of the supervoxel layers. Using belief propagation, marginal distributions for each node of the graph are found. The marginal

probabilities of all supervoxel layers are combined to create a distribution of potential displacements for each voxel. A Measurement of uncertainty is obtained by computing the standard deviation of this distribution.

The final method by Sedghi et al. is a special case since a Machine Learning Framework with an Image-based Approach is incorporated in a Parameter Exploration Approach (Sedghi et al., 2019). They proposed a probabilistic CNN classifier for image registration where corresponding patches around nodes of a B-spline transformation are used as input, and the classifier provides as output the class probabilities for each of 20 classes of discrete displacements for a node. The posterior probability of the classes is used to update the transformation parameters. At the final transformation, the probability distribution over displacements for each node is kept. The variance of this distribution is used as a Measurement of uncertainty. Although the CNN estimates the displacement between corresponding patches, which can be interpreted as an Image-based Approach in a Machine Learning Framework, the authors do not use this to get a final Measurement of estimated error. Instead, as mentioned, they use the variance of the probability distributions over the transformation parameters to yield a Measurement of uncertainty.

3.1.2.2. Transformation Plausibility

Six of the publications that met inclusion criteria use Transformation Plausibility Approaches (Figure 2.3).

Bender and Tomé utilized the Inverse Consistency Error, which quantitatively measures differences between the forward and reverse registrations between two images (Bender and Tomé, 2009). In the ideal case, the Inverse Consistency Error is zero everywhere. Li et al. used the unbalanced energy of a finite element model as a plausibility Approach (Li et al., 2013), which Zhong et al. previously proposed (Zhong et al., 2007). Essentially, the work from an external force is compared to the elastic energy stored in a tetrahedral element of a finite element model. Ideally, this value should be zero.

Kierkels et al. compared several different plausibility Approaches (Kierkels et al., 2018). They referred to two classes of transformations to describe these Approaches: numerically robust and physiologically realistic. The numerically robust measures were the inverse consistency error, the transitivity error and the distance discordance metric, introduced by Saleh et al. (Saleh

et al., 2014). The latter two Approaches require more than two images and therefore do not meet this review's inclusion criteria. The physiologically realistic metrics considered are the Jacobian determinant, the harmonic energy and the octahedral shear stress. Sokooti et al., as well as Lotfi et al., also used the Jacobian determinant of the deformation field (Lotfi et al., 2013; Sokooti et al., 2016, 2019b).

3.2. Framework

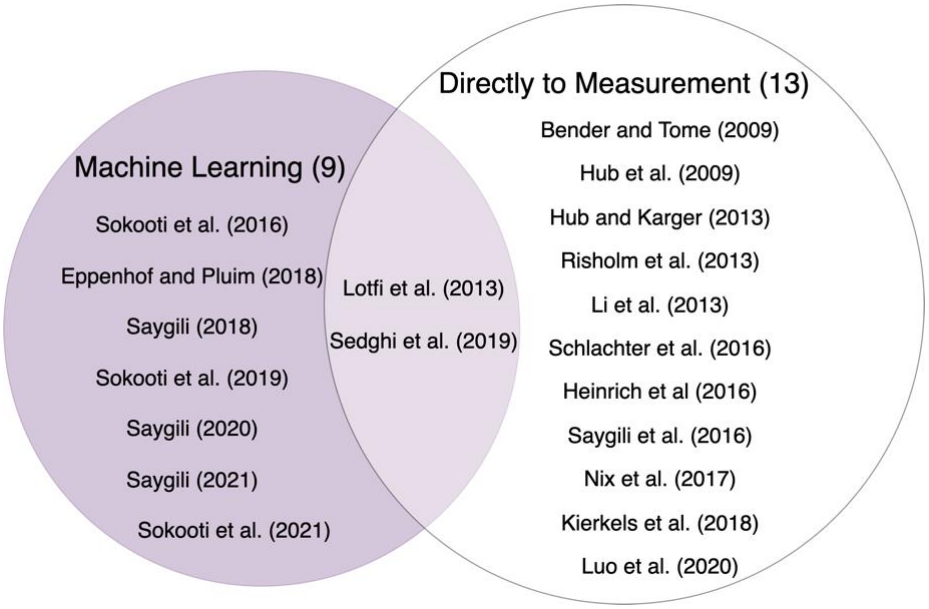

Figure 3.3. Publications classified by their Framework.

3.2.1. Directly to Measurement

Thirteen of the included publications use the results of the Approach directly (Bender and Tomé, 2009; Heinrich et al., 2016; Hub et al., 2009; Hub and Karger, 2013; Kierkels et al., 2018; Li et al., 2013; Lotfi et al., 2013; Luo et al., 2020; Nix et al., 2017; Risholm et al., 2013; Saygili et al., 2016; Schlachter et al., 2016; Sedghi et al., 2019) (Figure 3.3). The Parameter Exploration Approach accounts for the majority of the methods that directly produce a confidence Measurement (Heinrich et al., 2016; Hub et al., 2009; Hub and Karger, 2013; Lotfi et al., 2013; Luo et al., 2020; Risholm et al., 2013; Sedghi et al., 2019). Three methods are from each the Transformation Plausibility Approach (Bender and Tomé, 2009; Kierkels et al., 2018; Li et al., 2013) and the Image-based Approach (Nix et al., 2017; Saygili et al., 2016; Schlachter et al.,

2016). Note that the method of Lotfi et al. outputs a confidence Measurement which they also used in a Machine Learning Framework (Lotfi et al., 2013).

3.2.2. Machine Learning

Rather than directly using the confidence Measurement, the Machine Learning Framework can be used to provide error Measurements. Nine of the included publications used the Machine Learning Framework to transform the output of their Approach into an error Measurement (Eppenhof and Pluim, 2018; Lotfi et al., 2013; Saygili, 2018, 2020, 2021; Sedghi et al., 2019; Sokooti et al., 2016, 2019b, 2021) (Figure 3.3). The Machine Learning Framework is divided into three subcategories: the model, the features and the training data.

3.2.2.1. Model

The model can further be divided into classical machine learning and deep learning. The most popular machine learning model is random forest regression, used to estimate an error Measurement for each voxel (Lotfi et al., 2013; Saygili, 2018, 2020, 2021; Sokooti et al., 2016, 2019b). Lotfi et al. identified several benefits of the random forest model (Lotfi et al., 2013), which were echoed by the subsequent groups. Random forest models can handle a high number of inputs, are resistant to over-fitting and do not necessitate feature pre-processing or normalization (Lotfi et al., 2013; Saygili, 2018; Sokooti et al., 2019b). Furthermore, feature importance can easily be computed. Sokooti et al. made use of Random Forests to see the most relevant features in their model (Sokooti et al., 2016, 2019b).

More recently, sophisticated deep learning models are being adopted. Eppenhof and Pluim trained a 3D CNN where patches around corresponding voxels are used as input to the CNN, which learns to regress an error Measurement (Eppenhof and Pluim, 2018).

Sedghi et al. also used a patch-based CNN model but do not regress a continuous error estimate (Sedghi et al., 2019). Their CNN classifier considers corresponding patches in the two images and provides probabilities for each of 20 classes: unrelated, registered, ±2, ±4 or ±8 voxels (in the x, y, z directions). However, the CNN classifier is not used to provide error estimates. Instead, it is integrated in a registration algorithm that is classified as a Transformation-based Approach in a Directly to Measurement Framework, which uses the class

probabilities of the CNN prediction to obtain a Measurement of uncertainty. Given the method's implicit use of a Machine Learning Framework, it is included in this section as well.

Sokooti et al. use an Encoder-Decoder classification Machine Learning Framework (Sokooti et al., 2021). A latent representation of the registered images is created using the pre-trained RegNet network (the encoder), which is a CNN trained for image registration (Sokooti et al., 2019a, 2017). RegNet is selected since it is trained on similar data and preserves the spatial relation between the images. Three different encoders, corresponding to three different resolution scales of the images, are then created. Rather than simply concatenating the output of each resolution from the latent space and using a CNN as a decoder, they train a convolutional Long Short-Term Memory network as the decoder, which considers finer resolutions and classifications of error at each time step.

3.2.2.2. Features

The methods that use a random forest model make use of a variety of features. While Saygili sticks to the Image-based features (Saygili, 2021, 2020, 2018), Sokooti et al. and Lotfi et al. use a more diverse set of features, including those derived from Image-based, Parameter Exploration and Plausibility Approaches (Lotfi et al., 2013; Sokooti et al., 2019b, 2016). All of the above methods, excluding Lotfi et al., also perform feature pooling (i.e., include features from neighbouring voxels in addition to features from the voxel under consideration). The deep learning methods only use Image-based features. They directly take the voxel intensities as input (Eppenhof and Pluim, 2018; Sedghi et al., 2019; Sokooti et al., 2021).

3.2.2.3. Training data

Machine learning models require training data to learn the relationship between the features and the desired output. Two types of training data were identified in the papers reviewed: landmark points and artificial deformations.

Sokooti et al. and Saygili use landmarks to train their models (Saygili, 2021, 2020, 2018; Sokooti et al., 2019b, 2016). These are corresponding points in both images which can be identified manually, semi-automatically, or automatically, and may come with a published dataset. The actual error is known at such locations. To increase the number of samples available

for training, Sokooti et al. and Saygili also include the neighbourhood around the landmark as training data, assuming it has the same error as the central (landmark) voxel.

The remaining methods use artificially deformed data to train their models (Eppenhof and Pluim, 2018; Lotfi et al., 2013; Sedghi et al., 2019; Sokooti et al., 2021). In this case, a known deformation is applied to an image, or a region from an image. Since the artificial deformation is known, so too is the error between the images. Eppenhof and Pluim created training images from two sets of artificial deformations of one image, based on thin plate spline transformations with uniform random displacements, to generate errors between the images in a range of 0-4 mm (Eppenhof and Pluim, 2018). The data is further augmented by scaling and the addition of random offsets. Lotfi et al. randomly perturbed the nodes of a B-spline transformation grid to generate training data (Lotfi et al., 2013). Sedghi et al. created patch pairs for each class of misregistration they aimed to predict simply through translations (Sedghi et al., 2019). Finally, Sokooti et al. created four types of artificial deformations for their training data (Sokooti et al., 2021): single frequency (constant shift to B-spline knots), mixed frequency, respiratory motion (similar to (Hub et al., 2009)) and identity transform (no misalignment). Further augmentations made these deformations more realistic.

3.3. Measurement

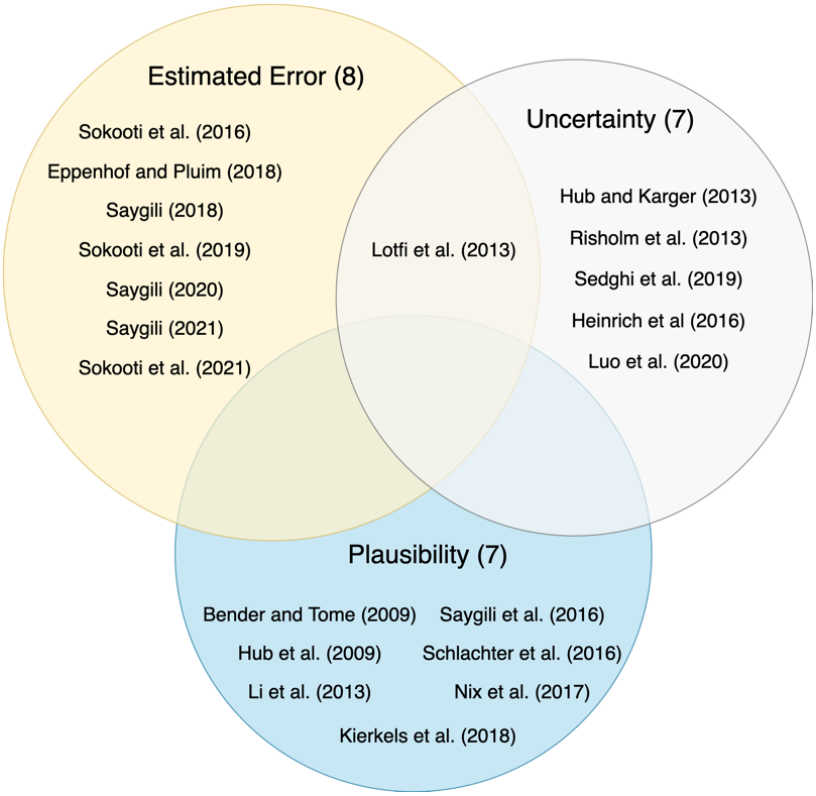

Figure 3.4. Publications classified by their Measurement.

The Measurements produced by the different methods are presented in this section (Figure 3.4). Select quantitative results are presented for context. Note, however, that the results from different methods are generally not directly comparable due to differences in validation choices, like type of Reference, data and Validation Metric. The Bias section below elaborates on these issues.

3.3.1. Estimated Error

All of the methods that yield an error Measurement come from the Machine Learning Framework (Eppenhof and Pluim, 2018; Lotfi et al., 2013; Saygili, 2021, 2020, 2018; Sokooti et al., 2021, 2019b, 2016). Landmarks are the more popular choice of Reference (Saygili, 2020, 2018; Sokooti et al., 2021, 2019b, 2016), followed by artificial deformations (Lotfi et al., 2013). Eppenhof and Pluim, and Saygili use both types of References (Eppenhof and Pluim, 2018; Saygili, 2021). Different Validation Metrics are used to compare the estimated error with the

known error: mean absolute error (Saygili, 2021, 2020, 2018; Sokooti et al., 2019b, 2016), root mean square difference (Eppenhof and Pluim, 2018), correlation (Lotfi et al., 2013; Saygili, 2021, 2020) and classification accuracy (Sokooti et al., 2021). The method Sokooti et al. introduced in 2021 estimates error as one of three classes: [0, 3), [3, 6) and [6, ∞) mm (Sokooti et al., 2021). Other authors also give results as a classification into these same classes, but also included continuous error Measurements (Saygili, 2021, 2020, 2018; Sokooti et al., 2019a, 2016). Saygili et al. is the only method to give vector Measurements of the error (Saygili, 2021).

Sokooti et al. obtained mean absolute errors between the estimated error and the known error of 1.07 ± 1.86 mm and 1.76 ± 2.59 mm in intra- and inter-database experiments, respectively (Sokooti et al., 2019b). Eppenhof and Pluim obtained root mean square differences between the estimated error and known error of 0.51 mm and 0.66 mm for artificial deformations and landmarks, respectively (Eppenhof and Pluim, 2018). Saygili reported mean absolute errors of 1.64 ± 1.81 mm (Saygili, 2018), 2.00 mm (Saygili, 2020) and 1.75 ± 2.47 mm (Saygili, 2021). For the classification method of Sokooti et al., an average classification and F1 score of 87% and 66%, respectively, was reported (Sokooti et al., 2021). Lotfi et al. received a correlation coefficient of 0.54 between the estimated and known error for one case of an artificial deformation (Lotfi et al., 2013).

3.3.2. Confidence

3.3.2.1. Plausibility

Some of the plausibility Measurements come from Transformation Plausibility Approaches (Bender and Tomé, 2009; Kierkels et al., 2018; Li et al., 2013), while the others comes from Image-based (Schlachter et al., 2016) and Parameter Exploration Approaches (Hub et al., 2009; Nix et al., 2017; Saygili et al., 2016). Each method relates the plausibility Measurement with the known error through correlation, except for Nix et al., who compare their mean confidence Measurement with the known error (Nix et al., 2017). Bender and Tomé use landmarks (Bender and Tomé, 2009), whereas others use artificial deformations (Hub et al., 2009; Kierkels et al., 2018; Nix et al., 2017; Saygili et al., 2016; Schlachter et al., 2016). Li et al. use both landmarks and artificial deformations (Li et al., 2013). The plausibility is presented as a scalar in all cases, except two (Hub et al., 2009; Nix et al., 2017).

Bender and Tomé obtained a Spearman's rank order correlations between their plausibility Measurement and the known error of 0.66 and 0.34 for different types of registrations (Bender and Tomé, 2009). Kierkels et al. obtained Pearson's correlations between the Harmonic Energy (a plausibility Measurement) and the known error of 0.44 and 0.49 for different anatomical regions (Kierkels et al., 2018). Schlachter et al. found a correlation coefficient of 0.71 between the known error and the Histogram Intersection similarity measure (Schlachter et al., 2016).

3.3.2.2. Uncertainty

The most common way of measuring uncertainty is to use a simple statistic that describes the spread of the estimated distribution (Heinrich et al., 2016; Hub and Karger, 2013; Luo et al., 2020; Risholm et al., 2013; Sedghi et al., 2019). Lotfi et al. investigated this topic in more depth (Lotfi Mahyari, 2013; Lotfi et al., 2013). Their probabilistic registration method results in probability distributions for the displacement of each voxel. To quantify the uncertainty, they improve on the simple use of Shannon's entropy of the probability distribution by considering the spatial information of the distribution (i.e., the displacement labels). Their uncertainty Measurement is high if the distribution is dispersed (as in Shannon's entropy) *and* if it has high probability for very different displacements (unlike Shannon's entropy). As an example, a distribution with 50% probability of *-4* mm and +5 mm displacements would have higher uncertainty than distribution with 50% probability of *+4* mm and +5 mm displacements.

The uncertainty Measurements come from Parameter Exploration Approaches (Heinrich et al., 2016; Hub and Karger, 2013; Lotfi et al., 2013; Luo et al., 2020; Risholm et al., 2013; Sedghi et al., 2019). Artificial deformations are the more popular Reference (Hub and Karger, 2013; Lotfi et al., 2013; Risholm et al., 2013; Sedghi et al., 2019), however, some authors opted for landmarks (Heinrich et al., 2016; Luo et al., 2020). Everyone except Risholm et al. correlated the known error with their Measurement of uncertainty. Risholm et al., instead, checked whether the ground truth deformation they applied was within their uncertainty level (Risholm et al., 2013). Two methods report the uncertainty as a vector (Hub and Karger, 2013; Risholm et al., 2013).

Hub and Karger obtained Pearson's correlations between uncertainty and known error in the range of 0.19-0.47 (Hub and Karger, 2013), while Lotfi et al. obtained a correlation

coefficient of 0.73 using a synthetic texture image with artificial deformations (Lotfi et al., 2013). Sedghi et al. achieved a correlation coefficient of 0.94 between uncertainty and the known error from artificial deformations (Sedghi et al., 2019). Luo et al. received Spearman's rank order correlations of 0.29 and 0.4 on different datasets (Luo et al., 2020). Heinrich et al. reported an R^2 correlation of 0.58 for one subject with two different registration settings (Heinrich et al., 2016).

3.4. Validation

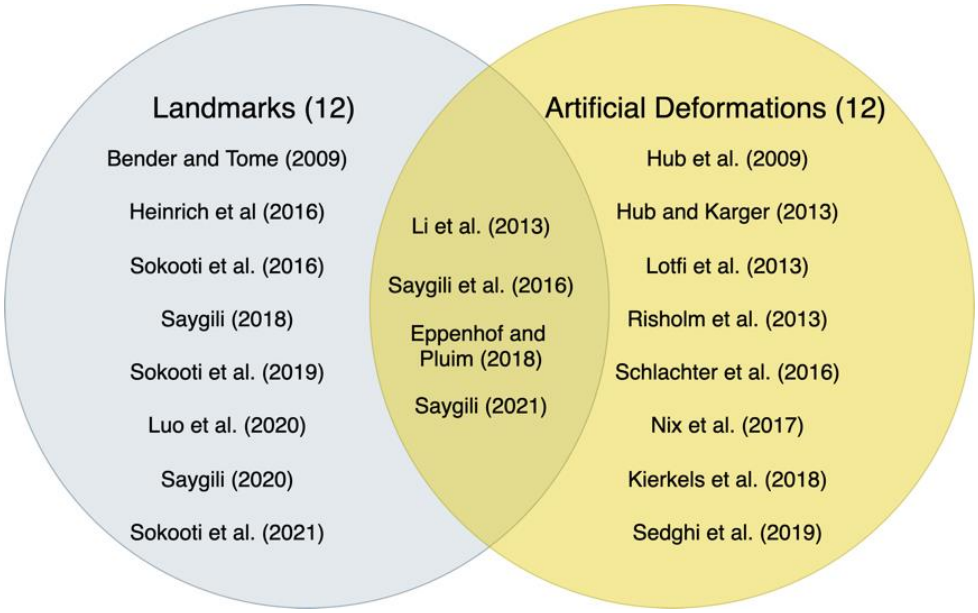

Figure 3.5. Publications classified by their Reference.

The different approaches to evaluating a method can be classified by the Reference data: landmarks or artificial deformations (Figure 3.5). Landmarks compose half the References (Bender and Tomé, 2009; Eppenhof and Pluim, 2018; Heinrich et al., 2016; Li et al., 2013; Luo et al., 2020; Saygili, 2021, 2020, 2018; Saygili et al., 2016; Sokooti et al., 2021, 2019b, 2016). Artificial deformations are alternatively or additionally used (Eppenhof and Pluim, 2018; Hub et al., 2009; Hub and Karger, 2013; Kierkels et al., 2018; Li et al., 2013; Lotfi et al., 2013; Nix et al., 2017; Risholm et al., 2013; Saygili, 2021; Saygili et al., 2016; Schlachter et al., 2016; Sedghi et al., 2019).

Different trends of artificial deformations are identifiable. In the first, a simple known deformation can directly be applied to an image to create the artificially deformed image. Nix et

al. apply translations and rotations applied to a reference image (Nix et al., 2017). Other approaches randomly perturb the control points of a deformation grid (Lotfi et al., 2013; Saygili, 2021; Saygili et al., 2016; Sedghi et al., 2019).

The second trend involves registering moving and fixed images and using the resulting deformation vector field as the ground truth artificial deformation. Kierkels et al. and Schlacter et al. create artificial deformations in this manner (Kierkels et al., 2018; Schlachter et al., 2016). Similarly, Eppenhof and Pluim register moving and fixed images with a B-spline coarse to fine registration scheme (Eppenhof and Pluim, 2018). The registered images after the coarse level and the fine level of the registration are used as the artificially deformed data pair.

The final trend uses realistic artificial deformations for validation. Li et al. and Hub et al. generate realistic deformations of lung motion (Hub et al., 2009; Hub and Karger, 2013; Li et al., 2013). Risholm et al. model the phenomenon of brain shift, with added sinusoidal deformations (Risholm et al., 2013). Sedghi et al., in addition to randomly perturbing B-spline nodes, incorporate image guided-surgery inspired resected tumours by removing areas of different shapes and sizes in random locations of an image (Sedghi et al., 2019).

3.5. Excluded Papers

In this section, we provide several examples of papers that did not meet one or more inclusion criterion but nonetheless present important or interesting methods.

Not all the methods we came across were for deformable image registration (Bansal et al., 2009; William R. Crum et al., 2004; Denis de Senneville et al., 2020; Hauler et al., 2016; Pennec et al., 1998; Pennec and Thirion, 1997). Some papers presented methods that are not dense (Fedorov et al., 2008; Garlapati et al., 2015, 2013; Li and Kurihara, 2014; Muenzing et al., 2012, 2009; Nanayakkara et al., 2009; Neylon et al., 2017; Pizzorni Ferrarese et al., 2014; Werner et al., 2013), including error estimation for point-based rigid registrations (Danilchenko and Fitzpatrick, 2011; Datteri and Dawant, 2012a; Fitzpatrick et al., 1998a, 1998b; Fitzpatrick and West, 2001; West and Maurer, 2002) and methods that perform classification without indicating the magnitude of estimated error (Armato et al., 2006; Galib et al., 2020; Shams et al., 2018; Wu et al., 2016; Wu and Murphy, 2010). Several papers did not compare their confidence Measurement to the known error (Gunay et al., 2018; Mazaheri et al., 2015; Schreibmann et al., 2012). This was especially common for papers with uncertainty Measurements from Bayesian

registrations (Agn and Van Leemput, 2019; Bayer et al., 2020; Grzech et al., 2020; Janoos et al., 2012a; Luo et al., 2019; Risholm et al., 2010a; Wang et al., 2019, 2018; Wassermann et al., 2014; Yang and Niethammer, 2015). Estimating registration error or confidence was not the focus of several publications (Dalca et al., 2019; Datteri et al., 2014; Fedorov et al., 2014; Gibson et al., 2012; Gil et al., 2021; Lin et al., 2019; Muenzing et al., 2014; Ren et al., 2017; Simpson et al., 2015, 2013a, 2013b, 2012, 2011; Sofka and Stewart, 2008; Yang et al., 2017), including those focused on dosimetry (Amir-Khalili et al., 2017; Azcona et al., 2019; Bai et al., 2019; Bender et al., 2012; Hub et al., 2012; Kim et al., 2017; Risholm et al., 2011; Vickress et al., 2017). Several papers did not perform 3D-3D registrations (Heiselman and Miga, 2021; Hu et al., 2016; Kybic, 2010, 2008; Kybic and Smutek, 2006; Le Folgoc et al., 2017; Schultz et al., 2019, 2018; Wang et al., 2001; Watanabe and Scott, 2012; Wu and Samant, 2007, 2004). Multiple methods required more than two images, notably those based on consistency measures (Datteri et al., 2015; Datteri and Dawant, 2012b, 2012c; Gass et al., 2015, 2014; Kim et al., 2013; Kirby et al., 2016; Park et al., 2012; Saleh et al., 2014; Schestowitz et al., 2006; Tyyger et al., 2020; Vaman et al., 2010; Vishnevskiy et al., 2015). Not all methods focused on biomedical image registration (Glocker et al., 2008). Finally, not all methods were automated (Thompson et al., 2018).

4. Discussion

In this section, we first consider current and future trends in method development. Then we present advantages and disadvantages of the Approaches, Frameworks and Measurements. Aspects of the Machine Learning Framework, as well as Parameter Exploration Approaches and uncertainty Measurements, are then considered. Subsequently, we highlight cases of thorough validation, and how different algorithm and validation choices can bias results. Finally, we address limitations of our systematic review. Suggestions for best practices and future research are provided throughout.

4.1. Trends

Machine Learning Frameworks compose most of the newer publications. The salient trend in Machine Learning Frameworks is towards more complex algorithms, namely deep learning. Earlier methods employed more classical machine learning models (Lotfi et al., 2013; Saygili, 2021, 2020, 2018; Sokooti et al., 2019b, 2016). Most commonly, random forest regressors were used. Now, however, deep learning models have displaced the classical machine learning models. CNNs, whether on their own (Eppenhof and Pluim, 2018; Sedghi et al., 2019) or in more complex architectures (Sokooti et al., 2021), are the base model of choice of the included publications. The move to direct use of image intensities, as opposed to features derived from them, seems to obviate the need of feature extraction from Parameter Exploration and Transformation Plausibility Approaches.

The deep learning models require more training data than classical machine learning models. This is reflected in the fact that the methods of creating data for these models are all through artificial deformations. Artificial deformations allow for the creation of arbitrarily large amounts of training data with minimal user input. Conversely, using landmark points as training data, as often done in the earlier uses of the Machine Learning Framework, can require significant amounts of user time and expertise.

Uncertainty Measurements are increasingly obtained from probabilistic image registrations (Heinrich et al., 2016; Lotfi et al., 2013; Luo et al., 2020; Risholm et al., 2013; Sedghi et al., 2019). While the recent deep learning-based registration methods of Yang et al. and Dalca et al. did not meet the inclusion criteria for this review, these methods can provide

uncertainty quantifications (Dalca et al., 2019; Yang et al., 2017). Compared to other Measurement types (error and plausibility), uncertainty is most frequently presented as a vector.

Estimated error is the most common Measurement of all the included publications. It comes from Machine Learning Frameworks. This trend is likely to continue, as other Frameworks have not yet been introduced to estimate error directly. Nearly all error Measurements are scalar. Recently, Saygili proposed a Machine Learning Framework that measures error as a vector (Saygili, 2021). This may signal the beginning of a trend towards directional error metrics. However, Sokooti et al. recently published a method restricting the output to scalar errors in the ranges of three classes (Sokooti et al., 2021), signalling an alternate trend for those who are only interested in a coarser error estimate.

Given the decreasing use of Transformation Plausibility Approaches and the increasing use of Image-based Approaches in Machine Learning Frameworks, the plausibility Measurement is decreasing in popularity.

The incidence of using both landmarks and artificial deformations to evaluate a method is increasing (Eppenhof and Pluim, 2018; Li et al., 2013; Saygili, 2021; Saygili et al., 2016). Recent publications seem to place more emphasis on validation.

4.2. Advantages and Disadvantages

The qualities of an ideal confidence/error estimation algorithm include minimal dependence on tuneable parameters, (training) data or a registration algorithm, as well as fast estimates of the error itself (Garlapati et al., 2015). Here, we cover general benefits and drawbacks of the different Approaches, Frameworks and Measurements.

4.2.1. Approach

Transformation Plausibility Approaches usually do not depend on parameters, data or the registration, meaning that they are widely applicable. Furthermore, they generally do not depend on the intensities of the images. This abstraction from the registration is desirable. However, Transformation Plausibility Approaches yield plausibility Measurements, which are not necessarily indicative of registration error (Bender and Tomé, 2009; Ribeiro et al., 2015).

Parameter Exploration Approaches do not require training data, and in some cases, are applicable to both multimodal (e.g., (Agn and Van Leemput, 2019; Janoos et al., 2012a)) and

inter-subject registrations. However, these approaches generally have tuneable parameters that can change the estimated Measurement. For example, the method of Hub et al. requires to set the maximum magnitude of a perturbation, which limits the estimate of confidence to values smaller than this threshold (Hub et al., 2009). Or, in Bayesian registration algorithms, the strength of the regularization parameter can affect the uncertainty Measurement (Schultz et al., 2019). Parameter Exploration Approaches can also be time consuming. The methods of Hub and Karger (Hub and Karger, 2013) and Sokooti et al. (Sokooti et al., 2019b, 2016) require multiple re-registrations. The sampling performed in Bayesian registration algorithms can require even more prohibitive computation time (Risholm et al., 2013; Schultz et al., 2019). Faster approximation techniques exist, but have their own challenges as discussed in Section 4.4.2 below. Parameter Exploration Approaches that obtain uncertainty through probabilistic registrations depend on the probabilistic registrations itself.

Image-based Approaches range in complexity and can be very intuitive to understand. However, their results may also be misleading. On the one hand, Schlachter et al. achieved decent correlations between a Histogram Intersection metric and the known error (Schlachter et al., 2016). On the other hand, the results of Castillo et al. show that simply using similarity metrics to assess the results of a registration showed very little correlation to the known error (R. Castillo et al., 2009). In fact, they showed cases where the similarity measure increased even though this corresponded to larger registration errors.

4.2.2. Framework

The main benefit of the Machine Learning Framework is that it can draw on complementary information from the different Approaches to directly estimate the registration error. Machine Learning Frameworks can yield fast error estimates once they are trained. However, their speed may be limited by the features they require (e.g., those requiring multiple registrations (Sokooti et al., 2019b, 2016)). The major drawback of the Machine Learning Framework is size of training dataset required to learn the relationship between the features and the error, which depending on the training data, may be biased or limited in its representation. Training data based on landmarks restricts available data to that which is already annotated by experts. The use of artificial deformations overcomes this restriction but introduces unanswered questions regarding the importance of systematic bias or the complexity of the deformations.

Furthermore, it may be harder to generate ground truth artificial deformations for multimodal data. In either case, the machine learning model is limited to the range of errors it is exposed to in the training data. For example, Eppenhof and Pluim only trained their model on errors up to 4 mm (Eppenhof and Pluim, 2018), while Sokooti et al., up to 17 mm (Sokooti et al., 2021). Additionally, the trained model is limited to clinical context of the training data (e.g., specific for the anatomy, modality and pathology), and there can be issues of generalizability to new data even within the same clinical context. Although it can be considered a benefit or a drawback, by the nature of their current training and implementation, machine learning methods view anatomical change between the images (e.g., tumour growth) as registration error (Sokooti et al., 2021, 2019b). Finally, machine learning models may be accompanied by their own uncertainty – similar to the uncertainty of a registration algorithm.

Directly outputting the confidence Measurement from an Approach (i.e., from Section 3.2.1) of course obviates the drawbacks of the Machine Learning Framework but does not directly lead to an error Measurement, and thus, is prone to interpretation errors.

4.2.3. Measurement

Given the task of interest is to assess the registration quality, error Measurements are preferred. However, confidence Measurements, in terms of both uncertainty and plausibility, can no less be valuable if they demonstrate high correlations to the known error.

4.3. Machine Learning Frameworks

4.3.1. Features in machine learning

Sokooti et al. introduced two groups of features in their Machine Learning Framework (Sokooti et al., 2019b, 2016). The first was ‘registration-based features,’ which include Parameter Exploration and Transformation Plausibility Approach-based features. The second was ‘intensity-based features,’ comprised of Image-based Approaches. They trained with both sets of features separately and combined. When trained and tested on the same database, they found that the intensity-based features outperformed the registration-based features. Conversely, in their inter-database experiment, the model trained on registration-based features had a lower mean absolute error than that trained on the intensity-based features. This may suggest that the registration features are more generalizable than the intensity features. With Machine Learning

Frameworks trending towards the sole use of Image-based features (albeit with more complex models), one may wonder if including Parameter Exploration and Transformation Plausibility features, which provide complementary sources of information, could help these models remain generalizable and robust in the face of new data.

Another finding of interest from Sokooti et al. is that feature pooling (i.e., including in the model features from neighbouring voxels in addition to features from the voxel under consideration) helps the regressor (Sokooti et al., 2019b, 2016). The mean absolute error goes from 1.24 ± 2.22 mm without pooling to 1.07 ± 1.86 mm when pooling is performed. Rather than only considering the features at a voxel, the regressor also gets information from the neighbourhood surrounding the voxel. This finding suggests that it is important for models to be given contextual information.

4.3.2. Estimating Error with Transformation Plausibility and Parameter Exploration Approaches

It is interesting to consider how well a Transformation Plausibility or Parameter Exploration Approach can estimate error on its own. Sokooti et al. (Sokooti et al., 2019b) trained machine learning models with individual features, including the Jacobian of the deformation vector field determinant (a Transformation Plausibility Approach) and a Parameter Exploration Approach similar to that introduced by Hub and Karger (Hub and Karger, 2013). The models trained only on the Transformation Plausibility-based feature and only on the Parameter Exploration-based feature achieved mean absolute errors (between the known error and their estimated error) of 2.15 ± 3.15 mm and 1.51 ± 2.40 mm, respectively (Sokooti et al., 2019b). Perhaps surprisingly, these are not that much worse than the model trained on all features (1.07 ± 1.86 mm).

Taken together, these results indicate that Transformation Plausibility and Parameter Exploration Approach-based features can potentially lead to acceptable estimations of error on their own. It is interesting to consider, however, that Kierkels et al. found no correlation between the ground truth error and the Jacobian of the determinant (Kierkels et al., 2018). Similarly, Ribeiro et al. found little correlation between field smoothness measures and the known error (Ribeiro et al., 2015). Hub and Karger found correlations in the range of 0.19-0.47 between the known error and their method (Hub and Karger, 2013). These discrepancies may be the result of different types of data, References or Validation Metrics. For example, Sokooti et al. used

landmarks (Sokooti et al., 2019b, 2016), whereas the other studies used artificial deformations (Hub and Karger, 2013; Kierkels et al., 2018; Ribeiro et al., 2015).

4.3.3. Training in Machine Learning Frameworks

4.3.3.1. Training with Artificial Deformations

There are several unique approaches to create training data for Machine Learning Frameworks. Sokooti et al. and Sedghi et al. represent opposite ends of the spectrum (Sedghi et al., 2019; Sokooti et al., 2021). Sedghi et al., on the one hand, restrict their artificial deformations to simple translations. Sokooti et al., on the other hand, not only use deformations of varying complexities, but they also include realistic lung motion deformations. The impact of having very simple deformations or a range of more complex deformations, has not yet been studied. Despite the simple approach taken by Sedghi et al., their resulting uncertainty estimates correlate well with the known error ($r = 0.94$), and their combined registration and uncertainty estimation method can recover deformations from B-spline transformed images. However, the task of the CNN used by Sokooti et al. and Sedghi et al. is different. The CNN trained on simple deformations by Sedghi et al. updates the registration iteratively and is not used to directly estimate error. As discussed in Section 3.2.2.1, their method instead provides a Measurement of uncertainty. On the other hand, the CNN employed by Sokooti et al. directly estimates the error. Regardless, further research is warranted to study the effect of the complexity of artificial deformations used to train error estimating neural networks.

4.3.3.2. Training through Transfer Learning

Creating training data for Machine Learning Frameworks remains a challenge due to the lack of available ground truth data. This is, in fact, the motivation to create artificial deformations. In 2020, Saygili (Saygili, 2020) adapted the stereo matching CNN developed by Žbontar and LeCun (Žbontar and LeCun, 2016), which was trained on the KITTI stereo matching dataset (Geiger et al., 2012), to provide features that were fed to a classical Machine Learning Framework for registration error estimation. This produced results on par with the difference of MIND features (Heinrich et al., 2012) that Saygili also explored. Such transfer learning approaches are potentially of great use and deserve more research in the future. One may hypothesize that first employing MIND features (Heinrich et al., 2012) to the training stereo

data, and subsequently on testing medical image data, may further improve the applicability of such approaches.

4.3.3.3. Distribution of Errors in Training

The distribution of errors used to train a model in a Machine Learning Framework can be uniform or nonuniform. While Eppenhof and Pluim and Sedghi et al. use a uniform distribution of errors (Eppenhof and Pluim, 2018; Sedghi et al., 2019), Sokooti et al. and Saygili use a distribution of errors that mimics what the model will likely see (Saygili, 2021, 2020, 2018; Sokooti et al., 2021, 2019b, 2016). In the 2021 publication of Sokooti et al., more training examples are explicitly put into the lower error class ([0,3) mm). In the publications of Sokooti et al. (2016, 2019) as well as those from Saygili (2018, 2020, 2021), it is hard to control the distribution of errors since they are from annotated landmarks. Sokooti et al. and Saygili used different approaches to ensure a large range of errors were included in their training set. Saygili registered the images only at a coarse level with a limited number of iterations. Sokooti et al. registered the images with a varying number of iterations, with lower errors expected after more iterations. It is not clear which approach, a uniform or nonuniform distribution of errors, is superior. However, it is worth noting that the mean absolute error for errors in the range of [6,∞) mm obtained by Sokooti et al. and Saygili are notably larger than the mean absolute error for errors in the range of [0,6) mm (Saygili, 2021, 2020, 2018; Sokooti et al., 2019b, 2016).

4.3.4. Missing Correspondences

Luo et al. contend that Image-based Approaches will run into trouble in cases where there are significant differences between the images being registered, making identification and alignment of corresponding features difficult (Luo et al., 2020), as may be the case of brain shift resulting from tumour resection in image guided surgery. The work of Sedghi et al. may potentially pose a solution for this issue (Sedghi et al., 2019). The CNN trained by Sedghi et al. is incorporated in a registration scheme that iteratively updates the registration based on the estimated misalignment into one of twenty classes. One such class is *unrelated*, and lets the registration know that the patches under consideration do not match. Interestingly, when the model is trained without this class, the registration performs much worse. Such an *unrelated* class could be visualized alongside an error or confidence estimate. It could also easily be added to the CNN of Eppenhof and Pluim (Eppenhof and Pluim, 2018). While this does not fully solve

the issue raised by Luo et al., it would help guide and warn surgeons where the registration may not be performing well.

4.4. Parameter Exploration Approaches and Uncertainty Measurements

4.4.1. Uncertainty and Error

Two studies investigated the relation between uncertainty and error explicitly (Luo et al., 2020; Schultz et al., 2019). In Luo et al. (Luo et al., 2020), uncertainty estimates were obtained through a Gaussian Process-based registration. They compared uncertainty to the known (landmark) error through Spearman's correlation. The average correlations of two separate datasets were 0.2899 and 0.4014. These results, as recognized by Luo et al., do not encourage the use of uncertainty as a surrogate for error. However, Luo et al. mention that the registration uncertainty is inversely related to the distance a voxel is from a feature used to drive the Gaussian Process registration. One may not expect this uncertainty to have a strong relation to error. It is possible that uncertainty derived from other probabilistic registration algorithms better relates to error.

Schultz et al. investigated the relation between uncertainty Measurements and known error from point-based probabilistic (variational Bayesian) registrations of simulated deformations, where the posterior of the transformation parameters is used to characterize uncertainty (Schultz et al., 2019). They found that when no artificial lesions were inserted during simulation, the correlation between error and uncertainty was very low (ρ = 0.0516). However, when artificial lesions were inserted, the correlation increased (ρ = 0.2243), even more so with larger lesions (ρ = 0.7397). The uncertainty did not correlate with error in the case of *Model Mismatch* (Hub et al., 2009), errors caused by differences in transformation models used to simulate deformations and those used to estimate the registration, but did correlate with error in the case of missing correspondences caused by the artificially inserted lesions. Given the higher correlation with large lesions, Schultz et al. purport that registration uncertainty can be used as a measure of error due to missing correspondences, such as in the case of pathology.

While this section puts the validity of registration uncertainty as a surrogate for error into question, Sedghi et al. provide an example of high correlation between error and uncertainty (r = 0.94) with the use of artificial deformations with simulated resected lesions (Sedghi et al., 2019). Similar to Schultz et al. (Schultz et al., 2019), however, this high correlation may be due to the

missing correspondences introduced by the resected lesions as there is no Model Mismatch in their experiments. The implications of Model Mismatch are discussed in Section 4.6.4.

4.4.2. Approximate and Exact Inference

Probabilistic registration algorithms rely on exploring a parameter space of a transformation to produce an estimate of registration uncertainty. In some cases, these methods can be computationally expensive. Risholm et al. used a Bayesian registration scheme to explore the posterior distribution with a MCMC approach (Risholm et al., 2013). Barring any approximations, such approaches are asymptotically exact. In their case, sampling the posterior took roughly 50 hours to evaluate a single registration thus limiting its applicability.

An alternative to sampling the posterior of the transformation parameters exactly is to approximate the posterior through, for example, a variational inference approach (Le Folgoc et al., 2017; Schultz et al., 2019; Simpson et al., 2012). This strategy provides computational improvements at the cost of assuming the form of the posterior distribution (typically Gaussian). However, Risholm et al. and Janoos et al. found that the posterior distribution can be non-Gaussian and multimodal (Janoos et al., 2012b; Risholm et al., 2013). Because of this, Risholm et al. cautioned the use of such approximate approaches until they were thoroughly compared to the results of asymptotically exact sampling methods. Le Folgoc et al. did compare a variational inference approach with asymptotically exact MCMC and found that the uncertainty obtained in the approximate sampling case did not adequately match that from the asymptotically exact sampling (Le Folgoc et al., 2017).

On the one hand, these findings suggest that variational approaches may not be useful for uncertainty quantification, despite their increase in speed. On the other hand, useful results have been achieved using variational approaches (Schultz et al., 2019; Simpson et al., 2012). Overall, caution should be taken when using a fast but approximate inference scheme.

4.4.3. Summarizing the Distribution on Deformations

Once a distribution on deformations is obtained, it must be summarized for further use. This aspect has been relatively underappreciated. Lotfi et al. proposed a new technique to summarize such distributions (Lotfi et al., 2013). While similar to Shannon's entropy, it is tailored to summarizing uncertainty in image registration (Lotfi Mahyari, 2013).

Risholm et al., in a precursor to their 2013 publication, also gave this aspect some thought (Risholm et al., 2010a). They concluded that the interquartile range is a simple yet robust way to summarize posteriors. They propose visualizing the interquartile range in 3D as ellipsoids or as scalars at each voxel, in which the maximum interquartile range of all three directions is used. Note that a scalar map cannot convey directional bias. Since the estimated nonlinear transformations can be used to update intra-operative images with functional data from the pre-operative images (e.g., functional MRI, fMRI or Diffusion Tensor Imaging, DTI), Risholm et al. proposed further information-specific measures to show the uncertainty for such data and their derived contoured volumes (Risholm et al., 2010a). In 2013, Risholm et al. visualized the posterior expected deformed image, the average deformation of all the drawn samples from the MCMC scheme, wherein blurrier regions represent higher uncertainty (Risholm et al., 2013).

4.4.4. Future Research in Parameter Exploration

There remains much to study about methods that produce uncertainty Measurements if their intent is to be used as surrogates for error. The relationship between error and uncertainty is murky. Further investigating this relationship, perhaps in light of the different types of error introduced by Schultz et al. (Schultz et al., 2019), is recommended. In particular, the impact of Model Mismatch on error. Artificial deformations based on different transformation types may help elucidate this aspect. In addition, the correlation between error and uncertainty of probabilistic registrations based in different types of transformations should be studied. The influence that particular sampling schemes have on the relationship also deserves future attention. We have not yet come across a case where the uncertainty from an asymptotically exact sampled posterior (e.g., by MCMC) is correlated with error. Finally, the effect of summarizing the distribution on deformations – whether as a simple dispersion statistic or a more intricate measure – on the correlation between error and uncertainty should be considered in more detail (see (Lotfi Mahyari, 2013)).

4.5. Validation

Properly validating a method can be as important as the method itself. Validation should ensure a method's performance in a variety of scenarios and ultimately provides end-users, like

clinicians, confidence in that method. Here, we highlight cases of thorough validation from recent publications.

Eppenhof and Pluim trained a machine learning model to predict error (Eppenhof and Pluim, 2018) and validated with artificial deformations and landmark points as References to leverage the benefits of both. In addition, they validated with a third independent dataset that was not used for training. Similarly, Sokooti et al. (Sokooti et al., 2021, 2019b) trained their machine learning model on one dataset and tested it on another. Such inter-database evaluations are especially important for Machine Learning Frameworks to assess generalizability. Going even further, Sokooti et al. (Sokooti et al., 2019b) used different registration algorithms to assess generalizability.

These publications also provide insightful figures to illustrate the results of their validation. A figure visualizing the estimated Measurement (error, uncertainty or plausibility), perhaps overlayed on the registered image, can be considered essential. Sokooti et al. show the estimated and predicted error across all samples in a graph of error vs. samples (Sokooti et al., 2019b). This is helpful to assess regions where the error is over- or underestimated. Correlation scatter plots are likewise helpful. Eppenhof and Pluim present error histograms to assess bias in the estimated error (Eppenhof and Pluim, 2018). Future inspiration may be drawn from the visualization field of positional uncertainty (Gillmann et al., 2021; Pang et al., 1997).

Schlachter et al. performed a user study with clinicians (Schlachter et al., 2016). Such studies will be increasingly important in the future if error estimation methods are to gain clinical acceptance.

4.6. Bias

As recently noted by Sokooti et al. (Sokooti et al., 2021), comparing the results of different methods is far from trivial due to differences in transformation models, imaging data, organ targets, clinical context and applications. In the following, we review how choices in validation can bias results and the implications that follow. Ultimately, this demonstrates the need for a standardized protocol and openly available data to assess and compare competing methods. Ideally, a challenge would be organized that follows a standardized protocol and assesses methods on the same dataset. While such a protocol and challenge is beyond the scope of this review, we provide suggestions where relevant.

4.6.1. Clinical Context and Data

Validation takes place on data from various modalities (e.g., MRI vs. CT), focussing on vastly different anatomies (e.g., brain vs. lung) and in different clinical contexts (e.g., exhale-inhale images vs. monthly follow-up; pathology vs. healthy). To fully appreciate a method's performance, it would ideally be rigorously tested in regard to the clinical context it will be used in. Some of the included publications restrict their validation to a single domain, while others include multiple domains. For example, Saygili et al. used lung CT and brain MRI data (Saygili et al., 2016). Methods that use Machine Learning Frameworks may be more restricted to the type of data they were trained on. Nonetheless, Sokooti et al. tested their method on multiple datasets with different clinical contexts (Sokooti et al., 2019b); inspiration and expiration and follow-up lung CT images, as well as different pathologies. While the results of a study that investigates a method in diverse settings may indicate that it is less accurate than if it is only considered in a specific clinical context, they provide a better representation of the method's overall capabilities.

The information used to compute results is also important to consider (Jannin et al., 2006). Some authors decided only to use certain parts of the registered images to calculate their results. For example, Eppenhof and Pluim only consider voxels within a lung mask for the root mean square difference between known and estimated error (Eppenhof and Pluim, 2018), and similarly Nix et al. only used data within the patient boundaries (Nix et al., 2017). Gunay et al., on the other hand, used all of the image data to compute their uncertainty Measurement (Gunay et al., 2018). Kierkels et al. excluded the brain region of head and neck cancer CT images from their correlation between plausibility and known error due to its poor results (Kierkels et al., 2018). That said, Kierkels et al. used CT data, where there is notably limited contrast in the brain. Focusing the analysis on a region of interest in the data is perfectly acceptable, but it is important to note that such a result may be biased to look better than a result on all the anatomy visible in the images. Therefore, this should be made clear by authors.

4.6.2. Validation Metrics

Different Validation Metrics compare and summarize the data differently. Mean absolute error and root mean square difference are used in error estimating Machine Learning Frameworks. Sokooti et al. and Saygili used mean absolute error (Saygili, 2021, 2020, 2018;

Sokooti et al., 2019b, 2016), whereas Eppenhof and Pluim used root mean square difference (Eppenhof and Pluim, 2018). The latter is more sensitive to large discrepancies (or outliers). Presenting a histogram of error is more informative but does not summarize the error in a single number.

Saygili reported both correlation and mean absolute error for several methods (Saygili, 2020). In general, the rankings for R^2 matched that the mean absolute error, but this was not always the case. Therefore, we recommend reporting both metrics.

While some authors used Pearson's correlation coefficient (e.g., (Hub and Karger, 2013; Kierkels et al., 2018)), others used Spearman's rank correlation coefficient (e.g., (Bender and Tomé, 2009; Luo et al., 2020)). Ribeiro used both Pearson's and Spearman's correlation (Ribeiro et al., 2015). Luo et al. used Spearman's rank correlation coefficient to assess the relationship between uncertainty and error (Luo et al., 2020) where they identified two desirable properties over Pearson's correlation. First, Pearson's correlation looks for strength of a linear relationship, whereas Spearman's correlation evaluates monotonic relationships. Second Spearman's correlation is not as sensitive to outliers.

Authors may opt to report their results in classes (e.g., [0, 3), [3, 6) and [6, ∞) mm), similar to Sokooti et al. (Sokooti et al., 2021). Sokooti et al. recommend using F1 and Cohen's Kappa scores over accuracy, as they are more robust metrics for distributions with unbalanced classes. They reported all three metrics for the several methods they compared. It is worth noting that there are cases where the accuracy of one method is higher than another, while the other method has a higher F1 score. This illustrates how different Validation Metrics can obfuscate the validation process.

Overall, it is important to recognize that the choice of Validation Metric matters. It can influence the perception of a result and make the comparison of methods more challenging. Given that most Machine Learning publications use mean absolute error, we recommend using it, along with Pearson's correlation. Luo et al. provide a strong case for the use of Spearman's rank correlation (Luo et al., 2020), however, this may be application dependent (e.g., whether or not a strictly linear relationship is sought). It can also be helpful to show the mean absolute error or correlation in different ranges of error (e.g., [0, 3), [3, 6) and [6, ∞) mm), regardless of whether the method performs classification or regression.

4.6.3. Validation References

The use of landmarks or artificial deformations as a Reference provides complementary views of the performance of a method. Landmarks are identified at distinct locations. As a result, they only assess how the algorithm does for these distinct (usually high contrast) locations of an image. An interesting result from Obeidat et al. is applicable here (Obeidat et al., 2016). They assessed the performance of different registration algorithms on artificially deformed data, which also had landmarks selected at high contrast locations. The mean registration error detected by the landmarks was up to 4 times smaller than the mean registration error obtained from the (dense) artificial deformations. This demonstrates that the errors obtained at distinct landmark locations are not necessarily representative of the errors throughout the registered images.

Artificial deformations, in contrast, provide a dense map that enables dense evaluation of an algorithm, but most likely represent a simplified version of the true deformations and thus result in a potentially unrealistic version of the true registration error between two images (Pluim et al., 2016). Fitzpatrick notes that "*simulations have the primary advantage that the transformation is known exactly, and the secondary advantage that any transformation is readily available… but they lack the realism arising from the sometimes subtle anatomical changes… validations based on simulations can, however, provide an upper bound on success*" (Fitzpatrick, 2001).

Both types of References impact the results. Landmarks may overestimate an algorithms performance since they only consider distinct locations; they do not indicate how an algorithm does in more homogeneous regions of the images. Artificial deformations may overestimate the performance of a method since they do not provide errors that are as complex as those from real deformations.

4.6.3.1. Landmarks

Most authors do not consider the inter- or intra-observer variability that exists from the manual selection of landmarks. This variability puts a limit on the accuracy of the landmark-based validation. Eppenhof and Pluim qualitatively showed the performance of their method on image pairs with the worst and best inter-observer variability (Eppenhof and Pluim, 2018). Meunzing et al. only considered landmarks that have less than 2 mm inter-observer variability (Muenzing et al., 2012).

In the context of lung CT registrations, Castillo et al. recommended that over 1000 uniformly spaced manually identified landmarks are used in the validation of registration algorithms (R. Castillo et al., 2009). Although image registration and estimating registration error (or confidence) are not synonymous, it is likely that a similar recommendation can be made to validate methods of error (or confidence) estimation. Of the included studies, the typical number of landmark points (from lung CT databases) ranges from 100 to more than 300 (see Table 3.1). Castillo et al. note that if landmarks are not uniformly spaced, or if they are "*restricted to highly selective features*," the true registration error is at risk of being underestimated (R. Castillo et al., 2009).

Lastly, ensuring a balanced or representative training set for different size errors is more challenging when using landmarks. This has implications for the results. For example, Sokooti et al. had many more lower error samples (Sokooti et al., 2019b). Their model did not perform as well estimating larger errors. The overall mean absolute error is the average for all landmarks, which is biased toward the better performance on the lower error samples. If this method was compared to an equally performing method that was evaluated on a balanced set of samples, it would appear better.

4.6.3.2. Artificial Deformations

Some authors opted to create simple deformations to validate their method, while others used more complex, and sometimes realistic-inspired, artificial deformations. The impact this has is yet to be studied. However, it is likely that the results obtained from simple artificial deformations overestimates the true ability of an algorithm since it is tested on simpler errors.

4.6.3.3. Future Work and Suggestions

There is a great deal left to be explored on the impact a Reference (landmarks or artificial deformations) has on the results of a validation procedure. The issue of the quality and quantity of landmarks, as well as the impact of inter-observer variability, should be investigated. How the results of a landmark-based validation generalize to the rest of the image would be a fruitful area for further work. The effect of the complexity of an artificial deformation should also be investigated. Furthermore, the degree to which realistic artificial deformations are actually realistic would be an interesting area to explore.

We suggest using both artificial deformations and landmarks in the validation of an algorithm. This reaps the benefits of both methods. Eppenhof and Pluim followed this approach

(Eppenhof and Pluim, 2018). Their results for artificial deformations and landmarks are comparable, with root mean square differences between landmark-based known error and estimated error being slightly larger. They ascribe this discrepancy to inter-observer uncertainty in the landmarks, similarity of the artificially deformed testing data to the training data and the artificially deformed images being from the same image.

Finally, the method of Murphy et al. to semi-automatically identify landmark points may be of great use to find large sets of landmarks in new data (Murphy et al., 2011).

4.6.4. Model Mismatch

Recall that Model Mismatch occurs when the transformation used to artificially deform images does not match the transformation used to register the images (Hub et al., 2009). Model Mismatch is desirable when validating an algorithm. If the transformations are the same, then the registration is more likely to easily revert the artificial deformation (Fitzpatrick, 2001) and yield simpler errors for the error or confidence estimation algorithm to be evaluated on. This would overestimate the ability of the error or confidence estimation algorithm, and not reveal how it does in the case of Model Mismatch, which is more likely in real registrations. When Hub et al. created their realistic (artificially deformed) lung CT data, they ensured that different types of transformations were used to create the artificial deformations than were used in the registration, thus ensuring Model Mismatch (Hub et al., 2009).

Schultz et al. demonstrated that in the case of registration errors resulting from Model Mismatch, there is very low correlation between their estimated uncertainty and the known error (Schultz et al., 2019). In real registrations, the two images differ not by an artificial deformation but by the deformation caused by, for example, a physiological process. If the type of transformation used in the registration does not match the real deformation, it can be considered a case of Model Mismatch. One can presume, therefore, that the uncertainty obtained by the method of Schultz et al. would not provide a good indication of error in such a case of Model Mismatch.

The concept of Model Mismatch also applies, and is particularly important, when creating training data from artificial deformations (i.e., Section 3.2.2.3.). If an algorithm is only trained on data from one type of transformation and is subsequently evaluated on errors from this same transformation, the algorithm's ability to generalize to different types of transformations,

and combinations thereof, is not tested. Eppenhof and Pluim mention that one explanation for their machine learning model not detecting high frequency errors is that it was trained on transformations that yield lower frequency errors than the transformation used to register the images (Eppenhof and Pluim, 2018).

We suggest future research to probe the effect of Model Mismatch. We further encourage authors to be aware of and use transformations with Model Mismatch when using artificial deformations for their training and/or validation.

4.7. Limitations

4.7.1. Limitations of the Evidence Included in the Review

This review is not without some limitations. Its purpose was not meant to be a scoping review of the field, but rather a systematic review of papers that met the specific inclusion criteria. We recognize that, as a result, several papers may be excluded from this review but are nonetheless interesting methods that deserve further attention. For example, a great many probabilistic, and in particular Bayesian, registration algorithms are available but do not meet the criteria of relating the obtained uncertainty to the known registration error. We addressed some of these excluded articles in Section 3.5. Other error and confidence estimation methods can be found in the review of Paganelli et al. (Paganelli et al., 2018). Overall, the inclusion criteria led to a representative and objectively selected set of articles for the field of registration error and confidence estimation in medical images.

4.7.2. Limitations of the Review Processes Used

While the review process closely followed the PRISMA guidelines, certain limitations exist. First, we restricted our database search to the Scopus database. While other databases were considered, it is possible that we missed certain publications. However, we included records from a preliminary literature review, as well as those identified by searching through references of the included publications. The same limitation applies to the advanced search string that we used; it may have left out some relevant records, but we are confident that we would have come across them eventually. Finally, one reviewer was responsible for the majority of inclusion and exclusion decisions, and a second reviewer was consulted when necessary.

5. Conclusion

Image registration is widely used in both medical research and clinical settings. Registration errors can have immediate and pernicious consequences on downstream tasks. We performed a systematic review of the literature on methods that automatically estimate registration error, registration confidence or both. The review was structured around a taxonomy to organize and classify these different methods. Trends, benefits and drawbacks, and sources of bias of the methods, their development and validation were discussed. We provided suggestions for best practices and future research. The field of registration error and confidence estimation is blossoming, and we hope our review will help guide and structure it as it continues to grow.

6. Acknowledgements

This study was funded by grants from the Canadian Institutes of Health Research and from the Natural Sciences and Engineering Research Council of Canada. JB acknowledges funding from the Canada First Research Excellence Fund and Fonds de recherche du Québec, awarded to the Healthy Brains, Healthy Lives initiative at McGill University, and the Department of Biomedical Engineering at McGill University.

The authors would like to thank Nu Ree Lee, an Assistant Librarian at McGill University, for help in structuring the advanced literature search, and Mohammadreza Eskandari for proof reading sections of the article.

7. References

Agn, M., Van Leemput, K., 2019. Fast Nonparametric Mutual-Information-based Registration and Uncertainty Estimation, in: Greenspan, H., Tanno, R., Erdt, M., Arbel, T., Baumgartner, C., Dalca, A., Sudre, C.H., Wells, W.M., Drechsler, K., Linguraru, M.G., Oyarzun Laura, C., Shekhar, R., Wesarg, S., González Ballester, M.Á. (Eds.), Uncertainty for Safe Utilization of Machine Learning in Medical Imaging and Clinical Image-Based Procedures, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 42–51. https://doi.org/10.1007/978-3-030-32689-0_5

Amir-Khalili, A., Hamarneh, G., Zakariaee, R., Spadinger, I., Abugharbieh, R., 2017. Propagation of registration uncertainty during multi-fraction cervical cancer brachytherapy. Phys. Med. Biol. 62, 8116–8135. https://doi.org/10.1088/1361-6560/aa8b37

Armato, S.G., Doshi, D.J., Engelmann, R., Croteau, C.L., MacMahon, H., 2006. Temporal subtraction in chest radiography: automated assessment of registration accuracy. Med Phys 33, 1239–1249. https://doi.org/10.1118/1.2184441

Azcona, J.D., Huesa-Berral, C., Moreno-Jiménez, M., Barbés, B., Aristu, J.J., Burguete, J., 2019. A novel concept to include uncertainties in the evaluation of stereotactic body radiation therapy after 4D dose accumulation using deformable image registration. Med Phys 46, 4346–4355. https://doi.org/10.1002/mp.13759

Bai, X., Wang, S., Wang, B., Zhang, J., 2019. The Accuracy Heart Dosimetric Study of Left-breast Cancer Radio-therapy using Deformable Image Registration, in: Proceedings of the 2019 6th International Conference on Bioinformatics Research and Applications, ICBRA '19. Association for Computing Machinery, New York, NY, USA, pp. 73–77. https://doi.org/10.1145/3383783.3383803

Bansal, R., Staib, L.H., Laine, A.F., Xu, D., Liu, J., Posecion, L.F., Peterson, B.S., 2009. Calculation of the confidence intervals for transformation parameters in the registration of medical images. Med Image Anal 13, 215–233. https://doi.org/10.1016/j.media.2008.09.002

Bayer, S., Spiske, U., Luo, J., Geimer, T., Wells III, W.M., Ostermeier, M., Fahrig, R., Nabavi, A., Bert, C., Eyüpoglo, I., Maier, A., 2020. Investigation of Feature-Based Nonrigid

Image Registration Using Gaussian Process, in: Tolxdorff, T., Deserno, T.M., Handels, H., Maier, A., Maier-Hein, K.H., Palm, C. (Eds.), Bildverarbeitung für die Medizin 2020, Informatik aktuell. Springer Fachmedien, Wiesbaden, pp. 156–162. https://doi.org/10.1007/978-3-658-29267-6_32

Bender, E.T., Hardcastle, N., Tomé, W.A., 2012. On the dosimetric effect and reduction of inverse consistency and transitivity errors in deformable image registration for dose accumulation. Med Phys 39, 272–280. https://doi.org/10.1118/1.3666948

Bender, E.T., Tomé, W.A., 2009. The utilization of consistency metrics for error analysis in deformable image registration. Phys. Med. Biol. 54, 5561–5577. https://doi.org/10.1088/0031-9155/54/18/014

Brock, K.K., Mutic, S., McNutt, T.R., Li, H., Kessler, M.L., 2017. Use of image registration and fusion algorithms and techniques in radiotherapy: Report of the AAPM Radiation Therapy Committee Task Group No. 132. Med. Phys. 44, e43–e76. https://doi.org/10.1002/mp.12256

Castillo, E., Castillo, R., Martinez, J., Shenoy, M., Guerrero, T., 2009. Four-dimensional deformable image registration using trajectory modeling. Phys. Med. Biol. 55, 305–327. https://doi.org/10.1088/0031-9155/55/1/018

Castillo, R., Castillo, E., Fuentes, D., Ahmad, M., Wood, A.M., Ludwig, M.S., Guerrero, T., 2013. A reference dataset for deformable image registration spatial accuracy evaluation using the COPDgene study archive. Phys. Med. Biol. 58, 2861–2877. https://doi.org/10.1088/0031-9155/58/9/2861

Castillo, R., Castillo, E., Guerra, R., Johnson, V.E., McPhail, T., Garg, A.K., Guerrero, T., 2009. A framework for evaluation of deformable image registration spatial accuracy using large landmark point sets. Phys. Med. Biol. 54, 1849–1870. https://doi.org/10.1088/0031-9155/54/7/001

Cleary, K., Anderson, J., Brazaitis, M., Devey, G., DiGioia, A., Freedman, M., Grönemeyer, D., Lathan, C., Lemke, H., Long, D., Mun, S.K., Taylor, R., 2000. Final report of the technical requirements for image-guided spine procedures Workshop, April 17-20, 1999, Ellicott City, Maryland, USA. Comput Aided Surg 5, 180–215. https://doi.org/10.1002/1097-0150(2000)5:3<180::AID-IGS6>3.0.CO;2-C

Cobzas, D., Sen, A., 2011. Random walks for deformable image registration. Med Image Comput Comput Assist Interv 14, 557–565. https://doi.org/10.1007/978-3-642-23629-7_68

Crum, William R., Griffin, L.D., Hawkes, D.J., 2004. Automatic Estimation of Error in Voxel-Based Registration, in: Barillot, C., Haynor, D.R., Hellier, P. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2004, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 821–828. https://doi.org/10.1007/978-3-540-30135-6_100

Crum, W.R., Griffin, L.D., Hill, D.L.G., Hawkes, D.J., 2003. Zen and the art of medical image registration: correspondence, homology, and quality. NeuroImage 20, 1425–1437. https://doi.org/10.1016/j.neuroimage.2003.07.014

Crum, W R, Hartkens, T., Hill, D.L.G., 2004. Non-rigid image registration: theory and practice. BJR 77, S140–S153. https://doi.org/10.1259/bjr/25329214

Dalca, A.V., Balakrishnan, G., Guttag, J., Sabuncu, M.R., 2019. Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. Medical Image Analysis 57, 226–236. https://doi.org/10.1016/j.media.2019.07.006

Danilchenko, A., Fitzpatrick, J.M., 2011. General Approach to First-Order Error Prediction in Rigid Point Registration. IEEE Trans Med Imaging 30, 679–693. https://doi.org/10.1109/TMI.2010.2091513

Datteri, R.D., Asman, A.J., Landman, B.A., Dawant, B.M., 2014. Applying the algorithm "assessing quality using image registration circuits" (AQUIRC) to multi-atlas segmentation, in: Medical Imaging 2014: Image Processing. Presented at the Medical Imaging 2014: Image Processing, SPIE, pp. 355–361. https://doi.org/10.1117/12.2043756

Datteri, R.D., Dawant, B.M., 2012a. Estimation and Reduction of Target Registration Error. Med Image Comput Comput Assist Interv 15, 139–146.

Datteri, R.D., Dawant, B.M., 2012b. Automatic Detection of the Magnitude and Spatial Location of Error in Non-rigid Registration, in: Dawant, B.M., Christensen, G.E., Fitzpatrick, J.M., Rueckert, D. (Eds.), Biomedical Image Registration, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 21–30. https://doi.org/10.1007/978-3-642-31340-0_3

Datteri, R.D., Dawant, B.M., 2012c. Estimation of rigid-body registration quality using registration networks, in: Medical Imaging 2012: Image Processing. Presented at the

Medical Imaging 2012: Image Processing, SPIE, pp. 366–377. https://doi.org/10.1117/12.911556

Datteri, R.D., Liu, Y., D'Haese, P.-F., Dawant, B.M., 2015. Validation of a nonrigid registration error detection algorithm using clinical MRI brain data. IEEE Trans Med Imaging 34, 86–96. https://doi.org/10.1109/TMI.2014.2344911

Denis de Senneville, B., Manjón, J.V., Coupé, P., 2020. RegQCNET: Deep quality control for image-to-template brain MRI affine registration. Phys Med Biol 65, 225022. https://doi.org/10.1088/1361-6560/abb6be

Eppenhof, K.A.J., Pluim, J.P.W., 2018. Error estimation of deformable image registration of pulmonary CT scans using convolutional neural networks. Journal of Medical Imaging 5, 1–1. https://doi.org/10.1117/1.jmi.5.2.024003

Fedorov, A., Billet, E., Prastawa, M., Gerig, G., Radmanesh, A., Warfield, S.K., Kikinis, R., Chrisochoides, N., 2008. Evaluation of Brain MRI Alignment with the Robust Hausdorff Distance Measures, in: Bebis, G., Boyle, R., Parvin, B., Koracin, D., Remagnino, P., Porikli, F., Peters, J., Klosowski, J., Arns, L., Chun, Y.K., Rhyne, T.-M., Monroe, L. (Eds.), Advances in Visual Computing, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 594–603. https://doi.org/10.1007/978-3-540-89639-5_57

Fedorov, A., Wells, W.M., Kikinis, R., Tempany, C.M., Vangel, M.G., 2014. Application of tolerance limits to the characterization of image registration performance. IEEE Trans Med Imaging 33, 1541–1550. https://doi.org/10.1109/TMI.2014.2317796

Fitzpatrick, J.M., 2001. Detecting Failure, Assessing Success, in: Hajnal, J., Hawkes, D., Hill, D. (Eds.), Medical Image Registration, Biomedical Engineering. CRC Press, pp. 117–139. https://doi.org/10.1201/9781420042474.ch6

Fitzpatrick, J.M., West, J.B., 2001. The distribution of target registration error in rigid-body point-based registration. IEEE Trans Med Imaging 20, 917–927. https://doi.org/10.1109/42.952729

Fitzpatrick, J.M., West, J.B., Maurer, C.R., 1998a. Predicting error in rigid-body point-based registration. IEEE Trans Med Imaging 17, 694–702. https://doi.org/10.1109/42.736021

Fitzpatrick, J.M., West, J.B., Maurer, C.R., 1998b. Derivation of expected registration error for point-based rigid-body registration, in: Medical Imaging 1998: Image Processing.

Presented at the Medical Imaging 1998: Image Processing, SPIE, pp. 16–27. https://doi.org/10.1117/12.310824

Galib, S.M., Lee, H.K., Guy, C.L., Riblett, M.J., Hugo, G.D., 2020. A fast and scalable method for quality assurance of deformable image registration on lung CT scans using convolutional neural networks. Med Phys 47, 99–109. https://doi.org/10.1002/mp.13890

Garlapati, R.R., Joldes, G.R., Wittek, A., Lam, J., Weisenfeld, N., Hans, A., Warfield, S.K., Kikinis, R., Miller, K., 2013. Objective Evaluation of Accuracy of Intra-Operative Neuroimage Registration, in: Wittek, A., Miller, K., Nielsen, P.M.F. (Eds.), Computational Biomechanics for Medicine. Springer, New York, NY, pp. 87–99. https://doi.org/10.1007/978-1-4614-6351-1_9

Garlapati, R.R., Mostayed, A., Joldes, G.R., Wittek, A., Doyle, B., Miller, K., 2015. Towards measuring neuroimage misalignment. Computers in Biology and Medicine 64, 12–23. https://doi.org/10.1016/j.compbiomed.2015.06.003

Gass, T., Székely, G., Goksel, O., 2015. Consistency-based rectification of nonrigid registrations. J Med Imaging (Bellingham) 2, 014005. https://doi.org/10.1117/1.JMI.2.1.014005

Gass, T., Székely, G., Goksel, O., 2014. Detection and correction of inconsistency-based errors in non-rigid registration, in: Medical Imaging 2014: Image Processing. Presented at the Medical Imaging 2014: Image Processing, SPIE, pp. 324–331. https://doi.org/10.1117/12.2042757

Geiger, A., Lenz, P., Urtasun, R., 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition. Presented at the 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3354–3361. https://doi.org/10.1109/CVPR.2012.6248074

Gibson, E., Fenster, A., Ward, A.D., 2012. Registration accuracy: how good is good enough? A statistical power calculation incorporating image registration uncertainty. Med Image Comput Comput Assist Interv 15, 643–650. https://doi.org/10.1007/978-3-642-33418-4_79

Gil, N., Lipton, M.L., Fleysher, R., 2021. Registration quality filtering improves robustness of voxel-wise analyses to the choice of brain template. NeuroImage 227, 117657. https://doi.org/10.1016/j.neuroimage.2020.117657

Gillmann, C., Saur, D., Wischgoll, T., Scheuermann, G., 2021. Uncertainty-aware Visualization in Medical Imaging - A Survey. Computer Graphics Forum 40, 665–689. https://doi.org/10.1111/cgf.14333

Glocker, B., Paragios, N., Komodakis, N., Tziritas, G., Navab, N., 2008. Optical flow estimation with uncertainties through dynamic MRFs, in: 2008 IEEE Conference on Computer Vision and Pattern Recognition. Presented at the 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. https://doi.org/10.1109/CVPR.2008.4587562

Grzech, D., Kainz, B., Glocker, B., le Folgoc, L., 2020. Image Registration via Stochastic Gradient Markov Chain Monte Carlo, in: Sudre, C.H., Fehri, H., Arbel, T., Baumgartner, C.F., Dalca, A., Tanno, R., Van Leemput, K., Wells, W.M., Sotiras, A., Papiez, B., Ferrante, E., Parisot, S. (Eds.), Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 3–12. https://doi.org/10.1007/978-3-030-60365-6_1

Gunay, G., van der Voort, S., Luu, M.H., Moelker, A., Klein, S., 2018. Local Image Registration Uncertainty Estimation Using Polynomial Chaos Expansions, in: Klein, S., Staring, M., Durrleman, S., Sommer, S. (Eds.), Biomedical Image Registration, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 115–125. https://doi.org/10.1007/978-3-319-92258-4_11

Hammers, A., Allom, R., Koepp, M.J., Free, S.L., Myers, R., Lemieux, L., Mitchell, T.N., Brooks, D.J., Duncan, J.S., 2003. Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe. Hum Brain Mapp 19, 224–247. https://doi.org/10.1002/hbm.10123

Hauler, F., Furtado, H., Jurisic, M., Polanec, S.H., Spick, C., Laprie, A., Nestle, U., Sabatini, U., Birkfellner, W., 2016. Automatic quantification of multi-modal rigid registration accuracy using feature detectors. Phys Med Biol 61, 5198–5214. https://doi.org/10.1088/0031-9155/61/14/5198

Heinrich, M.P., Jenkinson, M., Bhushan, M., Matin, T., Gleeson, F.V., Brady, S.M., Schnabel, J.A., 2012. MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration. Medical Image Analysis 16, 1423–1435. https://doi.org/10.1016/j.media.2012.05.008

Heinrich, M.P., Simpson, I.J.A., Papież, B.W., Brady, S.M., Schnabel, J.A., 2016. Deformable image registration by combining uncertainty estimates from supervoxel belief propagation. Medical Image Analysis 27, 57–71. https://doi.org/10.1016/j.media.2015.09.005

Heiselman, J.S., Miga, M.I., 2021. Strain Energy Decay Predicts Elastic Registration Accuracy From Intraoperative Data Constraints. IEEE Trans. Med. Imaging 40, 1290–1302. https://doi.org/10.1109/TMI.2021.3052523

Holden, M., 2008. A Review of Geometric Transformations for Nonrigid Body Registration. IEEE Trans. Med. Imaging 27, 111–128. https://doi.org/10.1109/TMI.2007.904691

Hu, Y., Bonmati, E., Gibson, E., Hipwell, J.H., Hawkes, D.J., Bandula, S., Pereira, S.P., Barratt, D.C., 2016. 2D-3D Registration Accuracy Estimation for Optimised Planning of Image-Guided Pancreatobiliary Interventions, in: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 516–524. https://doi.org/10.1007/978-3-319-46720-7_60

Hub, M., Karger, C.P., 2013. Estimation of the uncertainty of elastic image registration with the demons algorithm. Physics in Medicine and Biology 58, 3023–3036. https://doi.org/10.1088/0031-9155/58/9/3023

Hub, M., Kessler, M.L., Karger, C.P., 2009. A stochastic approach to estimate the uncertainty involved in B-spline image registration. IEEE Trans Med Imaging 28, 1708–1716. https://doi.org/10.1109/TMI.2009.2021063

Hub, M., Thieke, C., Kessler, M.L., Karger, C.P., 2012. A stochastic approach to estimate the uncertainty of dose mapping caused by uncertainties in b-spline registration. Med Phys 39, 2186–2192. https://doi.org/10.1118/1.3697524

IXI Dataset – Brain Development, n.d. URL http://brain-development.org/ixi-dataset/ (accessed 10.7.21).

Jannin, P., Grova, C., Maurer, C.R., 2006. Model for defining and reporting reference-based validation protocols in medical image processing. Int J CARS 1, 63–73. https://doi.org/10.1007/s11548-006-0044-6

Janoos, F., Risholm, P., Wells, W., 2012a. Bayesian Characterization of Uncertainty in Multi-modal Image Registration, in: Dawant, B.M., Christensen, G.E., Fitzpatrick, J.M., Rueckert, D. (Eds.), Biomedical Image Registration, Lecture Notes in Computer Science. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 50–59. https://doi.org/10.1007/978-3-642-31340-0_6

Janoos, F., Risholm, P., Wells, W., 2012b. Robust non-rigid registration and characterization of uncertainty, in: 2012 IEEE Workshop on Mathematical Methods in Biomedical Image Analysis. Presented at the 2012 IEEE Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA), IEEE, Breckenridge, CO, USA, pp. 4321–4326. https://doi.org/10.1109/MMBIA.2012.6164760

Kierkels, R.G.J., den Otter, L.A., Korevaar, E.W., Langendijk, J.A., van der Schaaf, A., Knopf, A.C., Sijtsema, N.M., 2018. An automated, quantitative, and case-specific evaluation of deformable image registration in computed tomography images. Phys. Med. Biol. 63, 045026. https://doi.org/10.1088/1361-6560/aa9dc2

Kim, H., Chen, J., Phillips, J., Pukala, J., Yom, S.S., Kirby, N., 2017. Validating Dose Uncertainty Estimates Produced by AUTODIRECT: An Automated Program to Evaluate Deformable Image Registration Accuracy. Technol Cancer Res Treat 16, 885–892. https://doi.org/10.1177/1533034617708076

Kim, H., Monroe, J., Yao, M., Lo, S., Ellis, R., Machtay, M., Sohn, J., 2013. SU-E-J-84: Use of Deformation Error Histogram as An Accuracy Indicator for Deformable Image Registration. Medical Physics 40, 169–169. https://doi.org/10.1118/1.4814296

Kirby, N., Chen, J., Kim, H., Morin, O., Nie, K., Pouliot, J., 2016. An automated deformable image registration evaluation of confidence tool. Phys. Med. Biol. 61, N203–N214. https://doi.org/10.1088/0031-9155/61/8/N203

Klein, A., Andersson, J., Ardekani, B.A., Ashburner, J., Avants, B., Chiang, M.-C., Christensen, G.E., Collins, D.L., Gee, J., Hellier, P., Song, J.H., Jenkinson, M., Lepage, C., Rueckert, D., Thompson, P., Vercauteren, T., Woods, R.P., Mann, J.J., Parsey, R.V., 2009. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. NeuroImage 46, 786–802. https://doi.org/10.1016/j.neuroimage.2008.12.037

Kybic, J., 2010. Bootstrap Resampling for Image Registration Uncertainty Estimation Without Ground Truth. IEEE Transactions on Image Processing 19, 64–73. https://doi.org/10.1109/TIP.2009.2030955

Kybic, J., 2008. Fast no ground truth image registration accuracy evaluation: Comparison of bootstrap and Hessian approaches, in: 2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro. Presented at the 2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, pp. 792–795. https://doi.org/10.1109/ISBI.2008.4541115

Kybic, J., Smutek, D., 2006. Image Registration Accuracy Estimation Without Ground Truth Using Bootstrap, in: Beichel, R.R., Sonka, M. (Eds.), Computer Vision Approaches to Medical Image Analysis, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 61–72. https://doi.org/10.1007/11889762_6

Le Folgoc, L., Delingette, H., Criminisi, A., Ayache, N., 2017. Quantifying Registration Uncertainty With Sparse Bayesian Modelling. IEEE Trans. Med. Imaging 36, 607–617. https://doi.org/10.1109/TMI.2016.2623608

Li, S., Glide-Hurst, C., Lu, M., Kim, J., Wen, N., Adams, J.N., Gordon, J., Chetty, I.J., Zhong, H., 2013. Voxel-based statistical analysis of uncertainties associated with deformable image registration. Phys. Med. Biol. 58, 6481–6494. https://doi.org/10.1088/0031-9155/58/18/6481

Li, Z., Kurihara, T., 2014. Evaluation of Medical Image Registration by Using High-Accuracy Image Matching Techniques, in: El-Baz, A.S., Saba, L., Suri, J. (Eds.), Abdomen and Thoracic Imaging: An Engineering & Clinical Perspective. Springer US, Boston, MA, pp. 489–508. https://doi.org/10.1007/978-1-4614-8498-1_19

Lin, X.-B., Li, X.-X., Guo, D.-M., 2019. Registration Error and Intensity Similarity Based Label Fusion for Segmentation. IRBM 40, 78–85. https://doi.org/10.1016/j.irbm.2019.02.001

Lotfi Mahyari, T., 2013. Uncertainty in probabilistic image registration (Thesis). Applied Sciences: School of Computing Science.

Lotfi, T., Tang, L., Andrews, S., Hamarneh, G., 2013. Improving Probabilistic Image Registration via Reinforcement Learning and Uncertainty Evaluation, in: Wu, G., Zhang, D., Shen, D., Yan, P., Suzuki, K., Wang, F. (Eds.), Machine Learning in Medical

Imaging, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 187–194. https://doi.org/10.1007/978-3-319-02267-3_24

Luo, J., Frisken, S., Wang, D., Golby, A., Sugiyama, M., Wells, W., 2020. Are Registration Uncertainty and Error Monotonically Associated? Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 12263 LNCS, 264–274. https://doi.org/10.1007/978-3-030-59716-0_26

Luo, J., Sedghi, A., Popuri, K., Cobzas, D., Zhang, M., Preiswerk, F., Toews, M., Golby, A., Sugiyama, M., Wells, W.M., Frisken, S., 2019. On the Applicability of Registration Uncertainty, in: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.-T., Khan, A. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 410–419. https://doi.org/10.1007/978-3-030-32245-8_46

Machado, I., Toews, M., Luo, J., Unadkat, P., Essayed, W., George, E., Teodoro, P., Carvalho, H., Martins, J., Golland, P., Pieper, S., Frisken, S., Golby, A., Wells, W., 2018. Non-rigid registration of 3D ultrasound for neurosurgery using automatic feature detection and matching. Int J Comput Assist Radiol Surg 13, 1525–1538. https://doi.org/10.1007/s11548-018-1786-7

Maintz, J.B.A., Viergever, M.A., 1998. A survey of medical image registration. Medical Image Analysis 2, 1–36. https://doi.org/10.1016/S1361-8415(01)80026-8

Maurer, C.R., Fitzpatrick, J.M., 1993. A review of medical image registration. Interactive image-guided neurosurgery 1, 17–44.

Mazaheri, S., Sulaiman, P.S., Wirza, R., Dimon, M.Z., Khalid, F., Tayebi, R.M., 2015. Uncertainty Estimation for Improving Accuracy of Non-rigid Registration in Cardiac Images, in: Chbeir, R., Manolopoulos, Y., Maglogiannis, I., Alhajj, R. (Eds.), Artificial Intelligence Applications and Innovations, IFIP Advances in Information and Communication Technology. Springer International Publishing, Cham, pp. 19–28. https://doi.org/10.1007/978-3-319-23868-5_2

MICCAI BRATS 2012 [WWW Document], n.d. URL http://www2.imm.dtu.dk/projects/BRATS2012/index.html (accessed 10.7.21).

Muenzing, S.E.A., Murphy, K., Ginneken, B. van, Pluim, J.P.W., 2009. Automatic detection of registration errors for quality assessment in medical image registration, in: Medical

Imaging 2009: Image Processing. Presented at the SPIE Medical Imaging, SPIE, Lake Buena Vista, FL, pp. 205–213. https://doi.org/10.1117/12.812659

Muenzing, S.E.A., van Ginneken, B., Murphy, K., Pluim, J.P.W., 2012. Supervised quality assessment of medical image registration: Application to intra-patient CT lung registration. Medical Image Analysis 16, 1521–1531. https://doi.org/10.1016/j.media.2012.06.010

Muenzing, S.E.A., van Ginneken, B., Viergever, M.A., Pluim, J.P.W., 2014. DIRBoost–An algorithm for boosting deformable image registration: Application to lung CT intra-subject registration. Medical Image Analysis 18, 449–459. https://doi.org/10.1016/j.media.2013.12.006

Murphy, K., van Ginneken, B., Klein, S., Staring, M., de Hoop, B.J., Viergever, M.A., Pluim, J.P.W., 2011. Semi-automatic construction of reference standards for evaluation of image registration. Med Image Anal 15, 71–84. https://doi.org/10.1016/j.media.2010.07.005

Nanayakkara, N.D., Chiu, B., Fenster, A., 2009. A surface-based metric for registration error quantification, in: 2009 International Conference on Industrial and Information Systems (ICIIS). Presented at the 2009 International Conference on Industrial and Information Systems (ICIIS), pp. 349–353. https://doi.org/10.1109/ICIINFS.2009.5429837

Neylon, J., Min, Y., Low, D.A., Santhanam, A., 2017. A neural network approach for fast, automated quantification of DIR performance. Med Phys 44, 4126–4138. https://doi.org/10.1002/mp.12321

Nix, M.G., Prestwich, R.J.D., Speight, R., 2017. Automated, reference-free local error assessment of multimodal deformable image registration for radiotherapy in the head and neck. Radiotherapy and Oncology 125, 478–484. https://doi.org/10.1016/j.radonc.2017.10.004

Obeidat, M., Narayanasamy, G., Cline, K., Stathakis, S., Pouliot, J., Kim, H., Kirby, N., 2016. Comparison of different QA methods for deformable image registration to the known errors for prostate and head-and-neck virtual phantoms. Biomed. Phys. Eng. Express 2, 067002. https://doi.org/10.1088/2057-1976/2/6/067002

Ouzzani, M., Hammady, H., Fedorowicz, Z., Elmagarmid, A., 2016. Rayyan—a web and mobile app for systematic reviews. Systematic Reviews 5. https://doi.org/10.1186/s13643-016-0384-4

Paganelli, C., Meschini, G., Molinelli, S., Riboldi, M., Baroni, G., 2018. "Patient-specific validation of deformable image registration in radiation therapy: Overview and caveats." Medical Physics 15.

Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., Chou, R., Glanville, J., Grimshaw, J.M., Hróbjartsson, A., Lalu, M.M., Li, T., Loder, E.W., Mayo-Wilson, E., McDonald, S., McGuinness, L.A., Stewart, L.A., Thomas, J., Tricco, A.C., Welch, V.A., Whiting, P., Moher, D., 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ n71. https://doi.org/10.1136/bmj.n71

Pang, A.T., Wittenbrink, C.M., Lodha, S.K., 1997. Approaches to uncertainty visualization. The Visual Computer 13, 370–390. https://doi.org/10.1007/s003710050111

Park, S.B., Kim, H., Yao, M., Ellis, R., Machtay, M., Sohn, J.W., 2012. SU-E-J-87: Building Deformation Error Histogram and Quality Assurance of Deformable Image Registration. Med Phys 39, 3672. https://doi.org/10.1118/1.4734922

Pennec, X., Guttmann, C.R.G., Thirion, J.-P., 1998. Feature-based registration of medical images: Estimation and validation of the pose accuracy, in: Wells, W.M., Colchester, A., Delp, S. (Eds.), Medical Image Computing and Computer-Assisted Intervention — MICCAI'98, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 1107–1114. https://doi.org/10.1007/BFb0056300

Pennec, X., Thirion, J.-P., 1997. A Framework for Uncertainty and Validation of 3-D Registration Methods Based on Points and Frames. International Journal of Computer Vision 25, 203–229. https://doi.org/10.1023/A:1007976002485

Pizzorni Ferrarese, F., Simonetti, F., Foroni, R.I., Menegaz, G., 2014. A Framework for the Objective Assessment of Registration Accuracy. International Journal of Biomedical Imaging 2014, e128324. https://doi.org/10.1155/2014/128324

Pluim, J.P.W., Muenzing, S.E.A., Eppenhof, K.A.J., Murphy, K., 2016. The truth is hard to make: Validation of medical image registration, in: 23rd International Conference on Pattern Recognition (ICPR). IEEE, Cancun, pp. 2294–2300. https://doi.org/10.1109/ICPR.2016.7899978

Rathbone, J., Carter, M., Hoffmann, T., Glasziou, P., 2015. Better duplicate detection for systematic reviewers: evaluation of Systematic Review Assistant-Deduplication Module. Syst Rev 4, 6. https://doi.org/10.1186/2046-4053-4-6

Ren, J., Green, M., Huang, X., Abdalbari, A., 2017. Automatic error correction using adaptive weighting for vessel-based deformable image registration. Biomed. Eng. Lett. 7, 173–181. https://doi.org/10.1007/s13534-017-0020-9

Ribeiro, A.S., Nutt, D.J., McGonigle, J., 2015. Which Metrics Should Be Used in Non-linear Registration Evaluation?, in: Navab, N., Hornegger, J., Wells, W.M., Frangi, A. (Eds.), Medical Image Computing and Computer-Assisted Intervention -- MICCAI 2015, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 388–395. https://doi.org/10.1007/978-3-319-24571-3_47

Risholm, P., Balter, J., Wells, W.M., 2011. Estimation of delivered dose in radiotherapy: the influence of registration uncertainty. Med Image Comput Comput Assist Interv 14, 548–555. https://doi.org/10.1007/978-3-642-23623-5_69

Risholm, P., Janoos, F., Norton, I., Golby, A.J., Wells, W.M., 2013. Bayesian characterization of uncertainty in intra-subject non-rigid registration. Medical Image Analysis 17, 538–555. https://doi.org/10.1016/j.media.2013.03.002

Risholm, P., Pieper, S., Samset, E., Wells, W.M., 2010a. Summarizing and Visualizing Uncertainty in Non-rigid Registration, in: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2010, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 554–561. https://doi.org/10.1007/978-3-642-15745-5_68

Risholm, P., Samset, E., Wells, W., 2010b. Bayesian estimation of deformation and elastic parameters in non-rigid registration. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 6204 LNCS, 104–115. https://doi.org/10.1007/978-3-642-14366-3_10

Rohlfing, T., 2012. Image Similarity and Tissue Overlaps as Surrogates for Image Registration Accuracy: Widely Used but Unreliable. IEEE Trans. Med. Imaging 31, 153–163. https://doi.org/10.1109/TMI.2011.2163944

Rohlfing, T., Avants, B., 2012. “Nonparametric Local Smoothing” is not image registration. BMC Res Notes 5, 610. https://doi.org/10.1186/1756-0500-5-610

Saleh, Z.H., Apte, A.P., Sharp, G.C., Shusharina, N.P., Wang, Y., Veeraraghavan, H., Thor, M., Muren, L.P., Rao, S.S., Lee, N.Y., Deasy, J.O., 2014. The distance discordance metric - A novel approach to quantifying spatial uncertainties in intra- and inter-patient deformable image registration. Phys Med Biol 59, 733–746. https://doi.org/10.1088/0031-9155/59/3/733

Saygili, G., 2021. Predicting medical image registration error through independent directions. SIViP 15, 223–230. https://doi.org/10.1007/s11760-020-01784-3

Saygili, G., 2020. Predicting medical image registration error with block-matching using three orthogonal planes approach. Signal, Image and Video Processing 14, 1099–1106. https://doi.org/10.1007/s11760-020-01650-2

Saygili, G., 2018. Local-search based prediction of medical image registration error, in: Nishikawa, R.M., Samuelson, F.W. (Eds.), Medical Imaging 2018: Image Perception, Observer Performance, and Technology Assessment. Presented at the Image Perception, Observer Performance, and Technology Assessment, SPIE, Houston, United States, p. 49. https://doi.org/10.1117/12.2293740

Saygili, G., Staring, M., Hendriks, E.A., 2016. Confidence Estimation for Medical Image Registration Based On Stereo Confidence. IEEE TRANSACTIONS ON MEDICAL IMAGING 35, 539–549.

Schestowitz, R., Twining, C.J., Cootes, T., Petrovic, V., Taylor, C.J., Crum, W.R., 2006. Assessing the accuracy of non-rigid registration with and without ground truth, in: 3rd IEEE International Symposium on Biomedical Imaging: Nano to Macro, 2006. Presented at the 3rd IEEE International Symposium on Biomedical Imaging: Nano to Macro, 2006., pp. 836–839. https://doi.org/10.1109/ISBI.2006.1625048

Schlachter, M., Fechter, T., Jurisic, M., Schimek-Jasch, T., Oehlke, O., Adebahr, S., Birkfellner, W., Nestle, U., Buhler, K., 2016. Visualization of Deformable Image Registration Quality Using Local Image Dissimilarity. IEEE Trans. Med. Imaging 35, 2319–2328. https://doi.org/10.1109/TMI.2016.2560942

Schreibmann, E., Pantalone, P., Waller, A., Fox, T., 2012. A measure to evaluate deformable registration fields in clinical settings. J Appl Clin Med Phys 13, 3829. https://doi.org/10.1120/jacmp.v13i5.3829

Schultz, S., Handels, H., Ehrhardt, J., 2018. A multilevel Markov Chain Monte Carlo approach for uncertainty quantification in deformable registration, in: Medical Imaging 2018: Image Processing. Presented at the Medical Imaging 2018: Image Processing, SPIE, Houston, United States, pp. 162–169. https://doi.org/10.1117/12.2293588

Schultz, S., Krüger, J., Handels, H., Ehrhardt, J., 2019. Bayesian inference for uncertainty quantification in point-based deformable image registration, in: Medical Imaging 2019: Image Processing. Presented at the Medical Imaging 2019: Image Processing, SPIE, San Diego, United States, pp. 459–466. https://doi.org/10.1117/12.2512988

Sedghi, A., Kapur, T., Luo, J., Mousavi, P., Wells, W.M., 2019. Probabilistic Image Registration via Deep Multi-class Classification: Characterizing Uncertainty, in: Greenspan, H., Tanno, R., Erdt, M., Arbel, T., Baumgartner, C., Dalca, A., Sudre, C.H., Wells, W.M., Drechsler, K., Linguraru, M.G., Oyarzun Laura, C., Shekhar, R., Wesarg, S., González Ballester, M.Á. (Eds.), Uncertainty for Safe Utilization of Machine Learning in Medical Imaging and Clinical Image-Based Procedures, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 12–22. https://doi.org/10.1007/978-3-030-32689-0_2

Shams, R., Xiao, Y., Hebert, F., Abramowitz, M., Brooks, R., Rivaz, H., 2018. Assessment of Rigid Registration Quality Measures in Ultrasound-Guided Radiotherapy. IEEE Trans Med Imaging 37, 428–437. https://doi.org/10.1109/TMI.2017.2755695

Shattuck, D.W., Mirza, M., Adisetiyo, V., Hojatkashani, C., Salamon, G., Narr, K.L., Poldrack, R.A., Bilder, R.M., Toga, A.W., 2008. Construction of a 3D probabilistic atlas of human cortical structures. NeuroImage 39, 1064–1080. https://doi.org/10.1016/j.neuroimage.2007.09.031

Simpson, I.J.A., Cardoso, M.J., Modat, M., Cash, D.M., Woolrich, M.W., Andersson, J.L.R., Schnabel, J.A., Ourselin, S., Alzheimer's Disease Neuroimaging Initiative, 2015. Probabilistic non-linear registration with spatially adaptive regularisation. Med Image Anal 26, 203–216. https://doi.org/10.1016/j.media.2015.08.006

Simpson, I.J.A., Schnabel, J.A., Groves, A.R., Andersson, J.L.R., Woolrich, M.W., 2012. Probabilistic inference of regularisation in non-rigid registration. Neuroimage 59, 2438–2451. https://doi.org/10.1016/j.neuroimage.2011.09.002

Simpson, I.J.A., Woolrich, M.W., Andersson, J.L.R., Groves, A.R., Schnabel, J.A., 2013a. Ensemble Learning Incorporating Uncertain Registration. IEEE Transactions on Medical Imaging 32, 748–756. https://doi.org/10.1109/TMI.2012.2236651

Simpson, I.J.A., Woolrich, M.W., Cardoso, M.J., Cash, D.M., Modat, M., Schnabel, J.A., Ourselin, S., 2013b. A Bayesian Approach for Spatially Adaptive Regularisation in Non-rigid Registration, in: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 10–18. https://doi.org/10.1007/978-3-642-40763-5_2

Simpson, I.J.A., Woolrich, M.W., Groves, A.R., Schnabel, J.A., 2011. Longitudinal Brain MRI Analysis with Uncertain Registration, in: Fichtinger, G., Martel, A., Peters, T. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2011, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 647–654. https://doi.org/10.1007/978-3-642-23629-7_79

Sofka, M., Stewart, C.V., 2008. Location Registration and Recognition (LRR) for Longitudinal Evaluation of Corresponding Regions in CT Volumes, in: Metaxas, D., Axel, L., Fichtinger, G., Székely, G. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2008, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 989–997. https://doi.org/10.1007/978-3-540-85990-1_119

Sokooti, H., de Vos, B., Berendsen, F., Ghafoorian, M., Yousefi, S., Lelieveldt, B.P.F., Isgum, I., Staring, M., 2019a. 3D Convolutional Neural Networks Image Registration Based on Efficient Supervised Learning from Artificial Deformations. arXiv:1908.10235 [cs, eess].

Sokooti, H., de Vos, B., Berendsen, F., Lelieveldt, B.P.F., Išgum, I., Staring, M., 2017. Nonrigid Image Registration Using Multi-scale 3D Convolutional Neural Networks, in: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (Eds.), Medical Image Computing and Computer Assisted Intervention − MICCAI 2017, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 232–239. https://doi.org/10.1007/978-3-319-66182-7_27

Sokooti, H., Saygili, G., Glocker, B., Lelieveldt, B.P.F., Staring, M., 2019b. Quantitative error prediction of medical image registration using regression forests. Medical Image Analysis 56, 110–121. https://doi.org/10.1016/j.media.2019.05.005

Sokooti, H., Saygili, G., Glocker, B., Lelieveldt, B.P.F., Staring, M., 2016. Accuracy estimation for medical image registration using regression forests. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 9902 LNCS, 107–115. https://doi.org/10.1007/978-3-319-46726-9_13

Sokooti, H., Yousefi, S., Elmahdy, M.S., Lelieveldt, B.P.F., Staring, M., 2021. Hierarchical Prediction of Registration Misalignment Using a Convolutional LSTM: Application to Chest CT Scans. IEEE Access 9, 62008–62020. https://doi.org/10.1109/ACCESS.2021.3074124

Sotiras, A., Davatzikos, C., Paragios, N., 2013. Deformable Medical Image Registration: A Survey. IEEE Trans. Med. Imaging 32, 1153–1190. https://doi.org/10.1109/TMI.2013.2265603

Staring, M., Bakker, M.E., Stolk, J., Shamonin, D.P., Reiber, J.H.C., Stoel, B.C., 2014. Towards local progression estimation of pulmonary emphysema using CT. Medical Physics 41, 021905. https://doi.org/10.1118/1.4851535

Stolk, J., Putter, H., Bakker, E.M., Shaker, S.B., Parr, D.G., Piitulainen, E., Russi, E.W., Grebski, E., Dirksen, A., Stockley, R.A., Reiber, J.H.C., Stoel, B.C., 2007. Progression parameters for emphysema: A clinical investigation. Respiratory Medicine 101, 1924–1930. https://doi.org/10.1016/j.rmed.2007.04.016

Thompson, S., Schneider, C., Bosi, M., Gurusamy, K., Ourselin, S., Davidson, B., Hawkes, D., Clarkson, M.J., 2018. In vivo estimation of target registration errors during augmented reality laparoscopic surgery. Int J Comput Assist Radiol Surg 13, 865–874. https://doi.org/10.1007/s11548-018-1761-3

Tyyger, M., Nix, M., Al-Qaisieh, B., Teo, M.T., Speight, R., 2020. Identification and separation of rigid image registration error sources, demonstrated for MRI-only image guided radiotherapy. Biomed Phys Eng Express 6, 035032. https://doi.org/10.1088/2057-1976/ab81ad

Vaman, C., Staub, D., Williamson, J., Murphy, M.J., 2010. A method to map errors in the deformable registration of 4DCT images. Med Phys 37, 5765–5776. https://doi.org/10.1118/1.3488983

Vandemeulebroucke, J., Rit, S., Kybic, J., Clarysse, P., Sarrut, D., 2011. Spatiotemporal motion estimation for respiratory-correlated imaging of the lungs. Med Phys 38, 166–178. https://doi.org/10.1118/1.3523619

Vandemeulebroucke, J., Sarrut, D., Clarysse, P., 2007. The POPI-model, a point-validated pixel-based breathing thorax model. Proceeding of the XVth ICCR Conference 8.

Vickress, J., Battista, J., Barnett, R., Yartsev, S., 2017. Representing the dosimetric impact of deformable image registration errors. Phys Med Biol 62, N391–N403. https://doi.org/10.1088/1361-6560/aa8133

Viergever, M.A., Maintz, J.B.A., Klein, S., Murphy, K., Staring, M., Pluim, J.P.W., 2016. A survey of medical image registration – under review. Medical Image Analysis 33, 140–144. https://doi.org/10.1016/j.media.2016.06.030

Vishnevskiy, V., Gass, T., Székely, G., Tanner, C., Goksel, O., 2015. Unsupervised detection of local errors in image registration, in: 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI). Presented at the 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI), pp. 841–844. https://doi.org/10.1109/ISBI.2015.7164002

Wang, H.S., Feng, D., Yeh, E., Huang, S.C., 2001. Objective assessment of image registration results using statistical confidence intervals. IEEE Transactions on Nuclear Science 48, 106–110. https://doi.org/10.1109/23.910839

Wang, J., Wells, W.M., Golland, P., Zhang, M., 2019. Registration uncertainty quantification via low-dimensional characterization of geometric deformations. Magnetic Resonance Imaging, Artificial Intelligence in MRI 64, 122–131. https://doi.org/10.1016/j.mri.2019.05.034

Wang, J., Wells, W.M., Golland, P., Zhang, M., 2018. Efficient Laplace Approximation for Bayesian Registration Uncertainty Quantification. Med Image Comput Comput Assist Interv 11070, 880–888. https://doi.org/10.1007/978-3-030-00928-1_99

Wassermann, D., Toews, M., Niethammer, M., Wells, W., 2014. Probabilistic Diffeomorphic Registration: Representing Uncertainty, in: Ourselin, S., Modat, M. (Eds.), Biomedical Image Registration, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 72–82. https://doi.org/10.1007/978-3-319-08554-8_8

Watanabe, T., Scott, C., 2012. Spatial Confidence Regions for Quantifying and Visualizing Registration Uncertainty. Biomed Image Regist Proc 7359, 120–130. https://doi.org/10.1007/978-3-642-31340-0_13

Werner, R., Duscha, C., Schmidt-Richberg, A., Ehrhardt, J., Handels, H., 2013. Assessing accuracy of non-linear registration in 4D image data using automatically detected landmark correspondences, in: Medical Imaging 2013: Image Processing. Presented at the Medical Imaging 2013: Image Processing, SPIE, pp. 264–272. https://doi.org/10.1117/12.2002454

West, J., Fitzpatrick, J.M., Wang, M.Y., Dawant, B.M., Maurer, C.R., Kessler, R.M., Maciunas, R.J., Barillot, C., Lemoine, D., Collignon, A., Maes, F., Suetens, P., Vandermeulen, D., van den Elsen, P.A., Napel, S., Sumanaweera, T.S., Harkness, B., Hemler, P.F., Hill, D.L., Hawkes, D.J., Studholme, C., Maintz, J.B., Viergever, M.A., Malandain, G., Woods, R.P., 1997. Comparison and evaluation of retrospective intermodality brain image registration techniques. J Comput Assist Tomogr 21, 554–566. https://doi.org/10.1097/00004728-199707000-00007

West, J.B., Maurer, C.R.J., 2002. Extension of target registration error theory to the composition of transforms, in: Medical Imaging 2002: Image Processing. Presented at the Medical Imaging 2002: Image Processing, SPIE, pp. 574–580. https://doi.org/10.1117/12.467200

Wu, J., Murphy, M.J., 2010. A neural network based 3D/3D image registration quality evaluator for the head-and-neck patient setup in the absence of a ground truth. Med Phys 37, 5756–5764. https://doi.org/10.1118/1.3502756

Wu, J., Samant, S.S., 2007. Novel image registration quality evaluator (RQE) with an implementation for automated patient positioning in cranial radiation therapy. Med Phys 34, 2099–2112. https://doi.org/10.1118/1.2736783

Wu, J., Samant, S.S., 2004. Registration quality evaluator: application to automated patient setup verification in radiotherapy, in: Medical Imaging 2004: Image Processing. Presented at the Medical Imaging 2004: Image Processing, SPIE, pp. 137–142. https://doi.org/10.1117/12.538080

Wu, J., Su, Z., Li, Z., 2016. A neural network-based 2D/3D image registration quality evaluator for pediatric patient setup in external beam radiotherapy. J Appl Clin Med Phys 17, 22–33. https://doi.org/10.1120/jacmp.v17i1.5235

Xiao, Y., Fortin, M., Unsgård, G., Rivaz, H., Reinertsen, I., 2017. REtroSpective Evaluation of Cerebral Tumors (RESECT): A clinical database of pre-operative MRI and intra-operative ultrasound in low-grade glioma surgeries. Medical Physics 44, 3875–3882. https://doi.org/10.1002/mp.12268

Yang, X., Kwitt, R., Styner, M., Niethammer, M., 2017. Quicksilver: Fast predictive image registration – A deep learning approach. NeuroImage 158, 378–396. https://doi.org/10.1016/j.neuroimage.2017.07.008

Yang, X., Niethammer, M., 2015. Uncertainty Quantification for LDDMM Using a Low-Rank Hessian Approximation, in: Navab, N., Hornegger, J., Wells, W.M., Frangi, A. (Eds.), Medical Image Computing and Computer-Assisted Intervention -- MICCAI 2015, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 289–296. https://doi.org/10.1007/978-3-319-24571-3_35

Žbontar, J., LeCun, Y., 2016. Stereo Matching by Training a Convolutional Neural Network to Compare Image Patches. Journal of Machine Learning Research 17, 1–32.

Zhong, H., Peters, T., Siebers, J.V., 2007. FEM-based evaluation of deformable image registration for radiation therapy. Phys. Med. Biol. 52, 4721–4738. https://doi.org/10.1088/0031-9155/52/16/001

Chapter 2: Estimating Medical Image Registration Error and Confidence: A Taxonomy and Systematic Review

– postface –

In Section 3.5 of the above chapter, we provided several examples of papers that were excluded from our review. Below, we provide a more comprehensive, though not exhaustive, list of error estimation papers for point-based rigid registrations that did not meet the criteria of our review.

Danilchenko, A., Fitzpatrick, J.M., 2011. General Approach to First-Order Error Prediction in Rigid Point Registration. IEEE Trans Med Imaging 30, 679–693. https://doi.org/10.1109/TMI.2010.2091513

Danilchenko, A., Fitzpatrick, J.M., 2010. General approach to error prediction in point registration, in: Medical Imaging 2010: Visualization, Image-Guided Procedures, and Modeling. Presented at the Medical Imaging 2010: Visualization, Image-Guided Procedures, and Modeling, SPIE, pp. 134–147. https://doi.org/10.1117/12.843847

Datteri, R.D., Dawant, B.M., 2012. Estimation and Reduction of Target Registration Error. Med Image Comput Comput Assist Interv 15, 139–146.

Ma, B., Moghari, M.H., Ellis, R.E., Abolmaesumi, P., 2010. Estimation of optimal fiducial target registration error in the presence of heteroscedastic noise. IEEE Trans Med Imaging 29, 708–723. https://doi.org/10.1109/TMI.2009.2034296

Moghari, M.H., Abolmaesumi, P., 2009. Distribution of Fiducial Registration Error in Rigid-Body Point-Based Registration. IEEE Transactions on Medical Imaging 28, 1791–1801. https://doi.org/10.1109/TMI.2009.2024208

Moghari, M.H., Abolmaesumi, P., 2008. Maximum likelihood estimation of the distribution of target registration error, in: Medical Imaging 2008: Visualization, Image-Guided Procedures, and Modeling. Presented at the Medical Imaging 2008: Visualization, Image-Guided Procedures, and Modeling, SPIE, pp. 190–201. https://doi.org/10.1117/12.768868

Moghari, M.H., Abolmaesumi, P., 2006. A High-Order Solution for the Distribution of Target Registration Error in Rigid-Body Point-Based Registration, in: Larsen, R., Nielsen, M., Sporring, J. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2006, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 603–611. https://doi.org/10.1007/11866763_74

Moghari, M.H., Ma, B., Abolmaesumi, P., 2008. A theoretical comparison of different target registration error estimators. Med Image Comput Comput Assist Interv 11, 1032–1040. https://doi.org/10.1007/978-3-540-85990-1_124

Shamir, R.R., Joskowicz, L., 2009. Worst-case analysis of target localization errors in fiducial-based rigid body registration, in: Medical Imaging 2009: Image Processing. Presented at the Medical Imaging 2009: Image Processing, SPIE, pp. 1072–1082. https://doi.org/10.1117/12.811038

Wiles, A.D., Likholyot, A., Frantz, D.D., Peters, T.M., 2008. A Statistical Model for Point-Based Target Registration Error With Anisotropic Fiducial Localizer Error. IEEE Transactions on Medical Imaging 27, 378–390. https://doi.org/10.1109/TMI.2007.908124

Chapter 3: Estimating MRI-Ultrasound Registration Error in Image-Guided Neurosurgery

– preface –

Chapter 2 systematically reviewed methods that densely estimate registration error and confidence. Trends, advantages and disadvantages, suggestions and directions for future research are provided, among several other topics, and are all presented in relation to the proposed taxonomy. An observation of Chapter 2 is that, while error estimation methods are increasingly popular, the application of these methods, though numerous in theory, is very limited. In Chapter 3, we extend this list of applications by implementing an error estimating algorithm for MRI-ultrasound registrations in Image-Guided Neurosurgery. This application represents an exciting development given the impact such an algorithm can have in Image-Guided Neurosurgery, as discussed in Chapter 3.

The algorithm's implementation, based on a proposed method discovered and reviewed in Chapter 2, is guided by the findings, discussions and recommendations of Chapter 2. An algorithm meant to verify MRI-ultrasound registrations in Image-Guided Neurosurgery should meet specific criteria. Namely, it should provide Measurements that are easily interpretable by surgeons and have fast runtimes. To meet the first requirement, error estimates are ideal given they are intuitively understood (as opposed to uncertainty or plausibility Measurements, which can be less intuitive and may only correlate with the error to a certain degree).

The taxonomy developed in Chapter 2 easily allows us to identify the appropriate class of methods: the desired Measurement type is estimated error, which only comes from Machine Learning Frameworks. As identified in Chapter 2, an advantage of Machine Learning Frameworks is that they can provide fast error estimates once trained (meeting the second criteria

listed above). Moreover, Machine Learning Frameworks and the corresponding estimates of registration error were the most popular type of method reviewed in Chapter 2. Among these methods, deep learning-based Frameworks are becoming more frequently used, with their only feature being from an Image-based Approach (directly using the image intensities). Considering all of this, we implemented the convolutional neural network proposed by Eppenhof and Pluim (Eppenhof and Pluim, 2018).

Accompanying deep learning models is a need for large datasets. The strategies for artificially deforming data, which yield the ground truth registration error for training and validation in Chapter 3, were informed by the results and discussions on artificial deformations in Chapter 2. Furthermore, Chapter 3 adheres to the suggestions of Chapter 2 regarding validating error and confidence estimation methods and the forms of bias that result from different validation choices.

Overall, Chapter 2 forms the basis for the implementation of Chapter 3.

Chapter 3: Estimating MRI-Ultrasound Registration Error in Image-Guided Neurosurgery

Joshua Bierbrier[a,b], D. Louis Collins[a,b,c]

[a] Department of Biomedical Engineering, McGill University, Montreal (QC), Canada
[b] McConnell Brain Imaging Center, Montreal Neurological Institute and Hospital, Montreal (QC), Canada
[c] Department of Neurology and Neurosurgery, McGill University, Montreal (QC), Canada

joshua.bierbrier@mail.mcgill.ca
louis.collins@mcgill.ca

Corresponding author

Joshua Bierbrier
joshua.bierbrier@mail.mcgill.ca
3801 University Street
Montreal, Quebec, Canada
H3A 2B4

Abstract

Image-Guided Neurosurgery allows surgeons to view their tools in relation to pre-operatively acquired patient images and models. Such neuronavigation systems enable safer and more efficient surgeries. Due to the phenomenon of brain shift (e.g., deformations of the brain during surgery), the pre-operative images cease to accurately reflect the true state of the brain. One technique to compensate for brain shift is to update the pre-operative image through registration to ultrasound images of the patient's brain obtained intra-operatively. Although nonlinear registration algorithms are more appropriate to model brain shift, they are not used clinically given challenges in validating these algorithms. We implemented a method to estimate MRI-ultrasound registration errors, with the goal of enabling surgeons to quantitatively assess the performance of nonlinear registrations. The algorithm is based on a sliding-window convolutional neural network that operates on a voxel-wise basis. To create training data where the true registration error is known, ultrasound images were simulated from ten pre-operative MRI images and artificially deformed. Experiments determined optimal training parameters. The model was evaluated on held-out test patients with artificial deformations up to 15 mm. The model achieved a mean absolute error of 0.849 mm and a Pearson correlation of 0.838. To the best of our knowledge, this is the first error estimating algorithm applied to MRI-ultrasound registrations. It lays the foundation for future developments and ultimately implementation on clinical neuronavigation systems. The next steps involve evaluating the model with real ultrasound data and implementing it on the neuronavigation system developed by our group.

Abbreviations

MRI: Magnetic Resonance Imaging
CT: Computed Tomography
MAE: Mean Absolute Error
US: Ultrasound
CNN: Convolutional Neural Network
mm: Millimetres
CSF: Cerebrospinal Fluid

1. Introduction

Neurosurgery is uniquely challenging. Fortunately, a combination of methodological innovation and technological advancement gave rise to Image-Guided Neurosurgery in the 1980s (Galloway, 2015, 2001). Image guided procedures allow surgeons to visualize their tools in relation to pre-operatively acquired images and patient models to enable safer, less invasive and more effective operations (Galloway, 2015, 2001; Grimson et al., 1999). The guidance provided by these tools can also permit more risky surgeries. Image-Guided Neurosurgery has become "*the standard of cared for intracranial neurosurgery*" (Galloway, 2015) and is consequently used in most neurosurgical departments (Unsgård et al., 2014).

Magnetic Resonance Imaging (MRI), given its high spatial resolution and ability to image contrast between soft tissues, is a common choice as a pre-operative image (Galloway, 2015; Miner, 2017; Sastry et al., 2017). While such pre-operative images are used in neuronavigation, their use is limited due to the phenomenon of 'brain shift' (Fedorov et al., 2014; Gerard et al., 2021) – deformations the brain undergoes during surgery (Dickhaus et al., 1997; Gerard et al., 2021, 2017; Hartkens et al., 2003; Hastreiter et al., 2004; Kelly et al., 1986; Nabavi et al., 2001).

One approach to mitigate the effect of brain shift is to image the patient again during the surgery. Intra-operative ultrasound can be used to update the pre-operative MRI image so that it matches the deformed brain (Gerard et al., 2021; Sastry et al., 2017; Unsgård et al., 2014). Central to this process is image registration. Rigid (a type of linear) image registration algorithms are used for this purpose (examples from our group include e.g., (De Nigris et al., 2012; Mercier et al., 2012b)). However, they suffer a limitation in not being able to adequately model the deformation caused by brain shift. Nonlinear registration algorithms are more appropriate for this task (Archip et al., 2007; Fedorov et al., 2014; Sastry et al., 2017) but are not clinically accepted due to the complexity of the problem and the challenges in validating and verifying these algorithms (Fedorov et al., 2014) (examples from our group include e.g., (Arbel et al., 2004; Rivaz et al., 2015, 2014; Rivaz and Collins, 2015)).

Any error in registration, whether by a linear or nonlinear registration, can have significant impacts on the patient's wellbeing. Consider the neurosurgical case of brain tumour resection. Registration error could mislead a surgeon into damaging healthy (namely, eloquent) tissue or miss resecting the full extent of the tumour. A lesser extent of resection is linked to decreased

patient survival (Lacroix et al., 2001) and damaging healthy tissue leads to functional deficits that impair the patient's quality of life.

With this in mind, an algorithm that is integrated in a neuronavigation system to estimate registration error would therefore be extremely beneficial. The development of algorithms that estimate registration error is becoming increasingly popular (Bierbrier et al., 2022). Not only could such an algorithm warn surgeons of potential error in linear registrations (making these registrations safer), but they can also instill trust in nonlinear registration algorithms (enabling these registrations to be clinically acceptable). To the best of our knowledge, a method for estimating nonlinear registration error has yet to be proposed for ultrasound-MRI registrations in Image-Guided Neurosurgery. To fill this void and to lay the groundwork for future research, we demonstrate the application of a suitable algorithm with a clinical dataset.

A registration error estimation algorithm for application in Image-Guided Neurosurgery must meet specific criteria:

- First, and most obvious, it needs to meet a certain accuracy in estimating the registration error – that is, the error in estimating the amount of mis-registration must be sufficiently small. Here, we expect to obtain sub-millimeter accuracy when estimating registration error with maximum mis-registrations of 5 millimeters (mm). This is in line with the registration errors from brain shift compensation reported in Gerard et al.'s recent review (Gerard et al., 2021).
- Second, the algorithm must run quickly enough so it does not interrupt the surgical workflow. After discussion with surgeons, delays less than a few minutes would be acceptable.
- Third, the algorithm must produce results that lend themselves to intuitive visualization by surgeons.
- Additionally, the algorithm should estimate registration error, which, by definition, is a measurement in physical units (e.g., millimeters) (Rohlfing, 2012), and should yield dense results (i.e., error estimate at every voxel location). These requirements ensure the results are interpretable by a surgeon and that they provide appropriate information for deformable (nonlinear) registrations.

Considering these criteria in more detail, the method of Eppenhof and Pluim stands out (Eppenhof and Pluim, 2018). Eppenhof and Pluim proposed a patch-wise 3D convolutional neural network (CNN) to regress registration error estimates. Their method is applied in the

context of lung computed tomography (CT) images and yields a dense error map that can easily be visualized alongside (or by overlaying with a colorwash) the results of a registration in a neuronavigation system. While Eppenhof and Pluim report a prohibitive execution time of their algorithm (~25 minutes), we believe that it can meet runtimes suitable for the operating room given appropriate computation power (i.e., a powerful graphics processing unit), efficient implementation and smaller input volumes (i.e., ultrasound vs CT). On top of providing an algorithm that meets our requirements, Eppenhof and Pluim thoroughly validated their algorithm using two types references (landmarks and artificial deformations), as we recommend in our review (Bierbrier et al., 2022). They obtained root mean square deviations (between the known and estimated registration error) of 0.51 mm and 0.61 mm for artificially deformed data and manually annotated landmarks, respectively.

Our goal is to achieve similar accuracy (mean of < |1| mm) as well as high correlations (> 0.75) by applying Eppenhof and Pluim's model to ultrasound-MRI data in the context of Image-Guided Neurosurgery. To do so, we simulate ultrasound data from MRI images of patients with brain tumours. We artificially deform the images to mimic registration error and use the resulting images to train the model. Experiments to determine the optimal number of epochs to train the model and to determine the amount of data augmentation during training are performed. Drawing on these results, the trained model is used to estimate error on the test subjects. Finally, we discuss the results of the trained model and identify areas of future research.

2. Methods

2.1. Model

Eppenhof and Pluim's network uses a sliding window CNN to produce estimates of the registration error. As input, it considers the patches around the corresponding (registered) voxels for which the error is to be estimated. Eppenhof and Pluim used patch sizes of 33x33x33, with voxels resampled to 1x1x1 mm (yielding patches of 33x33x33 mm) after empirically determined this to be an optimal size for the range and type of error their network expected. Their empirically optimized network consists of two sequences of two convolutional layers followed by max pooling and then three fully connected layers, with the final layer being the regressed (scalar) registration error. They regularized the network with batch normalization after each layer

and dropout in the fully connected layers. An advantage of following the method of Eppenhof and Pluim, over similar previous methods like that of Sokooti et al. or Muenzing et al. (Muenzing et al., 2012; Sokooti et al., 2019, 2016), is that Eppenhof and Pluim trained their network with artificially deformed data which obviates the need for manually selected landmark-based training data (Eppenhof and Pluim, 2018). In terms of the taxonomy introduced in our review article (Bierbrier et al., 2022), Eppenhof and Pluim's method uses an *Image-based Approach* that provides features for a *Machine Learning Framework* and yields a *Measurement* of the *estimated error*.

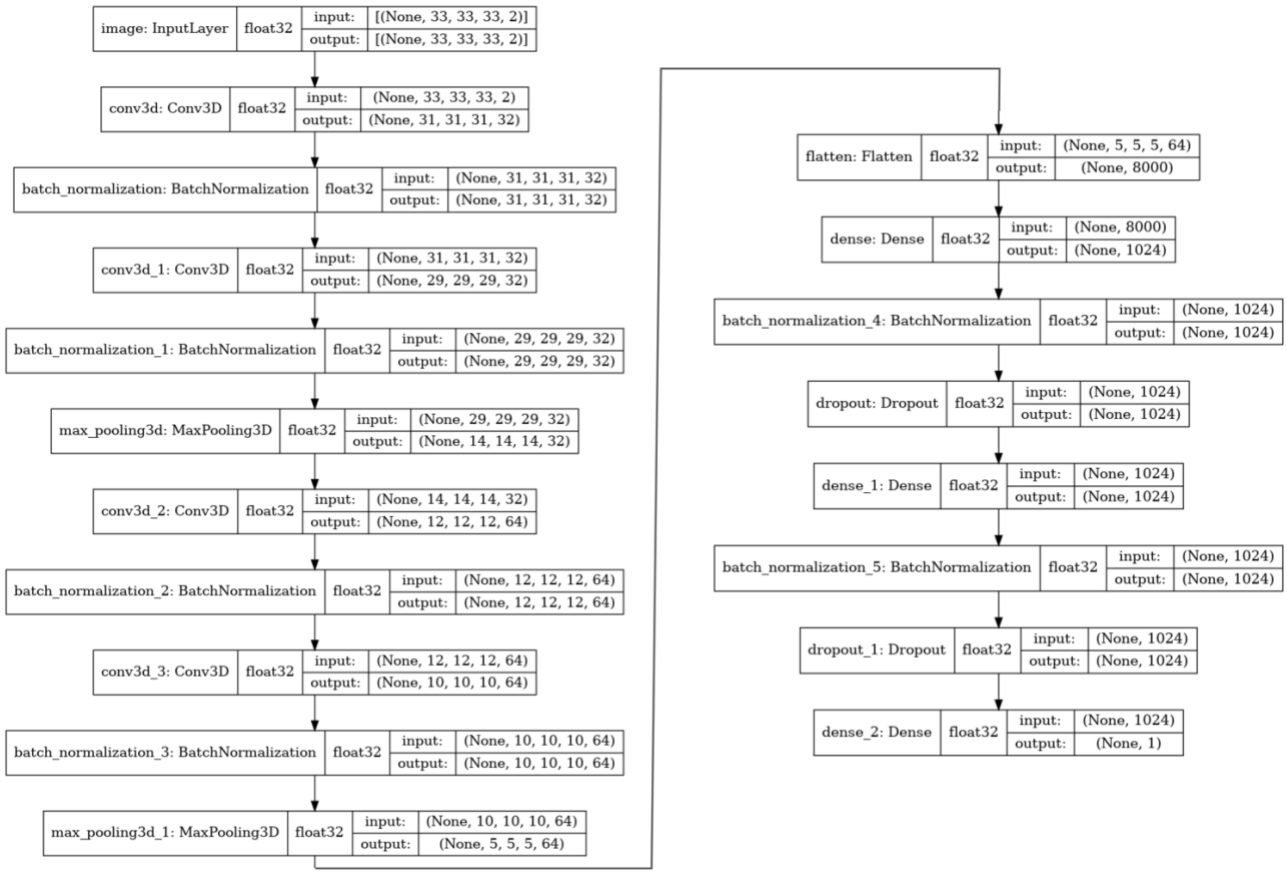

Figure 2.1. The architecture of the network proposed by Eppenhof and Pluim (adapted from (Eppenhof and Pluim, 2018)).

We make use of the architecture proposed by Eppenhof and Pluim (Eppenhof and Pluim, 2018) (Figure 2.1). However, there are several differences between our implementation and theirs. Our task – estimating error in ultrasound-MRI registrations – extends their method to the case of multimodal registrations. As noted by Eppenhof and Pluim, preparing training data for

the case of multimodal image registration following their procedure requires that one modality is simulated from the other modality to ensure that the registration error is perfectly known throughout the images. To this end, we simulate ultrasound images from the MRI images. Our process of creating artificially deformed data, both for training and validating our network, also differs. Additionally, whereas Eppenhof and Pluim restrict the errors their algorithm sees to the range of [0, 4] mm, we explicitly chose to train and test our algorithm on a much larger range of errors – [0, ≳15] mm – to robustly test its limits. Finally, on top of batch normalization and dropout, L2 regularization is employed to each layer to further ensure our model remains generalizable.

2.2. Simulated Ultrasound Images

Our dataset is composed of the BITE ('Brain Images of Tumors for Evaluation') database (Mercier et al., 2012a). It consists of pre- and post-operative MRI, and intra-operative ultrasound images of 14 patients with brain tumours. For this work, only the pre-operative T1-weighted MRI (with gadolinium) images are used. We arbitrarily split the data into training (subjects 1-10), validation (subjects 11-12) and independent test (subjects 13-14) sets. (Please note that the test subjects are used only to evaluate the final model and are never used in training or parameter optimization.)

As mentioned above, simulated ultrasound images are created from the pre-operative MRI images to ensure that the two 3D image volumes (i.e., MRI and ultrasound) are in perfect correspondence prior to being deformed. To create simulated ultrasound images we follow the approach of Mercier et al. (Mercier et al., 2012b), who simulated ultrasound images for rigid ultrasound-MRI registrations. While the technique is relatively straightforward, it was clearly sufficient for Mercier et al.; they achieved significantly improved MRI-ultrasound registration results when registering the MRI-based simulated ultrasound with the real ultrasound compared to directly registering the MRI and the real ultrasound. We summarize the method here.

The simulated ultrasound is generated by segmenting a masked MRI image into white matter, gray matter, cerebrospinal fluid (CSF) and background. The CSF is further categorized as ventricular CSF (which appears darker in ultrasound images) or sulcal CSF (which appears brighter in ultrasound images). A final class representing both vessels and high-grade tumours is obtained by thresholding the MRI image. Each class (white matter, gray matter, ventricular CSF,

sulcal CSF and vessels/high-grade tumours) is then mapped to an intensity to visually resemble ultrasound images. Finally, the intensity-remapped image is smoothed with a Gaussian filter.

Throughout, we used the same parameters as Mercier et al. (Mercier et al., 2012b). The MRI image intensities were corrected with the nonparametric nonuniform intensity normalization ('N3') algorithm (Sled et al., 1998) and were linearly normalized to match the ICBM152 template by histogram matching. The intensity corrected and normalized images were then registered to Talairach space using an affine transformation (Collins et al., 1994) and resampled to 1x1x1 mm. Brain masks were created for these images with the BEaST algorithm (Eskildsen et al., 2012). Within the brain mask, the images were nonlinearly registered to the ICBM152 template (Collins and Evans, 1997). The brain was then segmented into gray matter, white matter, CSF and background (Cocosco et al., 2003), and cortical structures (Collins et al., 1999). For all subsequent processing (here and below), we used the SimpleITK library in Python to interact with and manipulate images (Beare et al., 2018; Lowekamp et al., 2013; Yaniv et al., 2018). An example simulated ultrasound image is presented in Figure 2.2.

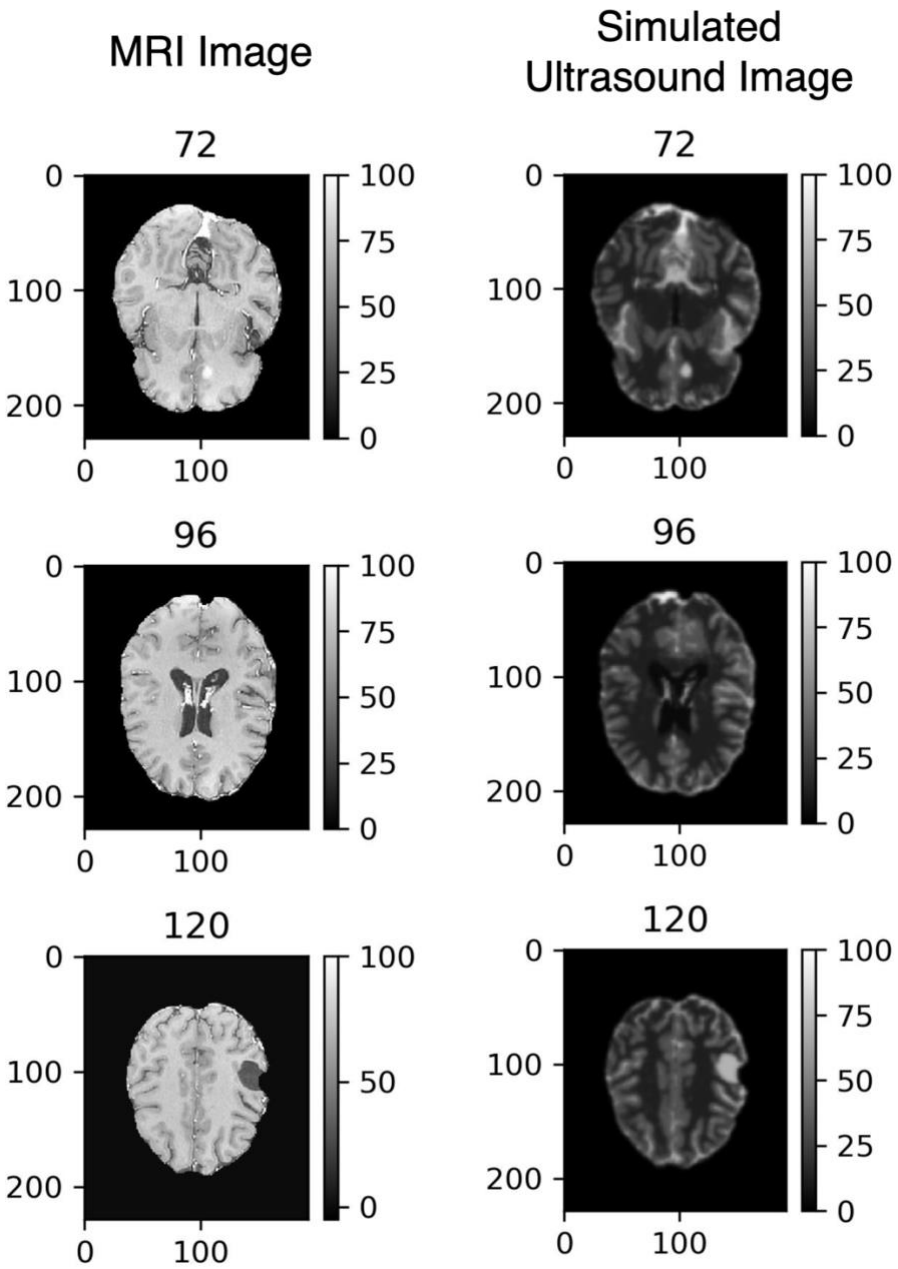

Figure 2.2. The MRI image (left) and the corresponding simulated ultrasound image (right) of subject 1. The numbers above each subplot denote their slice. The simulated ultrasound was created following the method proposed by Mercier et al. (Mercier et al., 2012b).

2.3. Training Data

2.3.1. Artificial Deformations

The original masked MRI and simulated ultrasound images, initially in perfect correspondence, are deformed using cubic B-spline-based free-form deformations (see e.g., (Rueckert et al., 1999; Rueckert and Schnabel, 2011) or (Pérez-García, 2020) for a helpful

tutorial with SimpleITK). B-spline transformations have been used by others to artificially deform training data for error estimating tasks (Lotfi et al., 2013; Sokooti et al., 2021). The B-spline transformations are defined by the displacements of a set of uniformly spaced control points (or nodes) in a 3D mesh. There are two parameters leveraged to create random artificial deformations (Figure 2.3). The first is the size of the control point mesh (see the vertical axis of Figure 2.3). A smaller number of mesh nodes implies lower frequency deformations, while a larger number of mesh nodes enables higher frequency deformations (Rueckert et al., 1999). We refer to this parameter as affecting the *frequency of the deformation*. The second parameter is the magnitude of the displacement (or perturbation) of the control points (see the horizontal axis of Figure 2.3). Larger displacements create larger deformations. We refer to this parameter as affecting the *size of the deformation*.

Given that the images are perfectly aligned to begin with, any deformation applied to either image results in misalignment which we use as a proxy for 'registration error.' Therefore, artificial deformations are applied to create training data in which the amount of registration error is known at every voxel in the two images. Figure 2.3 shows examples of deformed simulated ultrasound images.

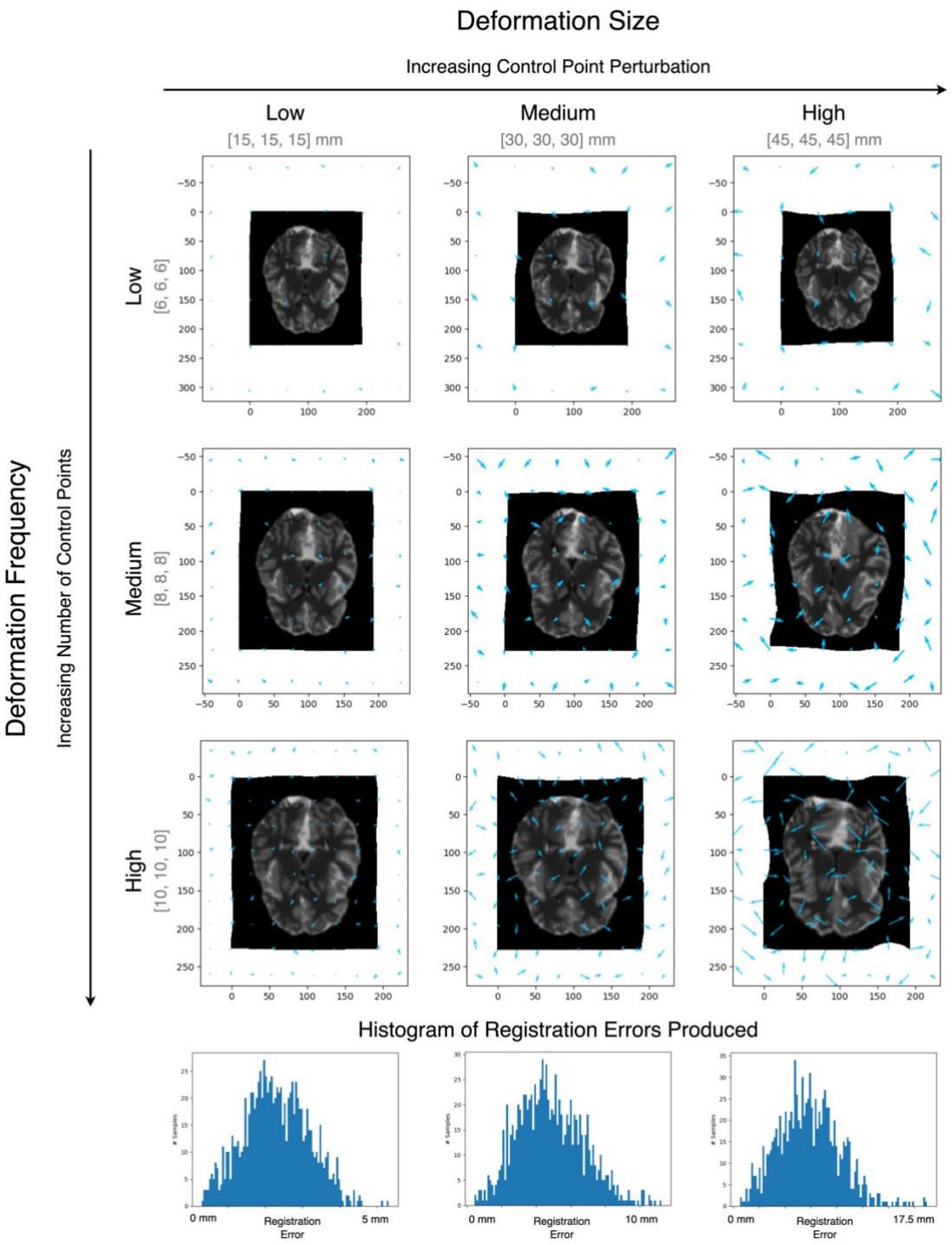

Figure 2.3. Top: Deforming the simulated ultrasound image with a different number of control points (low, medium and high frequency of the deformation shown in different rows) and varying control point perturbation (size of the deformation ranging from low, medium and high across columns). The control point displacements are shown in light blue in the X and Y directions (but note that the deformations are in 3D). Bottom: The histograms of registration errors that are produced by the levels of deformation size. Note that the horizontal axis changes for each histogram plot and that each plot contains the same number of samples. These distributions do not appreciably change for the levels of deformation frequency and, as such, they are only plotted for the levels of deformation size with medium deformation frequency.

2.3.2. Data Generator

One strategy to create training data is to apply (perhaps several) artificial deformations to the images and save corresponding patches (the samples) along with their known error (the labels) prior to training. The model is trained on this data during each epoch. This has the advantage that the training data can be generated prior to training. The disadvantage is that during each epoch, the model sees the same set of training data, which may limit its ability to generalize. Given this drawback, we opted for an online data generator strategy. In this approach, new data is generated at the beginning of each epoch. The primary advantage of using an online data generator is that the model is constantly exposed to previously unseen data, furthering its ability to generalize. It also does not require storing a large training dataset. The drawback of a data generator is the additional computational time required. With our implementation this accounted for 3-5x as much time as the training itself.

Our data generator was developed with the help of the tutorial from Amidi and Amidi (Amidi and Amidi, 2018) and works as follows. At the beginning of each epoch, a random artificial deformation is applied to the simulated ultrasound of each subject used for training and validation. This deformation is random in terms of the number of control points (affecting the frequency of the deformation) as well as the 3D displacement of each control point (affecting the size of the deformation). The control point grid is isotropic (i.e., the same size in each direction), with the size being randomly determined each deformation. The displacement of each control point (in each 3D direction) is randomly sampled from a uniform distribution of a given range centered on 0 mm. For each deformation, the range itself is random. This ensures simulated registration errors of different variety and size. For each subject, 100 corresponding patches are randomly selected from the MRI and ultrasound images (as shown in Figure 2.4), along with the known error of the central voxel of the patch pair. This yields (10 subjects x 100 patch pairs/subject = 1000) training samples per epoch. The order of the extracted patches is randomly shuffled each epoch.

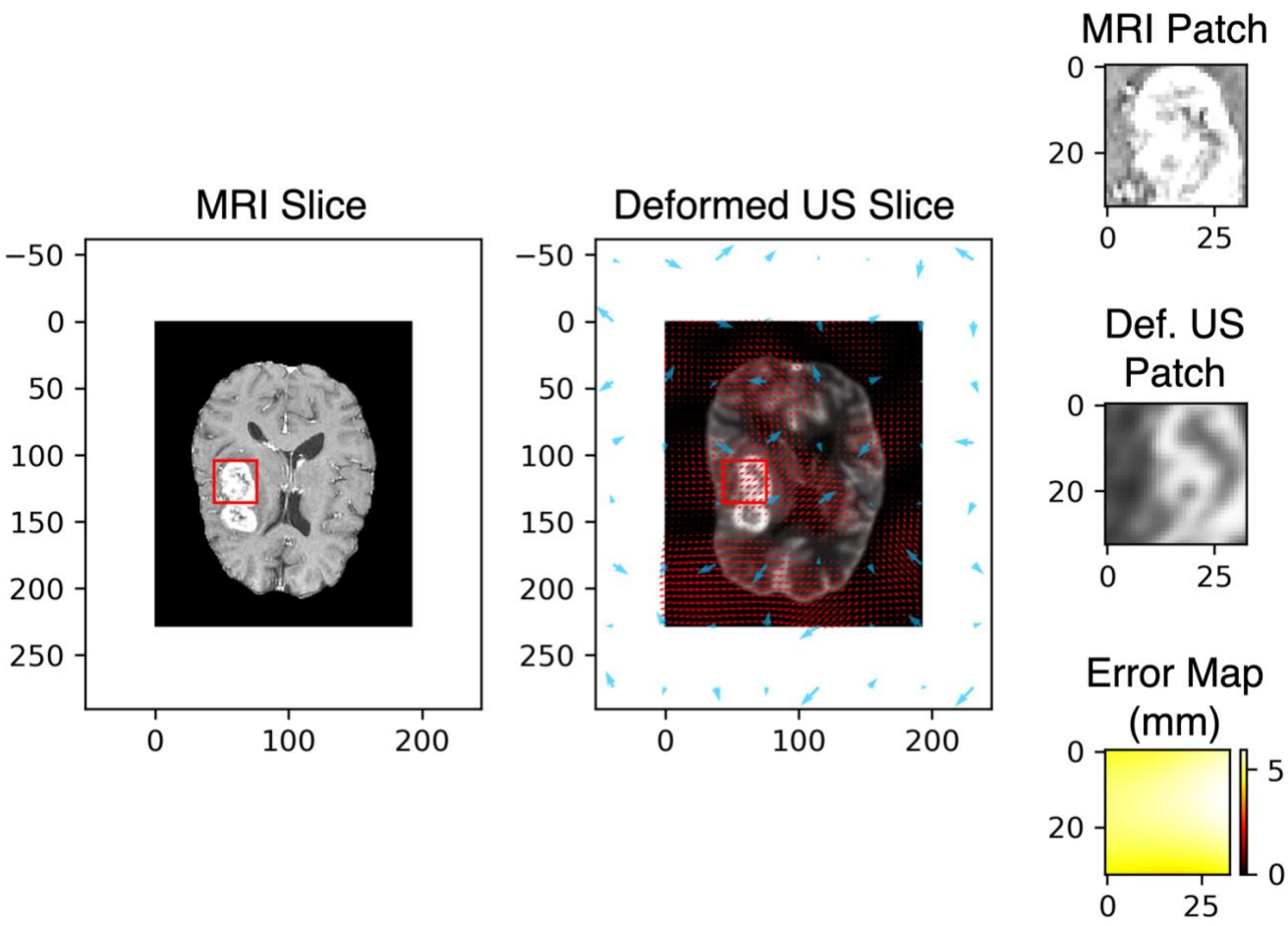

Figure 2.4. The MRI and deformed ultrasound ('US') images. The blue arrows illustrate the displacement of the control points and the red arrows display the deformation vector field of the deformation. The red boxes in the MRI and ultrasound images are the corresponding patches that are extracted. These patches are shown in the column on the right, along with the error map.

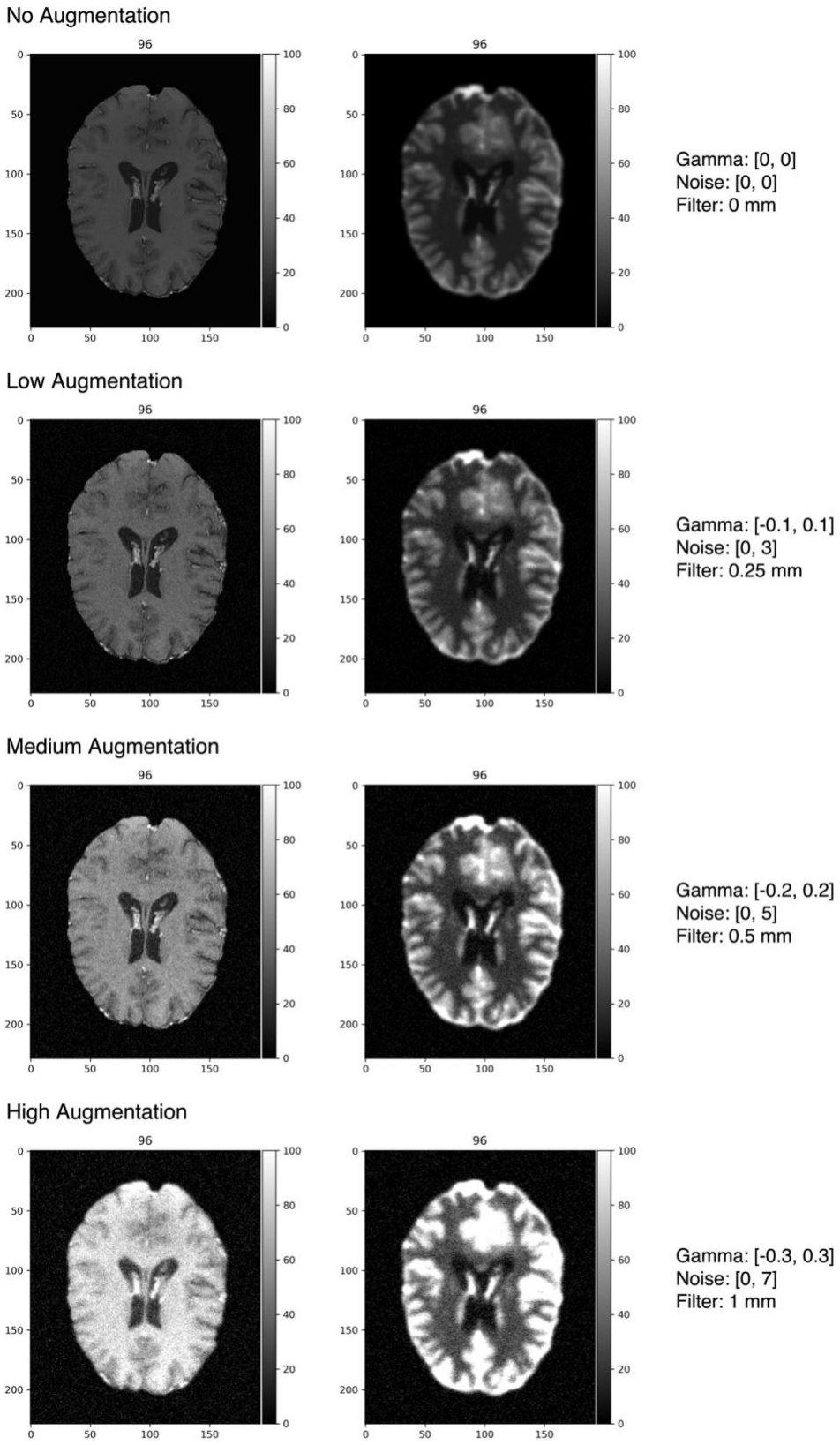

Figure 2.5. Levels of data augmentation (gamma shifting, Gaussian noise, Gaussian filter). For illustrative purposes, each plot is created with the maximum amount of augmentation (i.e., for low augmentation: gamma: 0.3, noise: 3, filter: 0.25mm).

2.3.3. Data Augmentation

Data augmentation is a popular strategy to avoid overfitting small datasets (Shorten and Khoshgoftaar, 2019). Eppenhof and Pluim found that data augmentation reduced the bias of their model (Eppenhof and Pluim, 2018). Therefore, in addition to the random artificial deformations, the images are augmented at the beginning of each epoch. The augmentation includes the addition of random Gaussian noise, random gamma shifting (to change image contrast) and Gaussian smoothing with a kernel with random standard deviation (see Figure 2.5), and are performed using the TorchIO Python library (Pérez-García et al., 2021). The augmentations are applied separately to the MRI and ultrasound images. Finally, the corresponding patches are randomly flipped in (X, Y, Z) directions. Note that the validation data does not undergo this augmentation. Thus, while the training data evidently comes from the same set of images at each epoch, it is 'new' to the model in terms of:

- The artificial deformation size and complexity:
    - Each training image is deformed with a randomly size control point mesh and randomly sized control point displacements.
- The intensity profile:
    - Each training image is augmented by random smoothing, noise and gamma shifting.
- The location of the patch within the image:
    - 100 corresponding patches are randomly selected from each of the training image pairs (the augmented MRI and augmented/deformed simulated ultrasound).
- The orientation:
    - Each augmented patch pair may randomly be flipped in the three cardinal directions.

Thus, during each epoch, 1000 patch pairs were used for training (10 training subjects x 100 patch pairs/subject). A similar approach is followed for the validation data that is used to assess the model *during* training, however, as mentioned, this data is not augmented. The model is assessed on the validation data again *after* training more thoroughly, as described below.

2.4. Evaluating the Model After Training

To validate the model *after* training, it is applied to the full field of view of both validation datasets (as opposed to randomly selecting a limited number of patches, as is done *during* training after each epoch). The validation datasets are randomly deformed following the same

procedure as above. The size and frequency of the deformations are each set to different levels: low, medium and high (see Figure 2.3). Therefore, each of the two validation subjects is deformed nine times. For example, a deformation with low frequency and low size has a [6, 6, 6] control point grid and randomly sized deformations in the range of [15, 15, 15] mm (i.e., 3D displacements sampled from [-7.5, 7.5) mm). Typical histograms of the registration error produced by each level of deformation are provided in Figure 2.3. The model then estimates the registration error for these images. For slices that are not being plotted, the model estimates error with a stride of 5 (i.e., skipping every 5 voxels) to increase computation efficiency. The error is only estimated for voxels within a mask of the brain. Note that the same procedure is performed on the two test datasets when the final model is being evaluated.

The large displacements we apply create large registration errors – that is to say, large 'residual' mis-alignment of MRI and ultrasound volumes. While we believe that nonlinear registration tools used in neurosurgery should result in much smaller errors (in line with the registration errors from brain shift compensation reported in Gerard et al.'s recent review (Gerard et al., 2021)), we wanted to know how robustly we could estimate such large errors, even if they were unlikely in practice.

2.5. Experiments

We performed two experiments to optimize hyperparameters prior to evaluating the model on the test set. First, we examined the optimal number of epochs to train the model. While this is a relatively trivial task, it depicts how the model's performance improves with more training. Second, we evaluated the effect of data augmentation (Gaussian noise, gamma shifting, Gaussian smoothing and random flips). The models from the experiments are evaluated on the validation subjects. Given the insights drawn from these two experiments, the model was trained and evaluated on the test data. The results of the hyperparameter optimization are presented in the next section, followed by the evaluation of the model on the test set.

These experiments are performed using a 'standard model' with the following parameters: Adam optimizer (Kingma and Ba, 2017) with a learning rate of 0.0001, mean squared error loss, 33x33x33 mm patches and 100 samples per subject for each epoch (10 subjects x 100 patches/subject = 1000 patches total). These values were chosen empirically or from the results of Eppenhof and Pluim (Eppenhof and Pluim, 2018). The parameters for the random artificial

deformations were set to [6, 12] for the number of control points and a range of [0, 60] mm for the perturbation of the control points. Linear interpolation was used to resample the transformed images. The model was trained using Keras (Chollet and others, 2015) on top of TensorFlow 2 (Martín Abadi et al., 2015) with an NVIDIA GeForce RTX 3090 graphics processing unit.

We refer to the difference between *estimated registration error* and the *known registration error* as the *regression error*; this is the error in estimating the registration error. We select the Mean Absolute Error (MAE) and Pearson correlation as the validation metrics, as suggested in our review (Bierbrier et al., 2022). The MAE (or Pearson correlation) is reported as the average of the MAEs (or Pearson correlations) for all deformation levels for both validation subjects.

3. Results

3.1. Number of Epochs

The model was trained on a varying number of epochs: 100, 500, 1000, 2000 and 3000. The learning curve for the model trained on 3000 epochs is presented in Figure 3.1.

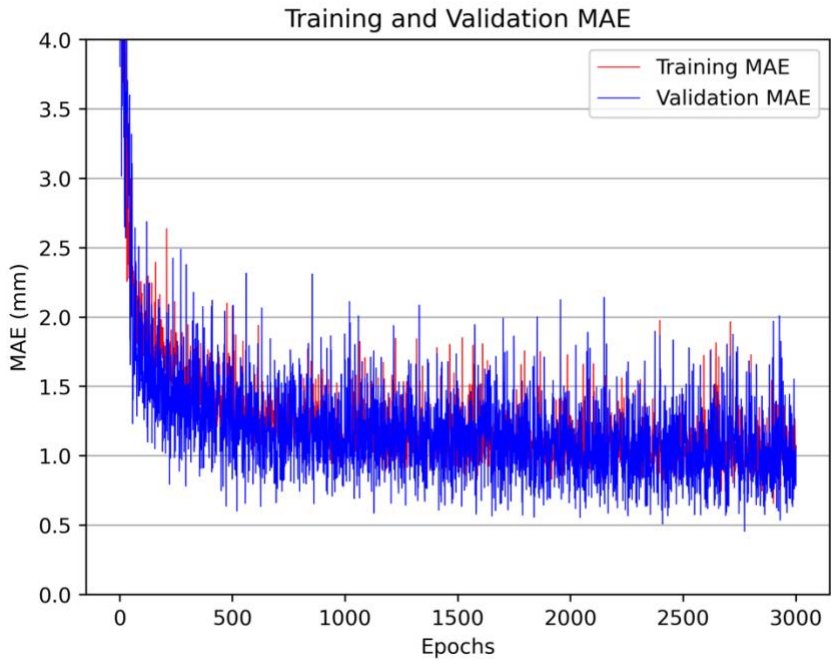

Figure 3.1. The learning curve for the model trained on 3000 epochs. The x-axis represents the epoch number and the y-axis illustrates the MAE in mm.

The experimental results of the models evaluated on the validation subjects for 100, 500, 1000, 2000 and 3000 epochs are presented in Figure 3.2 – Figure 3.4. The model with the lowest average MAE (0.807 mm) was trained for 3000 epochs. This model also had the highest mean Pearson correlation (0.775). The boxplots of Figure 3.2 illustrate that the distribution of absolute errors decreases as training progresses. The same trend is evident in Figure 3.4, where the violin plots illustrate a much larger range of regression errors for the model trained only on 100 epochs compared to that on 3000 epochs. The mean value of these violin plots shows the bias of the model (Eppenhof and Pluim, 2018). Particularly for the larger sized deformations (i.e., [45, 45, 45] mm), the model's bias is reduced as training progresses with more epochs. An example of the model trained on 3000 epochs estimating registration error is given in Figure 3.5.

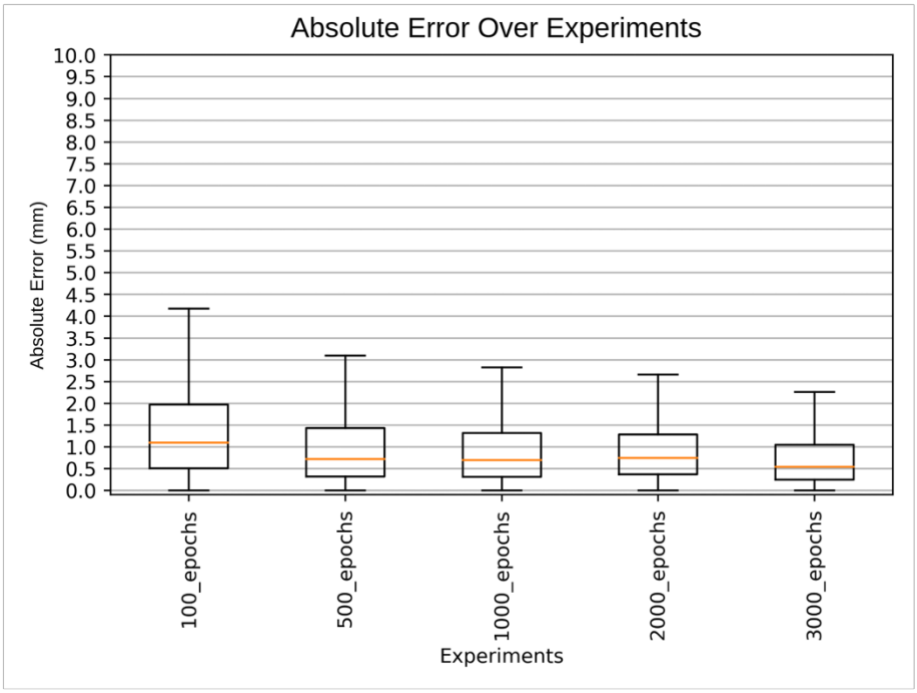

Figure 3.2. The absolute error of regression (|Estimated – True Error|) results from the Epochs experiment. The x-axis depicts the trained models. The y-axis displays the boxplots of the absolute error for all the evaluations performed for each experiment (all deformations for each validation subject). The orange bar represents the median. Note that outliers are not plotted to avoid cluttering the figure.

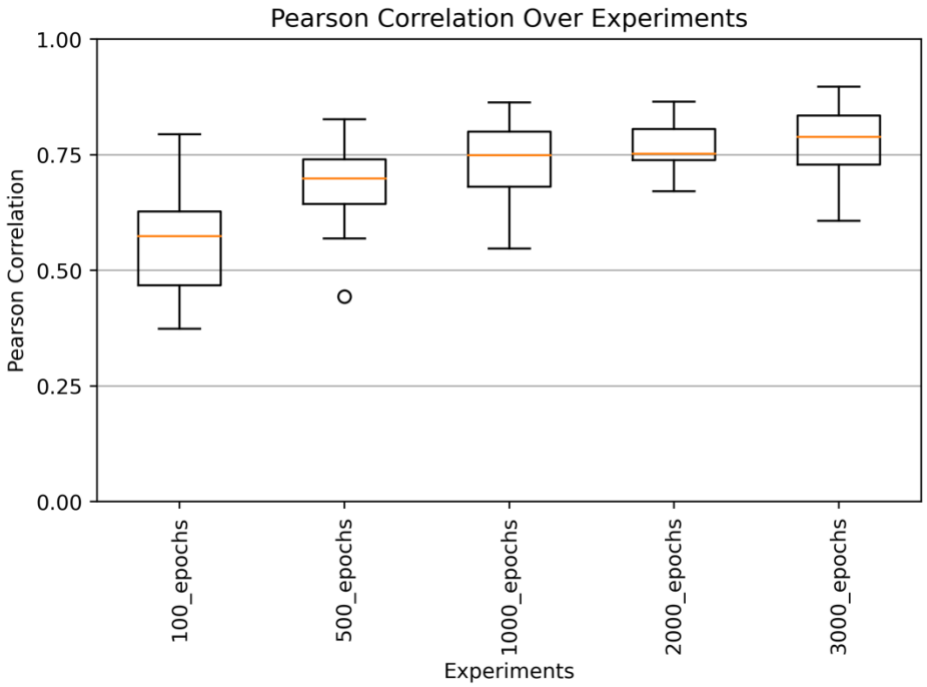

Figure 3.3. The Pearson correlation results for the Epochs experiment. The x-axis depicts the trained models. The y-axis shows boxplots of the Pearson correlation values for each of the evaluations performed (all deformations for each validation subject). The orange bars represent the median.

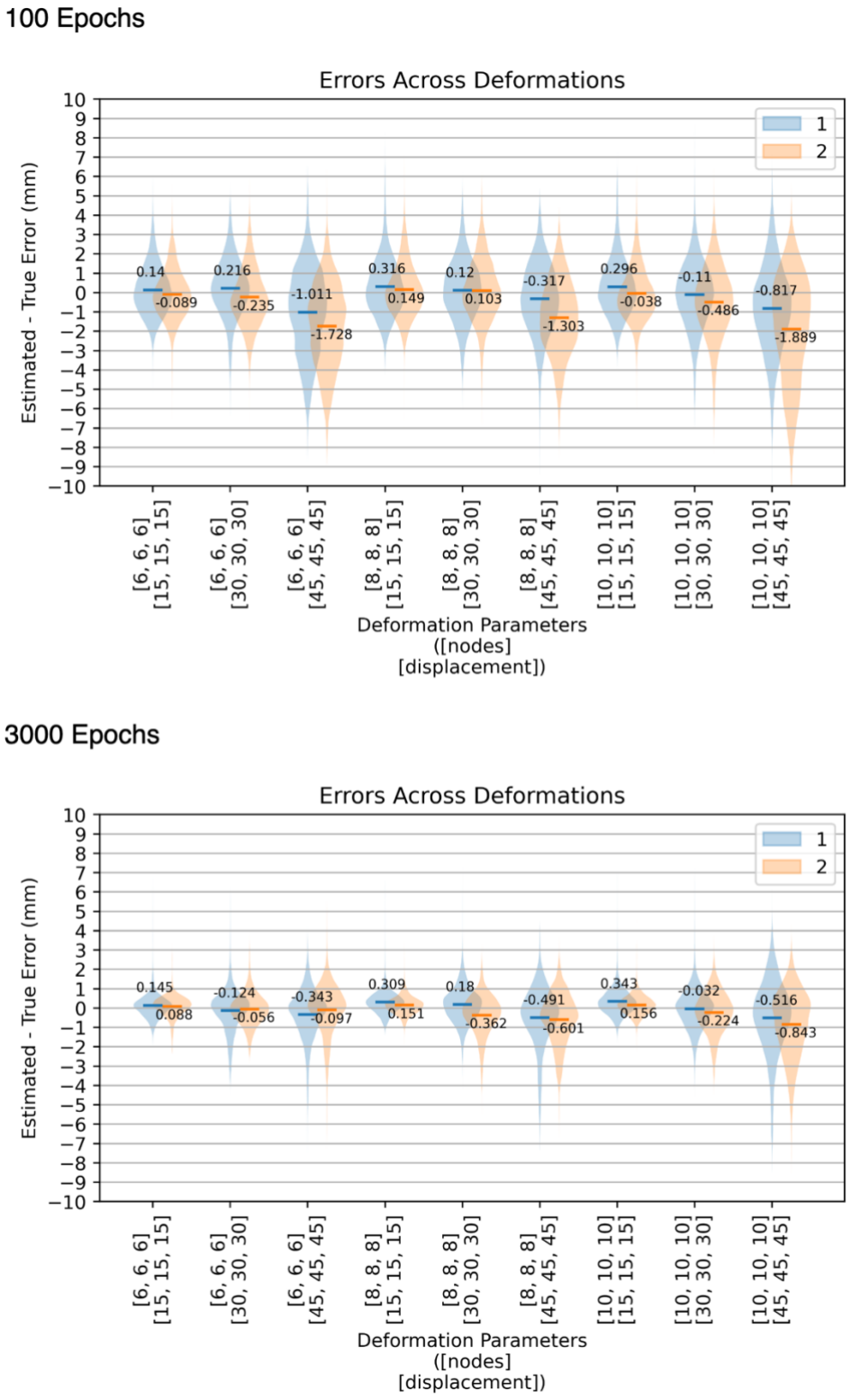

Figure 3.4. Violin plots of the regression error for the two validation subjects across the deformations for models trained on 100 and 3000 epochs. The x-axis represents the levels of the applied deformation. The legend displays the validation subjects (1: subject 11; 2: subject 12). The annotated blue and orange bars display the mean of the violin plots.

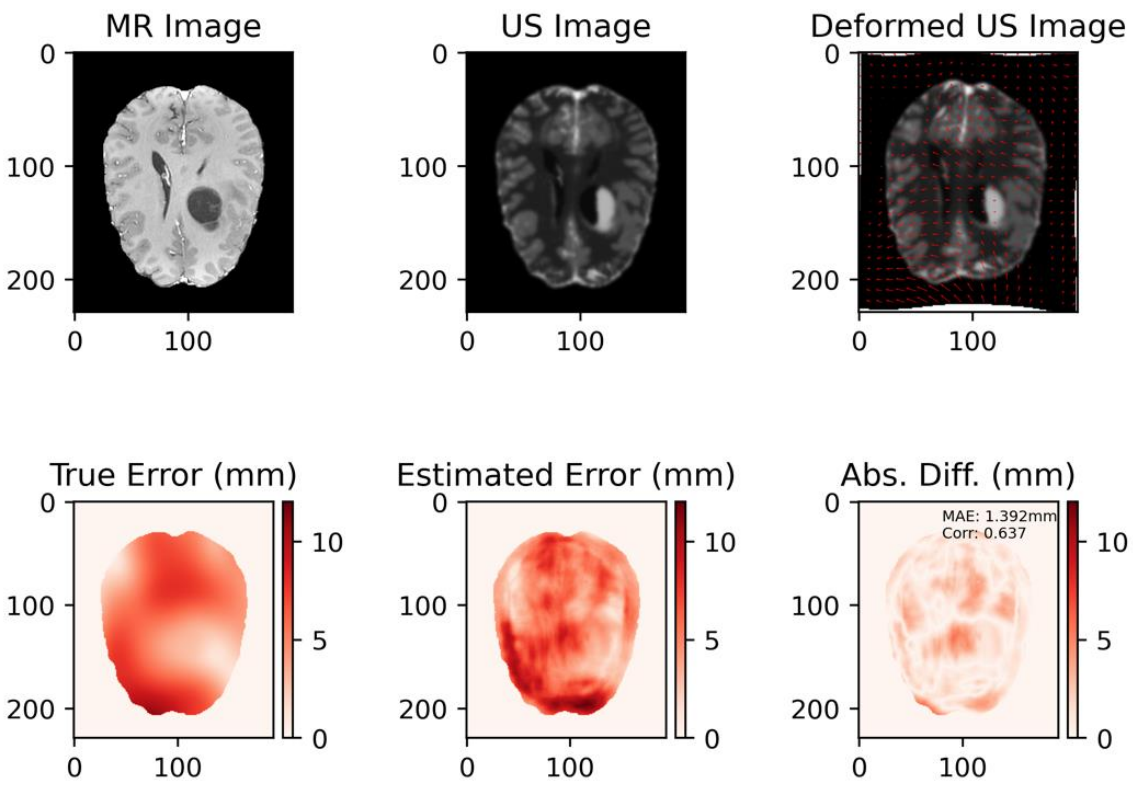

Figure 3.5. Estimated error of slice 100 of validation subject 11 from the model trained on 3000 epochs. The deformation is medium frequency (i.e., [8, 8, 8]) and medium size (i.e., [30, 30, 30] mm). The red arrows on the Deformed US Image (top right) are the displacement vectors of the deformation. Note that correlation and MAE values given in the Absolute Difference image (bottom right) are calculated only for this slice. (Abs. Diff.: Absolute Difference).

These results indicate that training the model on 3000 epochs is optimal. It is possible that training on even more epochs would produce better results (see the future work section). However, given the model's performance does not seem to be improving appreciably (based on the learning curve shown in Figure 3.1) and that it already meets our goals, we stop training after 3000 epochs. The primary reason for this is to avoid overfitting the model on the training data.

3.2. Data Augmentation

The model was trained with different levels of data augmentation (Figure 2.5): none, low, medium and high. It was trained for 1000 epochs for each level of data augmentation.

The experiment results on the two validation subjects for data augmentation are presented in Figure 3.6 – Figure 3.9. The model trained with the low level of data augmentation achieved the lowest average MAE (1.081 mm) followed closely by the model trained with medium data augmentation (1.083 mm). The average Pearson correlation was highest for the model trained with the low level of data augmentation (0.757), followed by no data augmentation (0.740). The MAE and Pearson correlation results plotted over the different levels of deformations confirm these findings (Figure 3.8 – Figure 3.9). The model trained with the low level of data augmentation performs the best most consistently in terms of MAE and Pearson correlation.

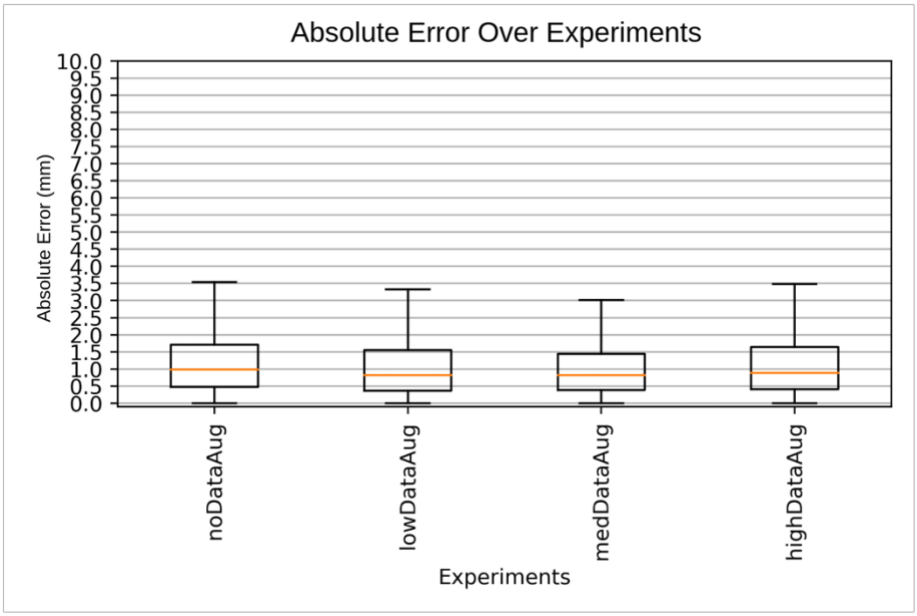

Figure 3.6. The absolute error of regression (\|Estimated – True Error\|) results from the Data Augmentation experiment. The x-axis depicts the trained models. The y-axis shows the boxplots absolute error for all the evaluations performed (over all deformations for both validation subjects). The orange bar represents the median. Note that outliers are not plotted to avoid cluttering the figure.

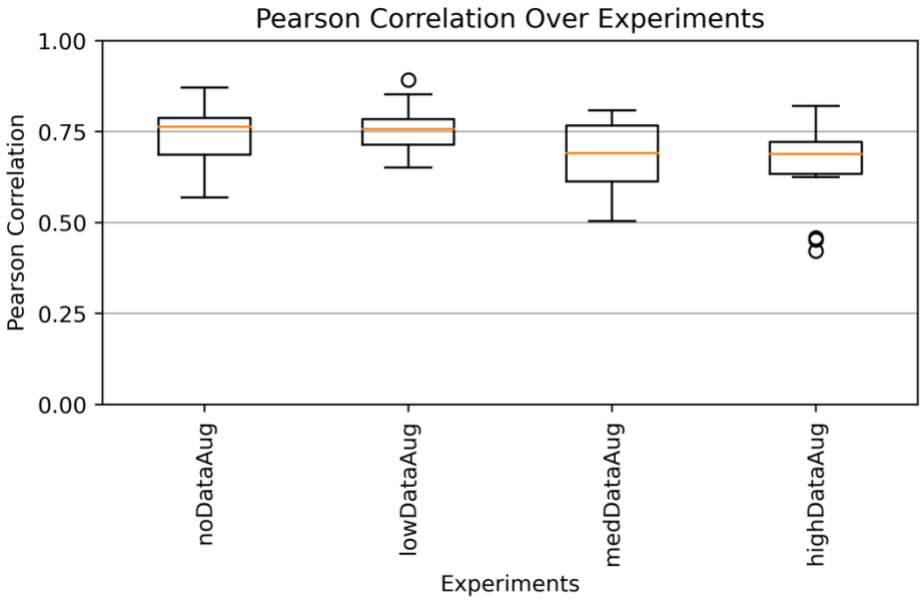

Figure 3.7. The Pearson correlation results for the Data Augmentation experiment. The x-axis depicts the trained models with different levels of data augmentation. The y-axis shows boxplots of the Pearson correlation values for each of the evaluations performed (all deformations for each validation subject). The orange bars represent the median.

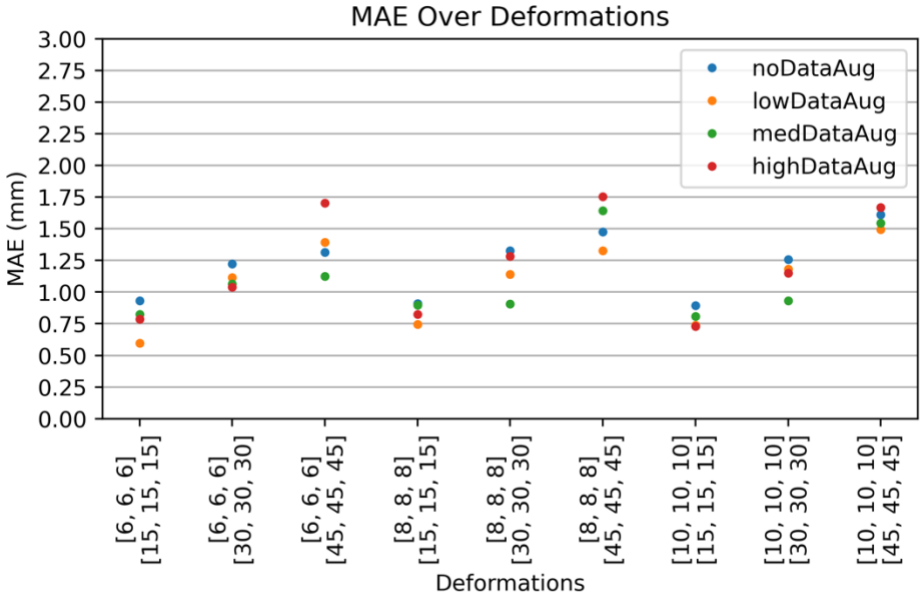

Figure 3.8. The MAE achieved by models trained on different levels of data augmentation across the levels of deformations. The x-axis shows the levels of deformation. The y-axis is the MAE in mm. Note that only the mean values are presented (not the entire distribution) to avoid cluttering the figure.

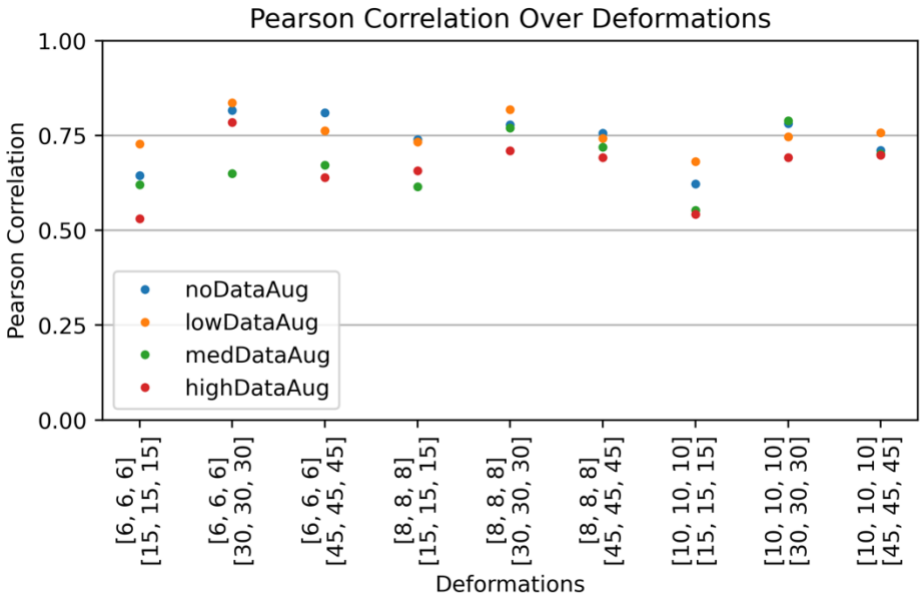

Figure 3.9. The Pearson correlation achieved by models trained on different levels of data augmentation across the levels of deformations. The x-axis shows the levels of deformation. The y-axis illustrates the Pearson correlation (only plotted in the range of [0, 1]).

The results of this experiment suggest that a low level of data augmentation is beneficial. It is possible that models trained on the medium and high levels of data augmentation would achieve the same or even better results if trained for more epochs, since these samples can be more challenging (e.g., more noise or blurring). However, based on the results of this experiment, we used a low level of data augmentation when training the final model.

3.3. Final Model

Drawing on the results of the two experiments, the final model was trained with 3000 epochs and with the low level of data augmentation. This model was evaluated on the independent test set (subjects 13-14) using the same procedure as performed on the validation set. The results for the final model on the test set are presented in Figure 3.10 – Figure 3.12. The model achieves an average MAE of 0.849 mm and an average Pearson correlation of 0.838. This is consistent with the findings on the validation set.

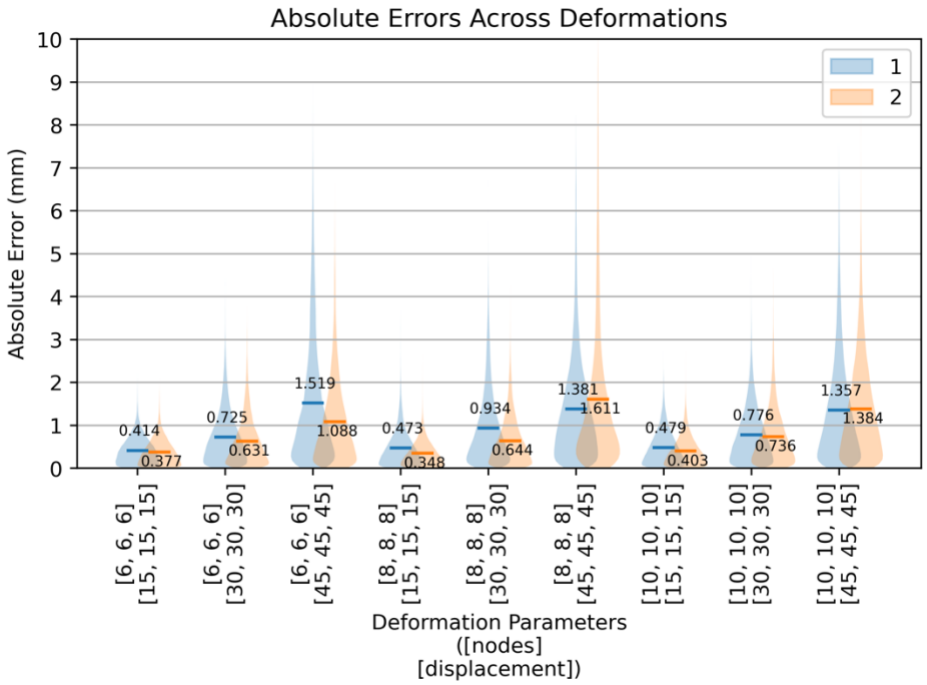

Figure 3.10. Violin plots of the absolute error of regression (|Estimated – True Error|) of the model on the test subjects (1: subject 13; 2: subject 14). The x-axis represents deformation level. The annotated blue and orange bars display the mean of the violin plots.

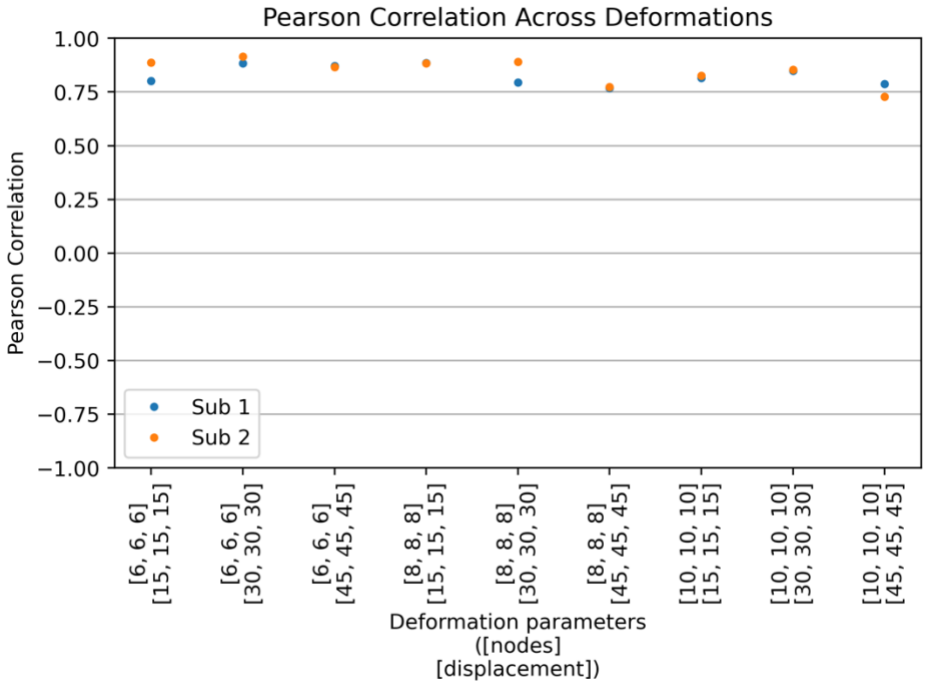

Figure 3.11. The Pearson correlation of the model on the test subjects (Sub 1: subject 13; Sub 2: subject 14). The x-axis represents deformation level.

Figure 3.12. Violin plots of the model's regression error on the test subjects (1: subject 13; 2: subject 14). The x-axis represents the deformation levels. The annotated blue and orange bars display the mean of the violin plots.

There are several discernible trends in the testing data results relating to the level of deformation. Consider the absolute error plot of Figure 3.10. The spread of the model's absolute error distribution increases as the size of the deformations increase from low to medium to high. Correspondingly, the MAE increases with the size of the deformations but remains under 1 mm for the two smallest sets of deformations. There is a similar yet less consistent trend of the MAE increasing as the frequency of the deformations increases. Figure 3.11 illustrates that the Pearson correlation coefficients are generally the highest for the medium size deformations, and lowest for the largest deformations. The correlation coefficients decrease as the frequency of the deformations increases. Finally, considering Figure 3.12, it is apparent that the model consistently underestimates the true registration error. Note that the difference between the two test subjects is relatively small, with the model overall performing better on subject 14.

Figure 3.13 demonstrates qualitatively and quantitatively the model's results for different levels of deformations of the test subjects. Qualitatively, the model performs well at identifying

regions with error. As seen in the Estimated Error and Absolute Difference subplots, the model's error estimate is generally not as smooth as the true error. Notably, the model's performance is not appreciatively affected by the presence of the lesion in Subject 13.

Figure 3.14 shows the model's performance with large deformations ([45, 45, 45] mm) and medium deformation frequency ([8, 8, 8]) for subject 14. The cyan boxes illustrate regions where the model underestimates the true registration error. The underestimation is confirmed by the violin plot. The relatively poor performance on this example results in a high MAE (1.578 mm) and low correlation (0.599), however it is important to note that one would expect smaller residual mis-registration from a robust nonlinear registration algorithm. In contrast, Figure 3.15 displays an instance where the model performed well (with slightly smaller simulated residual mis-registration), resulting in a MAE of 0.671 mm and correlation of 0.853, with no serious over- or underestimation of the true registration error.

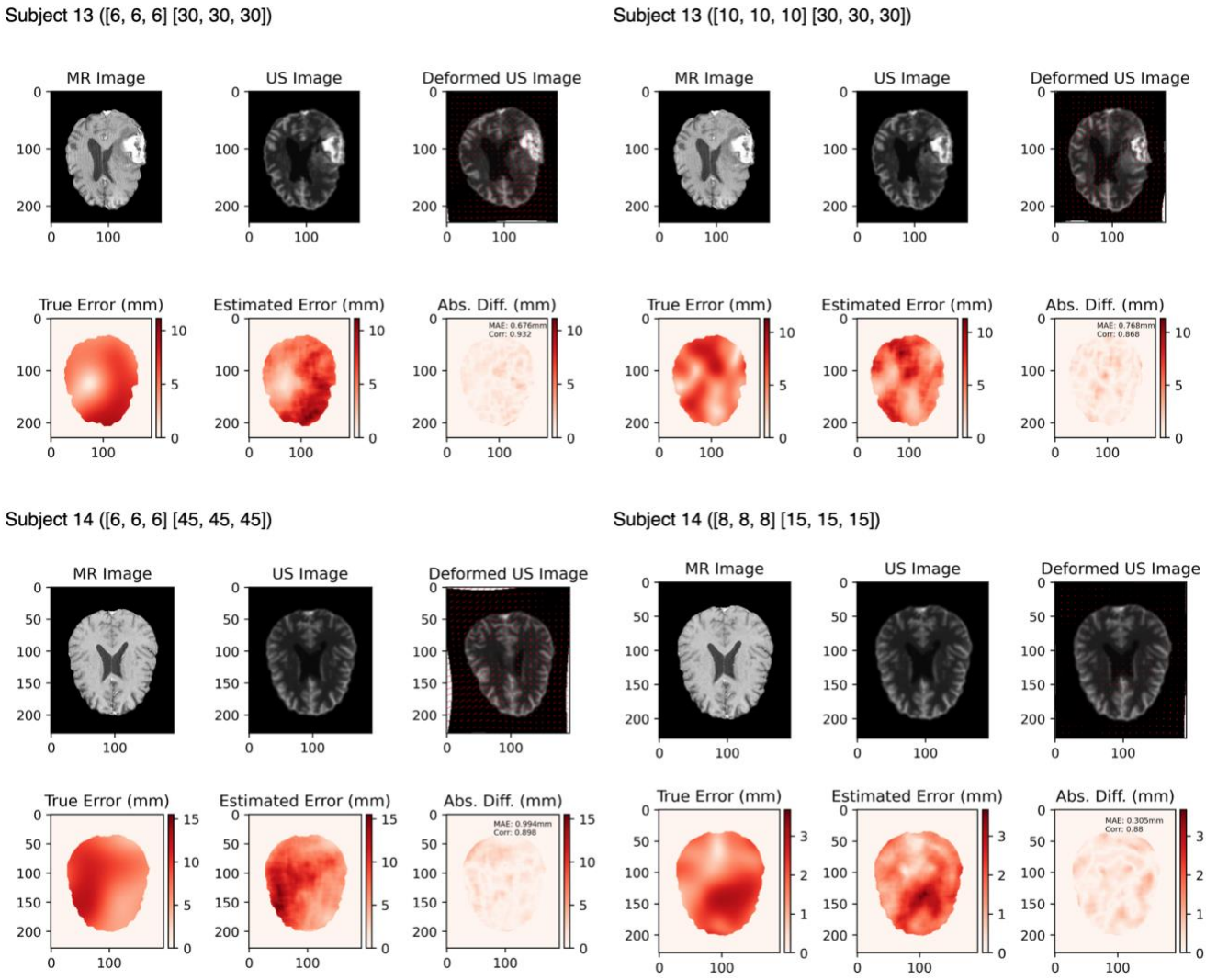

Figure 3.13. Estimated error on the test subjects at different deformation levels. The correlation and MAE values given in the Absolute Difference images are calculated only for presented slice, which in all cases is slice 100. Note that the error colour bar of the True, Estimated and Absolute Difference plots are consistent within each subfigure but vary across the subfigures. (Abs. Diff.: Absolute Difference).

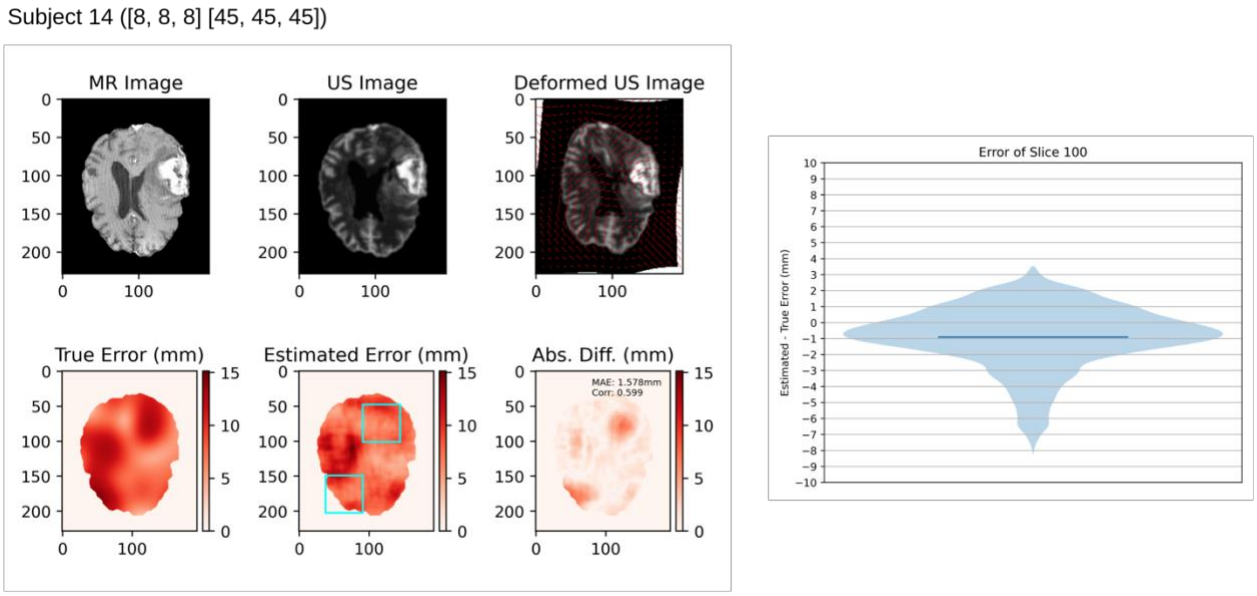

Figure 3.14. Estimated error on subject 14 with the corresponding violin plot displaying the regression error. The cyan boxes highlight regions where the model underestimated the registration error. The correlation and MAE values given in the Absolute Difference image, as well as all data in the violin plot, are calculated only for presented slice (slice 100). The blue bar in the violin plot represents the mean of the distribution. (Abs. Diff.: Absolute Difference).

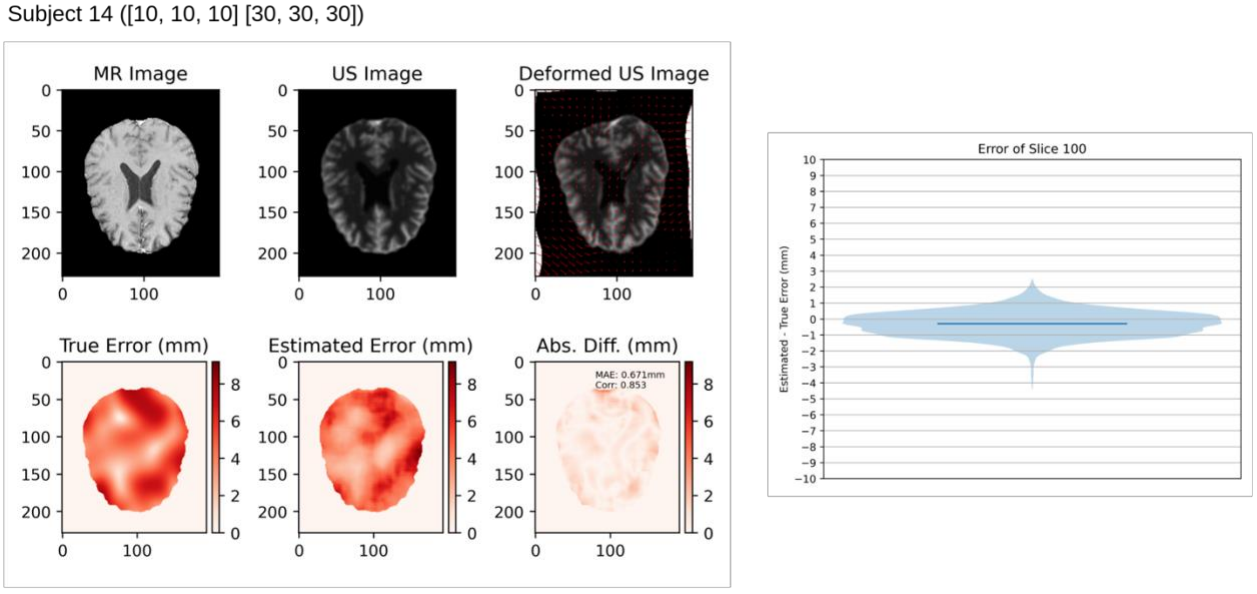

Figure 3.15. Estimated error on subject 14 with the corresponding violin plot displaying the regression error. The correlation and MAE values given in the Absolute Difference image, as well as all data in the violin plot, are calculated only for presented slice (slice 100). The blue bar in the violin plot represents the mean of the distribution. (Abs. Diff.: Absolute Difference).

4. Discussion

Registration error estimating algorithms are becoming more popular (Bierbrier et al., 2022). A successful algorithm would yield immediate benefits in Image-Guided Neurosurgery. We adapted the method of Eppenhof and Pluim (Eppenhof and Pluim, 2018) to estimate errors in MRI-ultrasound registrations in the context of Image-Guided Neurosurgery. We performed experiments to determine the optimal number of epochs and level of data augmentation to use in training. The results of these experiments informed the training of the final model.

The final model exceeded our goals. Evaluated on the test set, it obtained an average MAE of 0.849 mm (goal: mean of < |1| mm) and an average Pearson correlation of 0.838 (goal: > 0.75). Given the success of the model on the test set, it does not seem to be overfitted to the training data. This is likely due to several factors, including limiting the training to 3000 epochs, incorporating data augmentation techniques into the training and using a data generator to provide diverse training samples from the relatively limited training set. In terms of visualization, error maps similar to those in Figure 3.13 can be displayed on the neuronavigation system, perhaps overlayed as a colourwash on the registered image. Once the model is fully trained, it takes approximately 100 second to estimate error within the entire brain mask with an in-slice stride of five. We estimate it will take a similar amount of time for real ultrasound images given their smaller field of view, and thus is acceptable for Image-Guided Surgeries. However, there is room for improvement with more efficient implementation. Evaluating the final model revealed several trends in its performance, as we discuss below.

4.1. Model Trends

4.1.1. Deformation Size

Perhaps not surprisingly, the model performed better on smaller deformations in terms of the MAE (Figure 3.10). One possible explanation for this is the patch size. Eppenhof and Pluim selected the input size as 33x33x33 mm with the assumption that the errors their model would see are in the range of [0, 4] mm (Eppenhof and Pluim, 2018). They found 33x33x33 mm patches yielded the optimal trade-off between a model that has receives enough spatial context and a model that receives unnecessary information (e.g., deformations that do not inform on the error of the central voxel). Our range of deformations was much larger than theirs (up to ~15 mm) to account for the possibly large residual mis-registrations due to ineffective nonlinear

registrations when accounting for brain shift. For large deformations (e.g., > 10 mm), a significant portion of the input patches may not overlap. Using larger-sized patches may be more appropriate to estimate such larger registration errors, however one would expect a successful nonlinear registration algorithm to have much smaller residual alignment errors. Another possible explanation for the better performance on smaller deformations relates to the distribution of training samples; there were many fewer large error samples in the training data. Using a balanced training set could improve the model's ability to estimate large errors.

The correlation coefficients of the medium-sized deformations were the highest, followed by those for the small deformations (Figure 3.11). This may also be due to an unbalanced training set, with the most training samples representing the medium-sized deformations. The large deformations had the lowest correlations, likely due to the issues discussed above.

4.1.2. Deformation Frequency

Increasing the deformation frequency generally resulted in poorer performance in terms of the MAE (Figure 3.10) and (more consistently) the Pearson correlation (Figure 3.11). A higher number of control points allows for higher frequency deformations (i.e., less smooth simulated residual mis-registration). Such deformations seem to pose a greater challenge for the model to detect and estimate. This may relate to the size of the input patches, as Eppenhof and Pluim pointed out (Eppenhof and Pluim, 2018): with higher frequency deformations, the patches are more likely to contain deformations that do not provide information related only to the error of the central voxel of that patch.

4.1.3. Model Bias

Figure 3.12 demonstrates that the model consistently underestimates the true registration error. It is not clear whether this is a result of missing regions with large error, as Figure 3.14 illustrates, or if the model underestimates the error more generally. It is worth noting that in some cases the model overestimates the error; this was observed when evaluating the model on the validation data. It was particularly notable in regions of lesions, where high error would be estimated regardless of the true error. This may be the result of a lack of texture in these regions, which is generally exaggerated by the creation of simulated ultrasound data.

To ameliorate the issue of over- and underestimation of registration error, it is first important to understand (the likely many reasons) why it occurs. If it occurs due to the model's inability to detect large deformations (large registration errors), then the model can be trained with larger registration errors and a larger input patch size. If it is due to a lack contrast (e.g., in a lesion), then such regions can perhaps be preferentially sampled (e.g., based on a probability map) during the training process to ensure the model is sufficiently exposed to these more challenging cases.

4.2. Comparison to Other Methods

Overall, the model's performance meets the design criteria set out at the beginning of this chapter. The method faced significant challenges in estimating error from (a) multimodal registrations, which are generally more challenging that mono-modal registrations, of (b) brain anatomy, which is inherently complex, for (c) patients with tumours, which, as we noted, can present additional issues for algorithms. Beyond this, it estimates error from a relatively large range [0, ~15] mm – almost four times greater than that of Eppenhof and Pluim's model ([0, 4] mm) (Eppenhof and Pluim, 2018). On registration errors in the range similar to that of Eppenhof and Pluim, the model's results are impressive: < 0.5 mm MAE (low sized deformations on the test subjects; see Figure 3.10).

Comparing the performance of different registration error estimating methods is not trivial. In fact, in our review of the field, we call for a standardized analysis approach and datasets for evaluation of competing methods (Bierbrier et al., 2022). The publications in the literature that are most similar to ours are Eppenhof and Pluim (Eppenhof and Pluim, 2018), Sokooti et al. (Sokooti et al., 2021, 2019, 2016), Saygili (Saygili, 2021, 2020, 2018), Lotfi et al. (Lotfi et al., 2013) and Sedghi et al. (Sedghi et al., 2019). Each of these methods used a Machine Learning Framework to estimate registration error, with the exception of Sedghi et al. who integrated a Machine Learning Framework in a registration algorithm that provides uncertainty estimates. Table 4.1 provides the classification of other methods and their results in relation to ours. While the results are not directly comparable, it is clear that our results fall within the range of these other publications.

Author	Approach	Framework	Measure-ment	Reference	Data	Results
(Sokooti et al., 2016)	Parameter Exploration Image-based Transformation Plausibility	ML (RF; Landmarks + surrounding region)	Error	Landmarks	1. SPREAD: follow-up chest CT of 21 emphysema patients (~100 landmarks per case, chosen semi-automatically (Staring et al., 2014)) (Stolk et al., 2007)	Mean absolute error: 0.72 ± 0.96 mm
(Sokooti et al., 2019)	Parameter Exploration Image-based Transformation Plausibility	ML (RF; Landmarks + surrounding region)	Error	Landmarks	1. SPREAD 2. DIR-Lab-4DCT: 10 thoracic CT, half from patients with thoracic malignancies (>300 landmarks per image pair) (E. Castillo et al., 2009; R. Castillo et al., 2009) 3. DIR-Lab-COPDgene: 10 thoracic CT of patients with severe breathing disorders (>300 landmarks per case) (Castillo et al., 2013)	Mean absolute error: 1.07 ± 1.86 mm (intra-database; SPREAD data), 1.76 ± 2.59mm (inter-database)
(Eppenhof and Pluim, 2018)	Image-based	ML (CNN; Artificial deformations)	Error	Landmarks Artificial deformations	1. DIR-Lab-4DCT 2. DIR-Lab-COPDgene 3. POPI-model: 1 thorax 4DCT dataset (40 landmarks) (Vandemeulebroucke et al., 2007) 4. CREATIS: 6 thorax 4DCT of cancer patients (100 landmarks per case) (Vandemeulebroucke et al., 2011)	Root mean square difference: 0.51 mm (Artificial deformations), 0.66 mm (Landmarks)
(Saygili, 2018)	Parameter Exploration	ML (RF; Landmarks + surrounding region)	Error	Landmarks	1. DIR-Lab-4DCT	Mean absolute error: 1.64 ± 1.81 mm
(Saygili, 2020)	Parameter Exploration	ML (RF; Landmarks + surrounding region)	Error	Landmarks	1. DIR-Lab-4DCT 2. CREATIS	Mean absolute error: 2.00 mm (Three Orthogonal Planes approach with RF) $R^2 = 0.74$
(Saygili, 2021)	Parameter Exploration	ML (RF; Landmarks + surrounding region)	Error	Landmarks Artificial deformations	1. DIR-Lab-4DCT 2. CREATIS 3. HAMMERS: 30 healthy brain MRI (Hammers et al., 2003)	R^2 correlations: 0.63756, 0.58005 and 0.68825 (POPI data (x,y,z)) R^2 correlations: 0.65163, 0.56063, 0.77355 (DIRLab data (x,y,z)) Mean absolute error: 1.05 ± 1.28, 0.81 ± 0.87, 1.71 ± 2.91, 1.75 ± 2.47 mm (x, y, z, magnitude)
(Sokooti et al., 2021)	Image-based	ML (Encoder-Decoder; Artificial deformations)	Error	Landmarks	1. DIR-Lab-COPDgene 2. DIR-Lab-4DCT 3. SPREAD	Classification accuracy: 87.1% Average F1 score: 66.4%
(Lotfi et al., 2013)	Parameter Exploration Image-based Transformation Plausibility	Directly to Measurement ML (RF; Artificial deformations)	Uncertainty Error	Artificial deformations	1. Synthetic texture image 2. BRATS 2012: 6 brain tumour MRI ("MICCAI BRATS 2012," n.d.) 3. LBPA40: 40 brain MRI (Shattuck et al., 2008)	Correlation coefficient: 0.730 (uncertainty; synthetic data), 0.215 (est. error; without uncertainty as a feature), 0.542 (est. error; with uncertainty as a feature)
(Sedghi et al., 2019)	Image-based Parameter Exploration	Directly to Measurement ML (CNN; Artificial deformations)	Uncertainty	Artificial deformations	1. IXI: T1 and T2 brain MRI ("IXI Dataset – Brain Development," n.d.)	Correlation coefficient: 0.94
Our implementation	Image-based	ML (CNN; Artificial deformations)	Error	Artificial deformations	1. BITE database: 14 brain tumour MRI and ultrasound (Mercier et al., 2012a)	Mean absolute error: 0.849 mm Pearson correlation: 0.838

Table 4.1. Classification (according to the taxonomy we proposed (Bierbrier et al., 2022)) and results of similar methods. Taken and partially modified from (Bierbrier et al., 2022).

4.3. Future Work

Our progress marks a positive step towards the use of registration error estimating algorithms in the context of Image-Guided Neurosurgery. There are several areas of future research. They are grouped into three sections: model, evaluation and future experiments. Ultimately, addressing these areas provides the basis for our future work.

4.3.1. Model

There are numerous design considerations when crafting a machine learning model. Many of these were determined empirically (e.g., the learning rate) or based on the results of others (e.g., the model architecture and patch size), however a more thorough analysis of these parameters would likely prove useful. Such parameters include the choice of learning rate, optimizer, the patch size, number of samples per subject in each epoch, number of epochs to train and type and amount of regularization. In fact, the architecture itself can likely be improved. We implemented the model proposed by Eppenhof and Pluim (Eppenhof and Pluim, 2018) with minor improvements (such as the addition of L2 regularization). This model performed well for lung CT registrations, as presented by Eppenhof and Pluim, and adequately for brain MRI-ultrasound registrations, as presented above. It is possible that a more complex architecture would produce better results. For example, modifying the encoder-decoder model Sokooti et al. proposed to regress registration error estimates (as opposed to classifying the registration error into different ranges) (Sokooti et al., 2021). Such an architecture may reduce how noisy the error estimates can be. There are also several deep learning-based image registration algorithms that can be used as inspiration for registration error estimating algorithms (e.g., (Dalca et al., 2019; Yang et al., 2017)).

4.3.2. Evaluation

There are several areas where the model evaluation can be made more robust. First, the fully trained models are validated by being deployed on different levels of random artificial deformations on the validation subjects. These deformations are random for each evaluation. Ideally, they would be the same for across experiments. Additionally, the process of randomly deforming the validation data could be repeated several times for each model so that the model's average behaviour could be discerned, rather than its behaviour on a specific random

deformation. It is possible, for example, that the random deformations for one model are more complex (e.g., much larger) for one model than for another. As a result, the model that is evaluated on the more complex data performs (and is judged to be) worse.

Second, the validation process can be made more robust. Selecting only two subjects as the 'validation subjects' is not optimal. Performing k-fold cross validation, in which the model is trained with a different training set (and different validation set) each iteration, is more robust.

Finally, the model should be assessed with 'model mismatch,' which it currently is not. We discuss the implications of model mismatch in our review (Bierbrier et al., 2022). In our case, model mismatch occurs when the transformation model used is different for the training and the validation data. We used B-spline deformations for both the training and the validation data synthesis. Ideally the model should be assessed with model mismatch, i.e., with validation data created by a different type of artificial deformation (e.g., a finite element method or thin plate splines). Additionally, when implementing these future artificial deformations, as well as the B-spline deformations from above, more care will be taken to ensure the deformations are realistic and, in particular, diffeomorphic.

4.3.3. Future Experiments

The final avenue of future work is a more comprehensive validation. First, this could include more data. Given that the method of generating simulated ultrasound images only requires an MRI image, data from any database that contains MRI can be used (although the process would need to be modified if the MRI images are not gadolinium enhanced). Note, too, that the presence of a lesion (e.g., a tumour) can impact the performance of a model. Additional data should be selected with this in mind.

While the simulated ultrasound is designed to resemble real ultrasound acquisitions, it is not perfect. The model's performance on real ultrasound data – the ultimate goal of this work – is not guaranteed to match its performance on simulated ultrasound. Therefore, a more comprehensive validation involves evaluating the model on real ultrasound data with real registrations. Given that creating a dense correspondence map for real MRI-ultrasound data is not feasible, manually annotated landmarks are required. While manually annotated landmarks are not perfect (see a discussion in our review (Bierbrier et al., 2022)), they are the best case for real data. Fortunately, the BITE database, from which our data comes, contains a set of such landmarks.

It is possible that the model will not perform as well on real ultrasound data given the relative simplicity of the simulated ultrasound. For example, a major difference between the simulated and real ultrasound images is the field of view. While our simulated ultrasound is created for the entire brain mask, real ultrasound imaging captures a much smaller field of view. This difference alone is not a major challenge for the trained model given that it operates on a patch-wise basis – it is only concerned with the corresponding 33x33x33 mm patches it receives as input, not the field of view of the entire ultrasound and MRI images. Potential difficulties arise when there is an edge in the ultrasound patch caused by the limited field of view of the real ultrasound, since such edges will not be present in the MRI patch. Making the simulated ultrasound (and hence, training data) more realistic by extracting limited fields of view from the whole-brain simulated ultrasound may alleviate this potential problem. There are several other possible steps to create more realistic simulated ultrasound images to bridge the distribution shift between the training data (simulated ultrasound) and the data the model will see in practice (real ultrasound). These include adding speckle to the images (a type of noise present in ultrasound images), adding depth effects (ultrasound decays proportionally to depth) and creating more realistic ultrasound/tissue properties (e.g., high contrast at certain tissue junctions). While these improvements will undoubtedly make the simulated ultrasound more realistic, one wonders how realistic the simulations *need* to be for the purpose of training a model to estimate registration error.

5. Conclusion

A successful registration error estimating algorithm would benefit Image-Guided Neurosurgery by enabling surgeons to trust (linear or nonlinear) registrations performed on neuronavigation systems. We implemented an established method for MRI-ultrasound registrations. Our results provide a strong basis for future research and, eventually, implementation on neuronavigation systems like IBIS, the neuronavigation system developed by our lab (Drouin et al., 2017).

6. Acknowledgements

This study was funded by grants from the Canadian Institutes of Health Research and from the Natural Sciences and Engineering Research Council of Canada. JB acknowledges funding from the Canada First Research Excellence Fund and Fonds de recherche du Québec, awarded to the Healthy Brains, Healthy Lives initiative at McGill University, and the Department of Biomedical Engineering at McGill University.

7. References

Amidi, A., Amidi, S., 2018. A detailed example of data generators with Keras [WWW Document]. URL https://stanford.edu/~shervine/blog/keras-how-to-generate-data-on-the-fly (accessed 4.20.22).

Arbel, T., Morandi, X., Comeau, R.M., Collins, D.L., 2004. Automatic non-linear MRI-ultrasound registration for the correction of intra-operative brain deformations. Computer Aided Surgery 9, 123–136. https://doi.org/10.3109/10929080500079248

Archip, N., Clatz, O., Whalen, S., Kacher, D., Fedorov, A., Kot, A., Chrisochoides, N., Jolesz, F., Golby, A., Black, P.M., Warfield, S.K., 2007. Non-rigid alignment of pre-operative MRI, fMRI, and DT-MRI with intra-operative MRI for enhanced visualization and navigation in image-guided neurosurgery. NeuroImage 35, 609–624. https://doi.org/10.1016/j.neuroimage.2006.11.060

Beare, R., Lowekamp, B., Yaniv, Z., 2018. Image Segmentation, Registration and Characterization in R with SimpleITK. Journal of Statistical Software 86, 1–35. https://doi.org/10.18637/jss.v086.i08

Bierbrier, J., Gueziri, H.-E., Collins, D.L., 2022. Estimating Medical Image Registration Error and Confidence: A Taxonomy and Systematic Review. Medical Image Analysis.

Castillo, E., Castillo, R., Martinez, J., Shenoy, M., Guerrero, T., 2009. Four-dimensional deformable image registration using trajectory modeling. Phys. Med. Biol. 55, 305–327. https://doi.org/10.1088/0031-9155/55/1/018

Castillo, R., Castillo, E., Fuentes, D., Ahmad, M., Wood, A.M., Ludwig, M.S., Guerrero, T., 2013. A reference dataset for deformable image registration spatial accuracy evaluation using the COPDgene study archive. Phys. Med. Biol. 58, 2861–2877. https://doi.org/10.1088/0031-9155/58/9/2861

Castillo, R., Castillo, E., Guerra, R., Johnson, V.E., McPhail, T., Garg, A.K., Guerrero, T., 2009. A framework for evaluation of deformable image registration spatial accuracy using large landmark point sets. Phys. Med. Biol. 54, 1849–1870. https://doi.org/10.1088/0031-9155/54/7/001

Chollet, F., others, 2015. Keras.

Cocosco, C.A., Zijdenbos, A.P., Evans, A.C., 2003. A fully automatic and robust brain MRI tissue classification method. Med Image Anal 7, 513–527. https://doi.org/10.1016/s1361-8415(03)00037-9

Collins, D.L., Evans, A.C., 1997. Animal: Validation and Applications of Nonlinear Registration-Based Segmentation. Int. J. Patt. Recogn. Artif. Intell. 11, 1271–1294. https://doi.org/10.1142/S0218001497000597

Collins, D.L., Neelin, P., Peters, T.M., Evans, A.C., 1994. Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space. J Comput Assist Tomogr 18, 192–205.

Collins, D.L., Zijdenbos, A.P., Baaré, W.F.C., Evans, A.C., 1999. ANIMAL+INSECT: Improved Cortical Structure Segmentation, in: Kuba, A., Šáamal, M., Todd-Pokropek, A. (Eds.), Information Processing in Medical Imaging, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 210–223. https://doi.org/10.1007/3-540-48714-X_16

Dalca, A.V., Balakrishnan, G., Guttag, J., Sabuncu, M.R., 2019. Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. Medical Image Analysis 57, 226–236. https://doi.org/10.1016/j.media.2019.07.006

De Nigris, D., Collins, D.L., Arbel, T., 2012. Multi-Modal Image Registration Based on Gradient Orientations of Minimal Uncertainty. IEEE Transactions on Medical Imaging 31, 2343–2354. https://doi.org/10.1109/TMI.2012.2218116

Dickhaus, H., Ganser, K.A., Staubert, A., Bonsanto, M.M., Wirtz, C.R., Tronnier, V.M., Kunze, S., 1997. Quantification of brain shift effects by MR-imaging, in: Proceedings of the 19th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. "Magnificent Milestones and Emerging Opportunities in Medical Engineering" (Cat. No.97CH36136). Presented at the Proceedings of the 19th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. "Magnificent Milestones and Emerging Opportunities in Medical Engineering" (Cat. No.97CH36136), pp. 491–494 vol.2. https://doi.org/10.1109/IEMBS.1997.757652

Drouin, S., Kochanowska, A., Kersten-Oertel, M., Gerard, I.J., Zelmann, R., De Nigris, D., Bériault, S., Arbel, T., Sirhan, D., Sadikot, A.F., Hall, J.A., Sinclair, D.S., Petrecca, K., DelMaestro, R.F., Collins, D.L., 2017. IBIS: an OR ready open-source platform for

image-guided neurosurgery. Int J CARS 12, 363–378. https://doi.org/10.1007/s11548-016-1478-0

Eppenhof, K.A.J., Pluim, J.P.W., 2018. Error estimation of deformable image registration of pulmonary CT scans using convolutional neural networks. Journal of Medical Imaging 5, 1–1. https://doi.org/10.1117/1.jmi.5.2.024003

Eskildsen, S.F., Coupé, P., Fonov, V., Manjón, J.V., Leung, K.K., Guizard, N., Wassef, S.N., Østergaard, L.R., Collins, D.L., 2012. BEaST: Brain extraction based on nonlocal segmentation technique. NeuroImage 59, 2362–2373. https://doi.org/10.1016/j.neuroimage.2011.09.012

Fedorov, A., Risholm, P., Wells, W.M., 2014. Deformable Registration for IGT, in: Jolesz, F.A. (Ed.), Intraoperative Imaging and Image-Guided Therapy. Springer New York, New York, NY, pp. 211–223. https://doi.org/10.1007/978-1-4614-7657-3_14

Galloway, R.L., 2015. Introduction and Historical Perspectives on Image-Guided Surgery, in: Image-Guided Neurosurgery. Elsevier, pp. 1–22. https://doi.org/10.1016/B978-0-12-800870-6.00001-7

Galloway, R.L., 2001. The Process and Development of Image-Guided Procedures. Annual Review of Biomedical Engineering 3, 83–108. https://doi.org/10.1146/annurev.bioeng.3.1.83

Gerard, I.J., Kersten-Oertel, M., Hall, J.A., Sirhan, D., Collins, D.L., 2021. Brain Shift in Neuronavigation of Brain Tumors: An Updated Review of Intra-Operative Ultrasound Applications. Front. Oncol. 10, 618837. https://doi.org/10.3389/fonc.2020.618837

Gerard, I.J., Kersten-Oertel, M., Petrecca, K., Sirhan, D., Hall, J.A., Collins, D.L., 2017. Brain shift in neuronavigation of brain tumors: A review. Medical Image Analysis 35, 403–420. https://doi.org/10.1016/j.media.2016.08.007

Grimson, W.E., Kikinis, R., Jolesz, F.A., Black, P.M., 1999. Image-guided surgery. Sci Am 280, 62–69. https://doi.org/10.1038/scientificamerican0699-62

Hammers, A., Allom, R., Koepp, M.J., Free, S.L., Myers, R., Lemieux, L., Mitchell, T.N., Brooks, D.J., Duncan, J.S., 2003. Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe. Hum Brain Mapp 19, 224–247. https://doi.org/10.1002/hbm.10123

Hartkens, T., Hill, D.L.G., Castellano-Smith, A.D., Hawkes, D.J., Maurer, C.R., Martin, A.J., Hall, W.A., Liu, H., Truwit, C.L., 2003. Measurement and analysis of brain deformation during neurosurgery. IEEE Transactions on Medical Imaging 22, 82–92. https://doi.org/10.1109/TMI.2002.806596

Hastreiter, P., Rezk-Salama, C., Soza, G., Bauer, M., Greiner, G., Fahlbusch, R., Ganslandt, O., Nimsky, C., 2004. Strategies for brain shift evaluation. Medical Image Analysis 8, 447–464. https://doi.org/10.1016/j.media.2004.02.001

IXI Dataset – Brain Development, n.d. URL http://brain-development.org/ixi-dataset/ (accessed 10.7.21).

Kelly, P.J., Kall, B.A., Goerss, S., Earnest, F., 1986. Computer-assisted stereotaxic laser resection of intra-axial brain neoplasms. Journal of Neurosurgery 64, 427–439. https://doi.org/10.3171/jns.1986.64.3.0427

Kingma, D.P., Ba, J., 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs].

Lacroix, M., Abi-Said, D., Fourney, D.R., Gokaslan, Z.L., Shi, W., DeMonte, F., Lang, F.F., McCutcheon, I.E., Hassenbusch, S.J., Holland, E., Hess, K., Michael, C., Miller, D., Sawaya, R., 2001. A multivariate analysis of 416 patients with glioblastoma multiforme: prognosis, extent of resection, and survival. Journal of Neurosurgery 95, 190–198. https://doi.org/10.3171/jns.2001.95.2.0190

Lotfi, T., Tang, L., Andrews, S., Hamarneh, G., 2013. Improving Probabilistic Image Registration via Reinforcement Learning and Uncertainty Evaluation, in: Wu, G., Zhang, D., Shen, D., Yan, P., Suzuki, K., Wang, F. (Eds.), Machine Learning in Medical Imaging, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 187–194. https://doi.org/10.1007/978-3-319-02267-3_24

Lowekamp, B., Chen, D., Ibanez, L., Blezek, D., 2013. The Design of SimpleITK. Frontiers in Neuroinformatics 7.

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Jia, Y., Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay

Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, Xiaoqiang Zheng, 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems.

Mercier, L., Del Maestro, R.F., Petrecca, K., Araujo, D., Haegelen, C., Collins, D.L., 2012a. Online database of clinical MR and ultrasound images of brain tumors. Medical Physics 39, 3253–3261. https://doi.org/10.1118/1.4709600

Mercier, L., Fonov, V., Haegelen, C., Del Maestro, R.F., Petrecca, K., Collins, D.L., 2012b. Comparing two approaches to rigid registration of three-dimensional ultrasound and magnetic resonance images for neurosurgery. Int J CARS 7, 125–136. https://doi.org/10.1007/s11548-011-0620-2

MICCAI BRATS 2012 [WWW Document], n.d. URL http://www2.imm.dtu.dk/projects/BRATS2012/index.html (accessed 10.7.21).

Miner, R.C., 2017. Image-Guided Neurosurgery. Journal of Medical Imaging and Radiation Sciences 48, 328–335. https://doi.org/10.1016/j.jmir.2017.06.005

Muenzing, S.E.A., van Ginneken, B., Murphy, K., Pluim, J.P.W., 2012. Supervised quality assessment of medical image registration: Application to intra-patient CT lung registration. Medical Image Analysis 16, 1521–1531. https://doi.org/10.1016/j.media.2012.06.010

Nabavi, A., Black, P.M., Gering, D.T., Westin, C.F., Mehta, V., Pergolizzi, R.S., Ferrant, M., Warfield, S.K., Hata, N., Schwartz, R.B., Wells, W.M., Kikinis, R., Jolesz, F.A., 2001. Serial intraoperative magnetic resonance imaging of brain shift. Neurosurgery 48, 787–797; discussion 797-798. https://doi.org/10.1097/00006123-200104000-00019

Pérez-García, F., 2020. B-spline deformation [WWW Document]. Gist. URL https://gist.github.com/fepegar/b723d15de620cd2a3a4dbd71e491b59d (accessed 4.21.22).

Pérez-García, F., Sparks, R., Ourselin, S., 2021. TorchIO: A Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. Computer Methods and Programs in Biomedicine 208, 106236. https://doi.org/10.1016/j.cmpb.2021.106236

Rivaz, H., Chen, S.J.-S., Collins, D.L., 2015. Automatic Deformable MR-Ultrasound Registration for Image-Guided Neurosurgery. IEEE Transactions on Medical Imaging 34, 366–380. https://doi.org/10.1109/TMI.2014.2354352

Rivaz, H., Collins, D.L., 2015. Near real-time robust non-rigid registration of volumetric ultrasound images for neurosurgery. Ultrasound Med Biol 41, 574–587. https://doi.org/10.1016/j.ultrasmedbio.2014.08.013

Rivaz, H., Karimaghaloo, Z., Fonov, V.S., Collins, D.L., 2014. Nonrigid registration of ultrasound and MRI using contextual conditioned mutual information. IEEE Trans Med Imaging 33, 708–725. https://doi.org/10.1109/TMI.2013.2294630

Rohlfing, T., 2012. Image Similarity and Tissue Overlaps as Surrogates for Image Registration Accuracy: Widely Used but Unreliable. IEEE Trans. Med. Imaging 31, 153–163. https://doi.org/10.1109/TMI.2011.2163944

Rueckert, D., Schnabel, J.A., 2011. Medical Image Registration, in: Deserno, T.M. (Ed.), Biomedical Image Processing, Biological and Medical Physics, Biomedical Engineering. Springer Berlin Heidelberg, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-15816-2

Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L.G., Leach, M.O., Hawkes, D.J., 1999. Nonrigid registration using free-form deformations: application to breast MR images. IEEE Transactions on Medical Imaging 18, 712–721. https://doi.org/10.1109/42.796284

Sastry, R., Bi, W.L., Pieper, S., Frisken, S., Kapur, T., Wells, W., Golby, A.J., 2017. Applications of Ultrasound in the Resection of Brain Tumors: Ultrasound in Brain Tumor Resection. J Neuroimaging 27, 5–15. https://doi.org/10.1111/jon.12382

Saygili, G., 2021. Predicting medical image registration error through independent directions. SIViP 15, 223–230. https://doi.org/10.1007/s11760-020-01784-3

Saygili, G., 2020. Predicting medical image registration error with block-matching using three orthogonal planes approach. Signal, Image and Video Processing 14, 1099–1106. https://doi.org/10.1007/s11760-020-01650-2

Saygili, G., 2018. Local-search based prediction of medical image registration error, in: Nishikawa, R.M., Samuelson, F.W. (Eds.), Medical Imaging 2018: Image Perception, Observer Performance, and Technology Assessment. Presented at the Image Perception,

Observer Performance, and Technology Assessment, SPIE, Houston, United States, p. 49. https://doi.org/10.1117/12.2293740

Sedghi, A., Kapur, T., Luo, J., Mousavi, P., Wells, W.M., 2019. Probabilistic Image Registration via Deep Multi-class Classification: Characterizing Uncertainty, in: Greenspan, H., Tanno, R., Erdt, M., Arbel, T., Baumgartner, C., Dalca, A., Sudre, C.H., Wells, W.M., Drechsler, K., Linguraru, M.G., Oyarzun Laura, C., Shekhar, R., Wesarg, S., González Ballester, M.Á. (Eds.), Uncertainty for Safe Utilization of Machine Learning in Medical Imaging and Clinical Image-Based Procedures, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 12–22. https://doi.org/10.1007/978-3-030-32689-0_2

Shattuck, D.W., Mirza, M., Adisetiyo, V., Hojatkashani, C., Salamon, G., Narr, K.L., Poldrack, R.A., Bilder, R.M., Toga, A.W., 2008. Construction of a 3D probabilistic atlas of human cortical structures. NeuroImage 39, 1064–1080. https://doi.org/10.1016/j.neuroimage.2007.09.031

Shorten, C., Khoshgoftaar, T.M., 2019. A survey on Image Data Augmentation for Deep Learning. Journal of Big Data 6, 60. https://doi.org/10.1186/s40537-019-0197-0

Sled, J.G., Zijdenbos, A.P., Evans, A.C., 1998. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans Med Imaging 17, 87–97. https://doi.org/10.1109/42.668698

Sokooti, H., Saygili, G., Glocker, B., Lelieveldt, B.P.F., Staring, M., 2019. Quantitative error prediction of medical image registration using regression forests. Medical Image Analysis 56, 110–121. https://doi.org/10.1016/j.media.2019.05.005

Sokooti, H., Saygili, G., Glocker, B., Lelieveldt, B.P.F., Staring, M., 2016. Accuracy estimation for medical image registration using regression forests. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 9902 LNCS, 107–115. https://doi.org/10.1007/978-3-319-46726-9_13

Sokooti, H., Yousefi, S., Elmahdy, M.S., Lelieveldt, B.P.F., Staring, M., 2021. Hierarchical Prediction of Registration Misalignment Using a Convolutional LSTM: Application to Chest CT Scans. IEEE Access 9, 62008–62020. https://doi.org/10.1109/ACCESS.2021.3074124

Staring, M., Bakker, M.E., Stolk, J., Shamonin, D.P., Reiber, J.H.C., Stoel, B.C., 2014. Towards local progression estimation of pulmonary emphysema using CT. Medical Physics 41, 021905. https://doi.org/10.1118/1.4851535

Stolk, J., Putter, H., Bakker, E.M., Shaker, S.B., Parr, D.G., Piitulainen, E., Russi, E.W., Grebski, E., Dirksen, A., Stockley, R.A., Reiber, J.H.C., Stoel, B.C., 2007. Progression parameters for emphysema: A clinical investigation. Respiratory Medicine 101, 1924–1930. https://doi.org/10.1016/j.rmed.2007.04.016

Unsgård, G., Solheim, O., Selbekk, T., 2014. Intraoperative Ultrasound in Neurosurgery, in: Jolesz, F.A. (Ed.), Intraoperative Imaging and Image-Guided Therapy. Springer New York, New York, NY, pp. 549–565. https://doi.org/10.1007/978-1-4614-7657-3_41

Vandemeulebroucke, J., Rit, S., Kybic, J., Clarysse, P., Sarrut, D., 2011. Spatiotemporal motion estimation for respiratory-correlated imaging of the lungs. Med Phys 38, 166–178. https://doi.org/10.1118/1.3523619

Vandemeulebroucke, J., Sarrut, D., Clarysse, P., 2007. The POPI-model, a point-validated pixel-based breathing thorax model. Proceeding of the XVth ICCR Conference 8.

Yang, X., Kwitt, R., Styner, M., Niethammer, M., 2017. Quicksilver: Fast predictive image registration – A deep learning approach. NeuroImage 158, 378–396. https://doi.org/10.1016/j.neuroimage.2017.07.008

Yaniv, Z., Lowekamp, B.C., Johnson, H.J., Beare, R., 2018. SimpleITK Image-Analysis Notebooks: a Collaborative Environment for Education and Reproducible Research. J Digit Imaging 31, 290–303. https://doi.org/10.1007/s10278-017-0037-8

Chapter 4: Discussion, Future Work and Conclusion

1. Discussion

Image registration is increasingly popular in medicine (Viergever et al., 2016). Perfect automated registration performance is not guaranteed. Errors in registration can have serious salient or inconspicuous consequences. As a result, the field of registration error and confidence estimation began developing. These algorithms aim to verify registrations on a per-case basis. Their potential is as wide-ranging as image registration itself – from quality control of large-scale datasets (Sokooti et al., 2021, 2019) to ensuring the performance of a registration in image-guided therapies (Brock et al., 2017; Paganelli et al., 2018).

Upon turning to the literature, we came to two realizations. First, the field of registration error and confidence estimation is relatively new and there are few unifying resources for new researchers. Second, it is disparate; the methods it contains are very diverse and do not lend easy to comparison. These two factors motivated Chapter 2 of the thesis. In Chapter 2, we provided a compendious review of the field of registration error and confidence estimation. The review is objective in its breadth and structured by the PRISMA guidelines. To enable the comparison of individual methods – and to provide structure to the field – we proposed a taxonomy to classify the methods so that they can be evaluated and compared in an organized fashion. Chapter 2 serves as a resource for those new to the field and those already in it. To help further the development of the field, we analyzed trends, provided suggestions for best practices and identified areas of future research. In addition, Chapter 2 will help structure and stimulate the growth of the field of registration error and confidence estimation.

In Chapter 3, we drew on the insights gained from Chapter 2 to implement a method that estimates registration error for MRI-ultrasound registrations in the context of Image-Guided Neurosurgery for brain tumour resections. The method is based on the error estimating

convolutional neural network (CNN) proposed by Eppenhof and Pluim for the task of lung computed tomography (CT) registrations (Eppenhof and Pluim, 2018). A notable feature of our implementation is the data generator that curates the training (and validation) data at each epoch. Experiments determining the number of epochs and level of data augmentation to train the model informed the training of the final model. The model's performance on the test set surpassed our goals and demonstrated that it generalized to new data well. The results of Chapter 3 mark a significant step forward in the development of a registration error estimating algorithm for Image-Guided Neurosurgery. At the end of Chapter 3, we outlined the work that lies ahead to continue this development and will describe the possible future work in more detail in the discussion below.

2. Future Work

Much research remains towards the end goal of implementing an error estimating method that will be usable in the clinic. To conclude this thesis, we apply the trends, suggestions and areas of future research outlined in Chapter 2 to continue the progress of Chapter 3. This includes improvements to the model and its training data, further experimentation, steps towards implementing the model on a neuronavigation system and other possible exciting developments and applications.

2.1. Model

A trend observed in Chapter 2 is the increasing complexity of the machine learning models used to estimate registration error. We implemented the architecture that Eppenhof and Pluim proposed in 2018 (Eppenhof and Pluim, 2018). While the architecture is relatively complex compared previous models (e.g., to the random forest regressors of Saygili and Sokooti et al. (Saygili, 2018; Sokooti et al., 2016)), it is not as complex as the recent convolutional long-short term memory network of Sokooti et al. (Sokooti et al., 2021). Such an architecture would likely produce smoother error estimates than the results obtained in Chapter 3 given that it does not consider each voxel independently like the sliding window CNN of Eppenhof and Pluim. In fact, Eppenhof and Pluim suggested increasing the complexity of the network architecture for

applications in brain MRI, such as considering information from multiple scales. Such a potential improvement to the model's architecture is worth investigating.

Another possible improvement to the method implemented in Chapter 3 is the information it receives. In its current form, it is a Machine Learning Framework that relies only on an Image-based Approach for its features – it uses only the image intensities. This may not be optimal. Two potential improvements include additionally using as input the deformation vector field and registration uncertainty. Both could be added to the model as another input layer. The registration deformation field can potentially provide useful information for the model to understand the scale of deformations that occurred during the registration as well as the smoothness relating to the physical or physiological plausibility of such deformations. An interesting result from Sokooti et al., also highlighted in Chapter 3, is the better generalization performance of their model when it is trained with Transformation-based features as opposed to Image-based features (Sokooti et al., 2019). This suggests the importance of retaining deformation field information for the error estimation. Similarly, the registration uncertainty may be a useful input to the model if it is provided by the registration algorithm (which is the case for some new deep learning methods (e.g., (Dalca et al., 2019; Yang et al., 2017)). Lotfi et al.'s error estimating model performed much better when uncertainty was included as a feature (Lotfi et al., 2013). Likewise, Simpson et al. showed that the registration uncertainty had a degree of predictive power when classifying Alzheimer's Disease and controls (Simpson et al., 2011). To incorporate these additional sources of information in the model, the training procedure would need to be adapted so it provides the deformation field and registration uncertainty along with the image intensities and known error.

Finally, there are potentially interesting modifications to the model's output. The first is to have the model output a vector of registration error as opposed to a scalar, as recently performed by Saygili (Saygili, 2021). This task is likely much more challenging than estimating a scalar given it is effectively registering the two images again (since a 3D displacement is estimated for each voxel). The second modification, at the other end of the spectrum, is to perform classification into different ranges of registration error. Sokooti et al. recently trained a model to classify the registration error into one of three classes (Sokooti et al., 2021). This could be an interesting approach for the task of Chapter 3 – perhaps a surgeon only wants to know if the error is greater than some threshold, e.g., 2 millimetres (mm). A related modification could

include a classification component to the model's regression output. The model could regress errors of a certain range ([0, X] mm) and anything beyond that range can be treated as 'Error above X mm.' This is one strategy, along with those presented in Chapter 3, that may help the model deal with the larger errors that it struggled with. Such an approach is akin to the method of Sedghi et al., which classified misalignment up to 8 voxels and included an 'unrelated' class (Sedghi et al., 2019). The 'unrelated' class proved to be important to their registration algorithm's performance. Making these modifications would require updating the network architecture and training data labels.

2.2. Training Data Complexity

While the model architecture is one aspect that will influence its performance, another is the data it is trained on. There are improvements to be made in this domain as well, on top of simply adding more training data. They include the suggestions mentioned in Chapter 3 like creating a balanced training set and more realistic simulated ultrasound, as well as a more diverse set of transformations, including potentially realistic brain shifts, to deform the simulated ultrasound images. For example, drawing inspiration from Sokooti et al. who recently used four different types of artificial deformations to train their model, including realistic respiratory motion deformations (Sokooti et al., 2021). Such changes to the training data require the new types of artificial deformations to be added to the data generator.

2.3. Experiments

There are several additional experiments that could be performed to further assess the model's performance. The first involves the complexity of the transformations used to evaluate the model. In Chapter 3, the model is evaluated with artificial deformations from B-spline-based transformations. This is the same type of transformation model used to create the training data – there is no model mismatch. To further test the model's ability, it should be evaluated with different types of transformations with varying complexity, as mentioned in Chapter 3. Furthermore, rather than only using artificial deformations, the model should also be tested after the deformed images are registered. Registering the images may introduce more subtle and more complex errors than those only from artificial deformations. In this case, the registration transformation model should be different than the artificial deformation transformation model to

ensure model mismatch. The second set of experiments involve evaluating the model with real ultrasound data, as outlined in Chapter 3. It is of utmost importance that the model is robust and generalizes well to new data. Therefore, to fully validate the model, it should be tested on data from outside of the BITE database (Mercier et al., 2012). Another database to test the model is the RESECT database (Xiao et al., 2017). Throughout the above experiments, it will be interesting to investigate some more general research questions identified in Chapter 2. This includes comparing the model's performance on artificial deformation and landmarks (and the role image contrast plays in this), and assessing the impact of transformation model mismatch.

The end goal of this research is to provide a means to estimate registration error on clinical neuronavigation systems. To this end, there are two important steps. The first involves implementing the model within IBIS ('Intraoperative Brain Imaging System'), the open-source neuronavigation platform developed by members of our lab (Drouin et al., 2017). Once available on IBIS, the second step is to obtain feedback from surgeons and other clinicians that will be using the registration error estimating method. A user study similar to that performed by Schlachter et al. may be appropriate (Schlachter et al., 2016). These steps will provide a functioning prototype of the model and suggestions to improve it for clinical use.

2.4. Related Opportunities

Beyond estimating error of MRI-ultrasound registrations, there are several related and exciting research avenues. I briefly touch on such topics. The first is implementing a registration error estimating algorithm for brain MRI registrations. Brain MRI registrations are fundamental for a range of different applications. A successful error estimating method could have widespread use. However, there are several factors to consider – whether the algorithm is designed for inter- or intra-subject registrations, whether for mono- or multimodal registrations and the clinical context it is implemented for. A requirement is that the brain MRI data has annotated corresponding landmarks, which is not as common as e.g., segmentations in brain MRI databases. A very interesting and relevant starting point is with the Brain Tumor Sequence Registration (BraTS-Reg) Challenge dataset, which contains annotated landmarks (Baheti et al., 2021). The second avenue is transfer learning – repurposing a trained model for a related task (Goodfellow et al., 2016). Saygili used transfer learning with a model trained for stereo matching and pointed out that transfer learning can help alleviate some of the issues related to a lack of

ground truth data (Saygili, 2020). Another application of transfer learning would be to incorporate the trained model of Eppenhof and Pluim into the implementation of Chapter 3 (Eppenhof and Pluim, 2018). A related avenue is how generalizable the error estimating models are and can be. Investigating how well a model trained for one task (e.g., lung CT) does on another task (e.g., brain MRI) would be interesting. Training a model with data from both of these tasks would also be worthwhile (i.e., with lung CT and brain MRI data). Perhaps the images can first be transformed into modality-independent features (Heinrich et al., 2012), as also suggested by Eppenhof and Pluim (Eppenhof and Pluim, 2018). Each of these research avenues advance the field of registration error estimation.

In pursuit of advancing the field, an important aspect is that different methods can be fairly and objectively compared to one another. Given the potential sources of bias in differing validation procedures (Bierbrier et al., 2022), there is a strong need for available data, open-source code and standard validation protocols. Ideally, a challenge will be organized to facilitate these catalysts of research.

3. Conclusion

The field of registration error and confidence estimation is full of potential. Building off the efforts of many others, this thesis attempted to further unlock the potential by first reviewing and unifying the field as it is, providing it structure and direction, and then by applying it to an application where it can have immediate benefit to patients – Image-Guided Neurosurgery. I sincerely hope these developments help the field realize its potential and, ultimately, help researchers and clinicians care for patients.

References

Archip, N., Clatz, O., Whalen, S., Kacher, D., Fedorov, A., Kot, A., Chrisochoides, N., Jolesz, F., Golby, A., Black, P.M., Warfield, S.K., 2007. Non-rigid alignment of pre-operative MRI, fMRI, and DT-MRI with intra-operative MRI for enhanced visualization and navigation in image-guided neurosurgery. NeuroImage 35, 609–624. https://doi.org/10.1016/j.neuroimage.2006.11.060

Baheti, B., Waldmannstetter, D., Chakrabarty, S., Akbari, H., Bilello, M., Wiestler, B., Schwarting, J., Calabrese, E., Rudie, J., Abidi, S., Mousa, M., Villanueva-Meyer, J., Marcus, D.S., Davatzikos, C., Sotiras, A., Menze, B., Bakas, S., 2021. The Brain Tumor Sequence Registration Challenge: Establishing Correspondence between Pre-Operative and Follow-up MRI scans of diffuse glioma patients. arXiv:2112.06979 [cs, eess].

Bierbrier, J., Gueziri, H.-E., Collins, D.L., 2022. Estimating Medical Image Registration Error and Confidence: A Taxonomy and Systematic Review. Medical Image Analysis.

Brock, K.K., Mutic, S., McNutt, T.R., Li, H., Kessler, M.L., 2017. Use of image registration and fusion algorithms and techniques in radiotherapy: Report of the AAPM Radiation Therapy Committee Task Group No. 132. Med. Phys. 44, e43–e76. https://doi.org/10.1002/mp.12256

Crum, W.R., Hartkens, T., Hill, D.L.G., 2004. Non-rigid image registration: theory and practice. BJR 77, S140–S153. https://doi.org/10.1259/bjr/25329214

Dalca, A.V., Balakrishnan, G., Guttag, J., Sabuncu, M.R., 2019. Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. Medical Image Analysis 57, 226–236. https://doi.org/10.1016/j.media.2019.07.006

Dickhaus, H., Ganser, K.A., Staubert, A., Bonsanto, M.M., Wirtz, C.R., Tronnier, V.M., Kunze, S., 1997. Quantification of brain shift effects by MR-imaging, in: Proceedings of the 19th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. "Magnificent Milestones and Emerging Opportunities in Medical Engineering" (Cat. No.97CH36136). Presented at the Proceedings of the 19th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. "Magnificent

Milestones and Emerging Opportunities in Medical Engineering" (Cat. No.97CH36136), pp. 491–494 vol.2. https://doi.org/10.1109/IEMBS.1997.757652

Drouin, S., Kochanowska, A., Kersten-Oertel, M., Gerard, I.J., Zelmann, R., De Nigris, D., Bériault, S., Arbel, T., Sirhan, D., Sadikot, A.F., Hall, J.A., Sinclair, D.S., Petrecca, K., DelMaestro, R.F., Collins, D.L., 2017. IBIS: an OR ready open-source platform for image-guided neurosurgery. Int J CARS 12, 363–378. https://doi.org/10.1007/s11548-016-1478-0

Eppenhof, K.A.J., Pluim, J.P.W., 2018. Error estimation of deformable image registration of pulmonary CT scans using convolutional neural networks. Journal of Medical Imaging 5, 1–1. https://doi.org/10.1117/1.jmi.5.2.024003

Fedorov, A., Risholm, P., Wells, W.M., 2014. Deformable Registration for IGT, in: Jolesz, F.A. (Ed.), Intraoperative Imaging and Image-Guided Therapy. Springer New York, New York, NY, pp. 211–223. https://doi.org/10.1007/978-1-4614-7657-3_14

Gerard, I.J., Kersten-Oertel, M., Hall, J.A., Sirhan, D., Collins, D.L., 2021. Brain Shift in Neuronavigation of Brain Tumors: An Updated Review of Intra-Operative Ultrasound Applications. Front. Oncol. 10, 618837. https://doi.org/10.3389/fonc.2020.618837

Gerard, I.J., Kersten-Oertel, M., Petrecca, K., Sirhan, D., Hall, J.A., Collins, D.L., 2017. Brain shift in neuronavigation of brain tumors: A review. Medical Image Analysis 35, 403–420. https://doi.org/10.1016/j.media.2016.08.007

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press.

Grimson, E., Kikinis, R., 2009. Chapter 42 - Registration for Image-Guided Surgery, in: Bankman, I.N. (Ed.), Handbook of Medical Image Processing and Analysis (Second Edition). Academic Press, Burlington, pp. 695–705. https://doi.org/10.1016/B978-012373904-9.50052-0

Hartkens, T., Hill, D.L.G., Castellano-Smith, A.D., Hawkes, D.J., Maurer, C.R., Martin, A.J., Hall, W.A., Liu, H., Truwit, C.L., 2003. Measurement and analysis of brain deformation during neurosurgery. IEEE Transactions on Medical Imaging 22, 82–92. https://doi.org/10.1109/TMI.2002.806596

Hastreiter, P., Rezk-Salama, C., Soza, G., Bauer, M., Greiner, G., Fahlbusch, R., Ganslandt, O., Nimsky, C., 2004. Strategies for brain shift evaluation. Medical Image Analysis 8, 447–464. https://doi.org/10.1016/j.media.2004.02.001

Heinrich, M.P., Jenkinson, M., Bhushan, M., Matin, T., Gleeson, F.V., Brady, S.M., Schnabel, J.A., 2012. MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration. Medical Image Analysis 16, 1423–1435. https://doi.org/10.1016/j.media.2012.05.008

Holden, M., 2008. A Review of Geometric Transformations for Nonrigid Body Registration. IEEE Trans. Med. Imaging 27, 111–128. https://doi.org/10.1109/TMI.2007.904691

Jannin, P., Grova, C., Maurer, C.R., 2006. Model for defining and reporting reference-based validation protocols in medical image processing. Int J CARS 1, 63–73. https://doi.org/10.1007/s11548-006-0044-6

Kelly, P.J., Kall, B.A., Goerss, S., Earnest, F., 1986. Computer-assisted stereotaxic laser resection of intra-axial brain neoplasms. Journal of Neurosurgery 64, 427–439. https://doi.org/10.3171/jns.1986.64.3.0427

Knowlton, R., 2009. Chapter 41 - Clinical Applications of Image Registration, in: Bankman, I.N. (Ed.), Handbook of Medical Image Processing and Analysis (Second Edition). Academic Press, Burlington, pp. 685–694. https://doi.org/10.1016/B978-012373904-9.50051-9

Lotfi, T., Tang, L., Andrews, S., Hamarneh, G., 2013. Improving Probabilistic Image Registration via Reinforcement Learning and Uncertainty Evaluation, in: Wu, G., Zhang, D., Shen, D., Yan, P., Suzuki, K., Wang, F. (Eds.), Machine Learning in Medical Imaging, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 187–194. https://doi.org/10.1007/978-3-319-02267-3_24

Maintz, J.B.A., Viergever, M.A., 1998. A survey of medical image registration. Medical Image Analysis 2, 1–36. https://doi.org/10.1016/S1361-8415(01)80026-8

Maurer, C.R., Fitzpatrick, J.M., 1993. A review of medical image registration. Interactive image-guided neurosurgery 1, 17–44.

Mercier, L., Del Maestro, R.F., Petrecca, K., Araujo, D., Haegelen, C., Collins, D.L., 2012. Online database of clinical MR and ultrasound images of brain tumors. Medical Physics 39, 3253–3261. https://doi.org/10.1118/1.4709600

Nabavi, A., Black, P.M., Gering, D.T., Westin, C.F., Mehta, V., Pergolizzi, R.S., Ferrant, M., Warfield, S.K., Hata, N., Schwartz, R.B., Wells, W.M., Kikinis, R., Jolesz, F.A., 2001. Serial intraoperative magnetic resonance imaging of brain shift. Neurosurgery 48, 787–797; discussion 797-798. https://doi.org/10.1097/00006123-200104000-00019

Paganelli, C., Meschini, G., Molinelli, S., Riboldi, M., Baroni, G., 2018. "Patient-specific validation of deformable image registration in radiation therapy: Overview and caveats." Medical Physics 15.

Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., Chou, R., Glanville, J., Grimshaw, J.M., Hróbjartsson, A., Lalu, M.M., Li, T., Loder, E.W., Mayo-Wilson, E., McDonald, S., McGuinness, L.A., Stewart, L.A., Thomas, J., Tricco, A.C., Welch, V.A., Whiting, P., Moher, D., 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ n71. https://doi.org/10.1136/bmj.n71

Rueckert, D., Schnabel, J.A., 2011. Medical Image Registration, in: Deserno, T.M. (Ed.), Biomedical Image Processing, Biological and Medical Physics, Biomedical Engineering. Springer Berlin Heidelberg, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-15816-2

Sastry, R., Bi, W.L., Pieper, S., Frisken, S., Kapur, T., Wells, W., Golby, A.J., 2017. Applications of Ultrasound in the Resection of Brain Tumors: Ultrasound in Brain Tumor Resection. J Neuroimaging 27, 5–15. https://doi.org/10.1111/jon.12382

Saygili, G., 2021. Predicting medical image registration error through independent directions. SIViP 15, 223–230. https://doi.org/10.1007/s11760-020-01784-3

Saygili, G., 2020. Predicting medical image registration error with block-matching using three orthogonal planes approach. Signal, Image and Video Processing 14, 1099–1106. https://doi.org/10.1007/s11760-020-01650-2

Saygili, G., 2018. Local-search based prediction of medical image registration error, in: Nishikawa, R.M., Samuelson, F.W. (Eds.), Medical Imaging 2018: Image Perception, Observer Performance, and Technology Assessment. Presented at the Image Perception, Observer Performance, and Technology Assessment, SPIE, Houston, United States, p. 49. https://doi.org/10.1117/12.2293740

Schlachter, M., Fechter, T., Jurisic, M., Schimek-Jasch, T., Oehlke, O., Adebahr, S., Birkfellner, W., Nestle, U., Buhler, K., 2016. Visualization of Deformable Image Registration Quality Using Local Image Dissimilarity. IEEE Trans. Med. Imaging 35, 2319–2328. https://doi.org/10.1109/TMI.2016.2560942

Sedghi, A., Kapur, T., Luo, J., Mousavi, P., Wells, W.M., 2019. Probabilistic Image Registration via Deep Multi-class Classification: Characterizing Uncertainty, in: Greenspan, H., Tanno, R., Erdt, M., Arbel, T., Baumgartner, C., Dalca, A., Sudre, C.H., Wells, W.M., Drechsler, K., Linguraru, M.G., Oyarzun Laura, C., Shekhar, R., Wesarg, S., González Ballester, M.Á. (Eds.), Uncertainty for Safe Utilization of Machine Learning in Medical Imaging and Clinical Image-Based Procedures, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 12–22. https://doi.org/10.1007/978-3-030-32689-0_2

Simpson, I.J.A., Woolrich, M.W., Groves, A.R., Schnabel, J.A., 2011. Longitudinal Brain MRI Analysis with Uncertain Registration, in: Fichtinger, G., Martel, A., Peters, T. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2011, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 647–654. https://doi.org/10.1007/978-3-642-23629-7_79

Sokooti, H., Saygili, G., Glocker, B., Lelieveldt, B.P.F., Staring, M., 2019. Quantitative error prediction of medical image registration using regression forests. Medical Image Analysis 56, 110–121. https://doi.org/10.1016/j.media.2019.05.005

Sokooti, H., Saygili, G., Glocker, B., Lelieveldt, B.P.F., Staring, M., 2016. Accuracy estimation for medical image registration using regression forests. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 9902 LNCS, 107–115. https://doi.org/10.1007/978-3-319-46726-9_13

Sokooti, H., Yousefi, S., Elmahdy, M.S., Lelieveldt, B.P.F., Staring, M., 2021. Hierarchical Prediction of Registration Misalignment Using a Convolutional LSTM: Application to Chest CT Scans. IEEE Access 9, 62008–62020. https://doi.org/10.1109/ACCESS.2021.3074124

Sotiras, A., Davatzikos, C., Paragios, N., 2013. Deformable Medical Image Registration: A Survey. IEEE Trans. Med. Imaging 32, 1153–1190. https://doi.org/10.1109/TMI.2013.2265603

Viergever, M.A., Maintz, J.B.A., Klein, S., Murphy, K., Staring, M., Pluim, J.P.W., 2016. A survey of medical image registration – under review. Medical Image Analysis 33, 140–144. https://doi.org/10.1016/j.media.2016.06.030

Xiao, Y., Fortin, M., Unsgård, G., Rivaz, H., Reinertsen, I., 2017. REtroSpective Evaluation of Cerebral Tumors (RESECT): A clinical database of pre-operative MRI and intra-operative ultrasound in low-grade glioma surgeries. Medical Physics 44, 3875–3882. https://doi.org/10.1002/mp.12268

Yang, X., Kwitt, R., Styner, M., Niethammer, M., 2017. Quicksilver: Fast predictive image registration – A deep learning approach. NeuroImage 158, 378–396. https://doi.org/10.1016/j.neuroimage.2017.07.008