This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Medical Image Registration Metrics Examined for TRUS-Guided Prostate Interventions

Mengting Liu[1], Mengting Yang[1], Shixing Ma[1], Yipeng Hu[3], and Zhe Min[1,2,3⋆]

[1] School of Control Science and Engineering, Shandong University, Jinan, China
minzhe@sdu.edu.cn
[2] Shenzhen Loop Area Institute, Shenzhen, China
[3] UCL Hawkes Institute, Department of Medical Physics and Biomedical Engineering, University College London, London, UK

Abstract. Accurate image registration is central to image-guided interventions, yet the choice of evaluation metric can substantially influence how registration quality is interpreted, because region overlap, boundary alignment, and target localization capture different error dimensions. We use MR–TRUS prostate interventions as a concrete application to quantify the relationships, agreement, and residual disagreement among Dice similarity coefficient (DSC), target registration error (TRE), and 95th percentile Hausdorff distance (HD95). TRE reflects landmark-based local registration error that is relevant to target-directed tasks such as biopsy or focal therapy, but is often unavailable because landmark annotation is labor-intensive. This work addresses three questions: (1) how strongly DSC, TRE, and HD95 are correlated and agree at the case level; (2) whether metric choice changes the comparison of registration methods; and (3) whether DSC and HD95 can approximate landmark-based TRE when landmarks are unavailable. Seven deep learning-based registration methods were evaluated using a 108-case MR–TRUS prostate registration cohort. We analyzed case-wise metric relationships using Pearson and Spearman correlations and pairwise agreement with Cohen's kappa. We further assessed whether metric choice changes the comparison of registration methods by ranking the seven methods using mean DSC, TRE, and HD95, and by comparing the best-performing method selected by each metric for each case. Linear regression models were also fitted to predict landmark-based TRE from DSC and HD95. The metrics showed significant associations across methods, with strong-to-very-strong Pearson correlation coefficients (PCCs) between all metric pairs ($|r| \in [0.799, 0.970]$, all $p < 0.001$). However, agreement was imperfect. Cohen's kappa ranged from 0.457 to 0.866, and metric pairs involving TRE showed weaker agreement than DSC–HD95. When methods were ranked by mean DSC, TRE, and HD95, the top two methods remained consistent across metrics, but lower-ranked method ordering and case-wise best-method selection changed with the metric. DSC and HD95 also approximated TRE with $R^2 \in [0.69, 0.84]$ across individual methods and a pooled RMSE of 1.463 mm over 245 method-case predictions. These findings suggest that DSC, TRE, and HD95 are correlated but not

⋆ Corresponding author.

interchangeable, supporting task-oriented multi-metric evaluation when landmark-based TRE is unavailable.

Keywords: Medical image registration · Registration evaluation metrics · Target registration error

1 Introduction

Image registration is a key enabling technique for ultrasound-guided interventions, where preoperative imaging needs to be aligned with intraoperative ultrasound to support localization, navigation, and treatment planning [13]. In prostate interventions, such as targeted biopsy and focal therapy, magnetic resonance (MR) imaging provides high-contrast soft-tissue and lesion information, whereas transrectal ultrasound (TRUS) provides real-time intraoperative guidance [12,13].

Despite its clinical importance, validation of medical image registration remains difficult. The true spatial correspondence between two clinical images is rarely known, especially for deformable registration, where ground-truth dense pixels are generally unavailable [19]. Registration errors can mislead image-guided therapy and bias downstream image analysis [3]. As a result, registration studies often rely on indirect evaluation metrics, such as image similarity measures including mutual information (MI) and structural similarity index measure (SSIM), rather than direct comparison with a known ground-truth transformation [5]. In MR–TRUS prostate registration, commonly reported metrics include DSC, HD95, and TRE.

A potential pitfall is that these metrics are often reported side by side as if they were interchangeable summaries of registration accuracy. In fact, DSC reflects region-level agreement, HD95 emphasizes boundary alignment, and TRE captures landmark-based local error. Different prostate interventions place different emphasis on specific aspects of registration error. TRE supports target localization in biopsy and focal therapy [12,13], DSC reflects whole-gland alignment for prostatectomy-related mapping [22], and HD95 is relevant to inner-gland or transition-zone boundary assessment [1], which can be important for BPH treatment planning or vaporization [23]. This distinction is illustrated in Fig. 1, where a registration result with high DSC may still have clinically relevant TRE, whereas a result with low TRE does not necessarily achieve optimal HD95.

TRE is regarded as a clinically meaningful metric because it directly measures the spatial error at anatomical landmarks or targets [11,7]. Classical point-based registration theory modeled TRE from fiducial localization and registration errors [10], and later work introduced variants such as total TRE to account for localization uncertainty in both registered spaces [17]. However, TRE requires corresponding anatomical landmarks, which are labor-intensive to annotate and may be unavailable in many clinical datasets. In such cases, overlap- and boundary-based metrics such as DSC and HD95 are commonly used as more

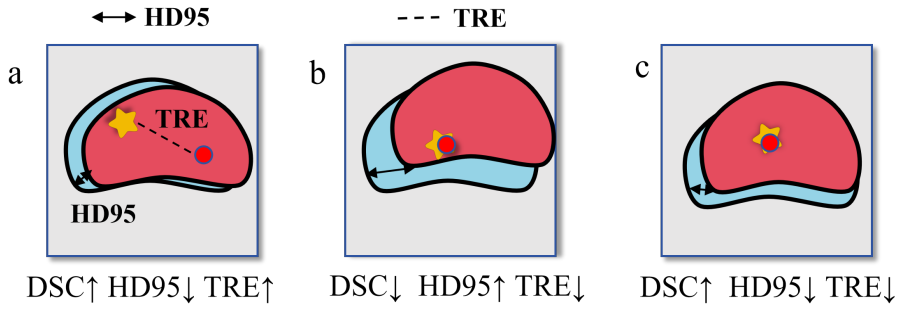

Fig. 1. Illustration of task-oriented differences among registration metrics. Blue and red contours denote the reference and warped prostate boundaries, respectively; the yellow star and red dot denote the reference and transformed target locations. (a) Good overlap but poor target matching; (b) accurate target matching but poor boundary alignment; (c) consistent overlap, target localization, and boundary alignment.

accessible alternatives. The key question is therefore not whether these metrics are useful, but whether they support interpretations that are sufficiently consistent with landmark-based TRE.

This concern is consistent with recent discussions in biomedical image analysis that metrics should be selected according to the intended clinical or scientific use rather than treated as interchangeable summaries [16,21]. For example, voxel-level segmentation scores may remain high despite missed lesions, just as good gland overlap in registration may still leave substantial local landmark error. For prostate cancer localization, Yan et al. showed that voxel-level segmentation metrics such as DSC and Hausdorff distance can disagree with lesion-level localization outcomes, even when statistical correlations are present [27]. A similar issue has been raised for point-set registration in computer-assisted surgery, where distance-based metrics may not consistently reflect clinically meaningful target registration error under partial overlap or anisotropic noise [8]. However, it remains unclear how this issue manifests in MR–TRUS prostate image registration, where DSC, HD95, and TRE are often reported together but may lead to different case-level or method-level conclusions.

In this work, we systematically examine the relationships among DSC, TRE, and HD95 for MR–TRUS prostate registration. Using 108 cases and seven learning-based registration methods, we investigate the following specific questions in this application. **(1) How strongly are DSC, TRE, and HD95 associated at the case level? (2) Do different metrics lead to consistent method rankings and case-wise best-method selections? (3) Can overlap- and boundary-based metrics approximate TRE when landmark annotations are unavailable?** Our aim is not to replace one metric with another, but to support task-aware interpretation of registration performance according to whether the downstream intervention depends primarily on regional overlap, boundary alignment, or target localization.

2 Materials and Methods

2.1 Dataset and Registration Methods

We used a prostate MR-TRUS registration dataset derived from the SmartTarget biopsy trial [12]. The data consist of paired T2-weighted MR and reconstructed TRUS volumes with prostate gland masks. In all experiments, TRUS was used as the fixed image and MR was used as the moving image; registration therefore warped MR into the TRUS fixed image space. Following the experimental split used in the source study, the dataset was divided into 65 training, 8 validation, and 35 test cases. All volumes were resampled to isotropic 0.8 mm spacing, intensity-normalized to $[0, 1]$, and center-padded to $128 \times 128 \times 128$ voxels.

Seven representative learning-based registration methods were evaluated: LocalNet [14], VoxelMorph [2], MIDIR [20], TransMorph [4], ModeT [24], RDP [25], and MambaMorph [26]. For a controlled comparison, all learning-based methods were retrained on the same training split using a common weakly supervised objective consisting of Dice loss on the warped prostate masks and L2 gradient regularization of the predicted displacement field to encourage smooth deformations. All models were optimized using Adam with a learning rate of 1×10^{-4} for 200 epochs. These methods cover weakly supervised and unsupervised training paradigms, convolutional architectures, Transformer-based designs, diffeomorphic or pyramid-based deformation modeling, and recent state-space-model-based registration.

2.2 Evaluation Metrics

For each registered test pair, we computed DSC, TRE, and HD95 in the fixed image space. Let $\Omega \subset \mathbb{R}^3$ denote the 3D image domain, and let $A, B \subset \Omega$ be the fixed prostate mask and the warped moving prostate mask, respectively. DSC measures region overlap and is defined as $\mathrm{DSC}(A, B) = 2|A \cap B|/(|A| + |B|)$, where $|\cdot|$ denotes the number of foreground voxels. For a case with K available landmark pairs, let $\mathbf{x}_i^f \in \mathbb{R}^3$ and $\mathbf{x}_i^w \in \mathbb{R}^3$ be the physical coordinates of the i-th fixed landmark and the corresponding warped moving landmark, respectively. The mean TRE is computed as $\mathrm{TRE} = K^{-1} \sum_{i=1}^{K} \|\mathbf{x}_i^f - \mathbf{x}_i^w\|_2$. HD95 measures boundary discrepancy between the fixed and warped moving masks. Let $\partial A, \partial B \subset \mathbb{R}^3$ denote their 2D boundary surfaces embedded in the 3D physical space, and let $d(\mathbf{x}, S) = \min_{\mathbf{y} \in S} \|\mathbf{x} - \mathbf{y}\|_2$ be the point-to-surface distance. HD95 is computed as the 95th percentile of the bidirectional surface distance set $\{d(\mathbf{x}, \partial B) \mid \mathbf{x} \in \partial A\} \cup \{d(\mathbf{y}, \partial A) \mid \mathbf{y} \in \partial B\}$. TRE and HD95 were reported in millimeters. Higher DSC indicates better overlap, whereas lower TRE and HD95 indicate smaller target localization and boundary errors.

2.3 Metric-Relationship Analysis

Pairwise correlation. For each registration method, Pearson and Spearman correlations were computed across the 35 test cases. Each case contributed one

post-registration DSC, TRE, and HD95 value, which were used to compute correlations for DSC–TRE, DSC–HD95, and TRE–HD95. Pearson correlation quantifies linear association, whereas Spearman correlation quantifies monotonic agreement in case ranking. Following commonly used interpretation thresholds [18,9], the absolute values of correlation coefficients were described as negligible (0.00–0.19), weak (0.20–0.39), moderate (0.40–0.59), strong (0.60–0.79), or very strong (0.80–1.00), respectively.

Pairwise case-wise agreement. To assess whether two metrics induced consistent relative rankings of test cases, we exhaustively compared all unique pairs among the 35 post-registration cases for each method, resulting in $\binom{35}{2} = 595$ case pairs. Metrics were converted to a higher-is-better direction as $s_{\mathrm{DSC}} = \mathrm{DSC}$, $s_{\mathrm{TRE}} = -\mathrm{TRE}$, and $s_{\mathrm{HD95}} = -\mathrm{HD95}$. For each pair of metrics, a case pair (i, j) was excluded if either metric assigned equal scores to the two cases. A metric m preferred case i over case j if $s_m(i) > s_m(j)$. Two metrics were considered to agree when they preferred the same case. We reported the proportion of concordant case pairs as raw agreement together with Cohen's kappa. Kappa values were interpreted as slight (0.00–0.20), fair (0.21–0.40), moderate (0.41–0.60), substantial (0.61–0.80), or almost perfect (0.81–1.00) agreement [15].

Method ranking consistency. We examined whether metric choice affects method selection. Methods were ranked by mean DSC, mean TRE, and mean HD95 on the test set. Agreement among rankings was measured using Spearman correlation, Kendall correlation, Kendall's coefficient of concordance W.

Regression-based TRE approximation. Since landmark annotation is often unavailable in clinical data, we tested whether DSC and HD95 can jointly approximate TRE. For each method, a linear model was fitted across the 35 test cases:

$$\mathrm{TRE} = \beta_0 + \beta_1 \mathrm{DSC} + \beta_2 \mathrm{HD95} + \epsilon. \tag{1}$$

In this model, the parameters $\beta_0, \beta_1, \beta_2 \in \mathbb{R}$ are the fitted regression coefficients, where β_0 is the intercept and β_1 and β_2 quantify the contributions of DSC and HD95, respectively. The residual term $\epsilon \in \mathbb{R}$ represents the unexplained difference between the actual TRE and the predicted TRE. The predicted value was therefore computed as $\mathrm{TRE}_{\mathrm{pred}} = \beta_0 + \beta_1 \mathrm{DSC} + \beta_2 \mathrm{HD95}$. We report the coefficient of determination (R^2), standardized coefficients, and the pooled actual-vs-predicted TRE distribution across all method-case predictions. Following Cohen's commonly used effect-size interpretation, $R^2 < 0.02$ was considered very weak, $0.02 \leq R^2 < 0.13$ weak, $0.13 \leq R^2 < 0.26$ moderate, and $R^2 \geq 0.26$ substantial [6].

3 Results

3.1 Test-Set Performance of Registration Methods

Table 1 summarizes test-set performance. ModeT [24] achieved the best mean performance across all three metrics, followed by RDP [25]. However, not all

Table 1. Test-set performance and metric-specific rankings of seven registration methods. Values are reported as mean ± SD over 35 test cases. Numbers in parentheses denote ranks based on the mean value, with rank 1 indicating the best performance.

Method	DSC ↑	TRE (mm) ↓	HD95 (mm) ↓
ModeT [24]	$\mathbf{0.80 \pm 0.11}$ (1)	$\mathbf{6.43 \pm 2.52}$ (1)	$\mathbf{7.05 \pm 3.82}$ (1)
RDP [25]	0.77 ± 0.12 (2)	7.95 ± 2.85 (2)	8.08 ± 3.71 (2)
LocalNet [14]	0.72 ± 0.12 (4)	8.17 ± 2.80 (3)	8.89 ± 3.07 (3)
TransMorph [4]	0.72 ± 0.15 (3)	8.90 ± 3.53 (4)	9.33 ± 4.38 (4)
MIDIR [20]	0.71 ± 0.11 (5)	9.15 ± 2.85 (5)	9.70 ± 3.09 (5)
VoxelMorph [2]	0.67 ± 0.14 (7)	9.65 ± 3.37 (6)	10.22 ± 3.75 (6)
MambaMorph [26]	0.68 ± 0.13 (6)	10.03 ± 3.52 (7)	10.52 ± 3.86 (7)

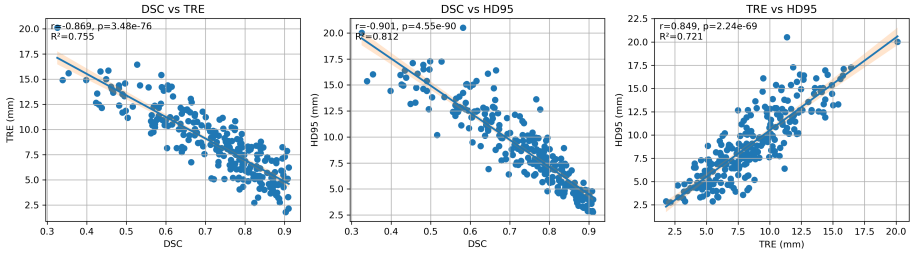

Fig. 2. Pooled relationships among DSC, TRE, and HD95 across all $7 \times 35 = 245$ method-case registration results. Each point represents one method-case output.

method rankings were identical across metrics. For example, TransMorph [4] ranked third by DSC but fourth by TRE and HD95, whereas LocalNet [14] ranked fourth by DSC but third by TRE and HD95. These rank changes suggest that average method comparison already depends to some extent on the chosen metric, motivating the metric-consistency analysis below.

3.2 Correlation and Agreement Among Metrics

Across the seven registration methods, DSC was negatively correlated with TRE and HD95, whereas TRE was positively correlated with HD95 (Fig. 2). The DSC–TRE and DSC–HD95 PCCs were very strong and negative, with $r \in [-0.912, -0.820]$ and $r \in [-0.970, -0.816]$, respectively, while TRE–HD95 showed strong-to-very-strong positive PCCs with $r \in [0.799, 0.911]$ (all $p < 0.001$). Spearman's ρ showed similar case-ordering trends for DSC–TRE, DSC–HD95, and TRE–HD95, with ranges of $[-0.910, -0.609]$, $[-0.969, -0.809]$, and $[0.673, 0.898]$, respectively. These results indicate that the metrics were directionally consistent in many cases, with the strongest consistency observed between DSC and HD95, while TRE retained partially distinct target-localization information.

Pairwise case-wise agreement further showed that the metrics were related but not interchangeable. Raw agreement ranged from approximately 0.73 to 0.93,

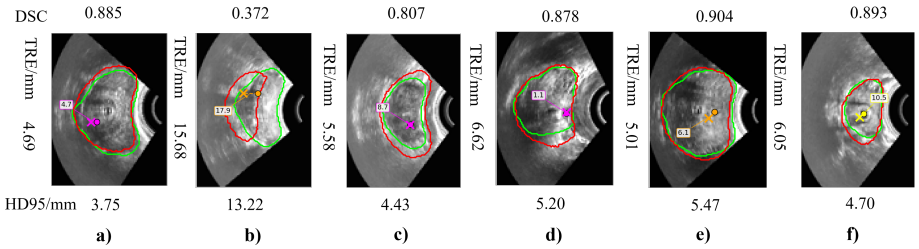

Fig. 3. Representative case-pair examples. Green/red contours denote fixed/warped boundaries; circles/crosses denote fixed/warped landmarks; boxed numbers show per-landmark TRE. (a,b) all metrics select (a); (c,d) DSC selects (d) but TRE selects (c); (e,f) HD95 selects (f) but TRE and DSC select (e).

and Cohen's kappa ranged from 0.46 to 0.87. DSC–HD95 achieved the highest agreement, with $\kappa \in [0.62, 0.87]$, whereas DSC–TRE and TRE–HD95 showed weaker agreement, mainly in the moderate-to-substantial range. Representative case pairs in Fig. 3 illustrate both complete agreement and disagreement patterns involving TRE. We therefore conclude that, despite significant correlations, overlap- and boundary-based metrics may give different case-level preference judgments from landmark-based TRE and should be interpreted with caution.

3.3 Metric Choice Can Affect Registration-Method Ranking

We ranked the seven registration methods using their mean DSC, TRE, and HD95 over the 35 test cases. The overall rankings were largely consistent: ModeT [24] ranked first and RDP [25] ranked second under all three metrics, and TRE and HD95 produced identical rankings. DSC-based rankings also agreed strongly with error-based rankings, with Spearman's $\rho = 0.929$ and Kendall's $\tau = 0.810$. However, the rankings were not identical, as LocalNet [14] and TransMorph [4], as well as VoxelMorph [2] and MambaMorph [26], changed their relative order across metrics.

We also compared case-wise best-method selection. For each test case, the best method was selected separately according to DSC, TRE, and HD95. All three metrics selected the same best-performing method in 21 of 35 cases (60%), while the selected method differed in the remaining 14 cases (40%). The average Kendall's W across cases was 0.83, indicating high but imperfect agreement. These results suggest that metric choice has limited influence on the top average-performing methods, but can affect lower-ranked method ordering and case-wise method selection.

3.4 Can DSC and HD95 Approximate TRE When Landmarks Are Missing?

The regression analysis showed that DSC and HD95 jointly explained much of the variation in TRE. Across individual registration methods, R^2 ranged from 0.69

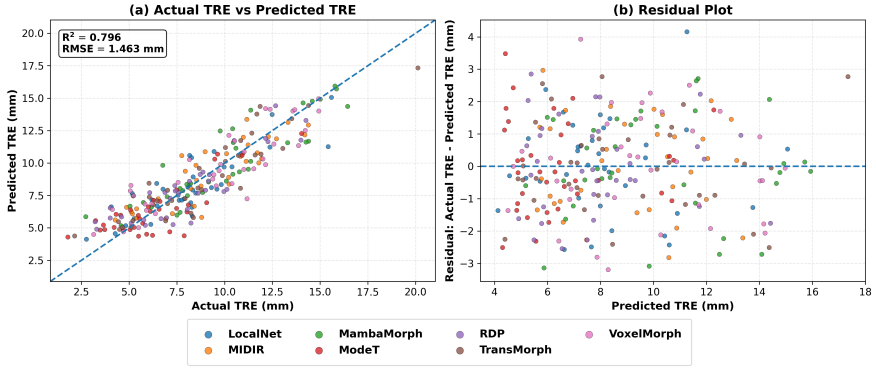

Fig. 4. Regression-based approximation of TRE using DSC and HD95. (a) Actual versus predicted TRE values; the dashed line denotes perfect prediction. (b) Residual plot, where residual is defined as actual TRE minus predicted TRE.

to 0.84. Standardized coefficients indicated that DSC generally had a stronger effect ($\beta_{\text{std}} \approx -0.39$ to -0.67), while HD95 provided stable complementary information ($\beta_{\text{std}} \approx 0.25$ to 0.50). In the pooled regression analysis across all 245 method-case predictions, the actual-versus-predicted TRE relationship reached $R^2 = 0.796$ with an RMSE of 1.463 mm (Fig. 4).

As shown in Fig. 4(b), residual errors remained after regression-based TRE approximation. Across all 245 method-case predictions, the absolute value of residual was 1.18 ± 0.86 mm, with signed residuals ranging from -3.19 to 4.16 mm. These results suggest that DSC and HD95 can help approximate TRE trends when landmark annotations are unavailable, but they cannot fully substitute for landmark-based TRE, which may capture landmark-specific or local anatomical errors not reflected by overlap or boundary distance alone.

4 Discussion and Conclusion

This study addressed three questions about whether DSC, TRE, and HD95 are associated, whether metric choice affects method comparison, and whether DSC and HD95 can approximate TRE in MR–TRUS prostate registration. Significant correlations among all metric pairs indicate that overlap, boundary alignment, and landmark-based local-error measures are often directionally consistent. However, imperfect kappa agreement shows that high correlation is insufficient to guarantee interchangeable case-level judgments.

Metric choice had limited influence on identifying the strongest methods, as the top two methods were consistent across DSC, TRE, and HD95. Nevertheless, lower-ranked method ordering and case-wise best-method selection changed with the metric, indicating that method comparison can depend on whether the task emphasizes region-level overlap, boundary alignment, or landmark-based local error. DSC and HD95 captured substantial TRE variation and may provide

auxiliary information when landmarks are unavailable. However, residual errors and case-wise inconsistencies show that they should not be treated as direct substitutes for landmark-based TRE, especially for biopsy or ablation guidance.

This analysis was limited to prostate MR–TRUS registration and seven learning-based methods, requiring validation across other organs, modalities, and workflows. Overall, DSC, TRE, and HD95 are correlated but not interchangeable, and registration evaluation should select and report metrics according to the downstream clinical task.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grants 62303275, U25A20145, and 22IAA01849, and in part by the Jinan Science and Technology Bureau under Grant 202333011.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Aldoj, N., Biavati, F., Michallek, F., Stober, S., Dewey, M.: Automatic prostate and prostate zones segmentation of magnetic resonance images using densenet-like u-net. Scientific Reports **10**(1), 14315 (2020). https://doi.org/10.1038/s41598-020-71080-0
2. Balakrishnan, G., Zhao, A., Sabuncu, M.R., et al.: VoxelMorph: A learning framework for deformable medical image registration. IEEE Transactions on Medical Imaging **38**(8), 1788–1800 (2019)
3. Bierbrier, J., Gueziri, H.E., Collins, D.L.: Estimating medical image registration error and confidence: A taxonomy and scoping review. Medical Image Analysis **81**, 102531 (2022)
4. Chen, J., Frey, E.C., He, Y., et al.: TransMorph: Transformer for unsupervised medical image registration. Medical Image Analysis **82**, 102615 (2022)
5. Chen, J., Liu, Y., Wei, S., Bian, Z., Subramanian, S., Carass, A., Prince, J.L., Du, Y.: A survey on deep learning in medical image registration: New technologies, uncertainty, evaluation metrics, and beyond. Medical Image Analysis **100**, 103385 (2025)
6. Cohen, J.: Statistical Power Analysis for the Behavioral Sciences. Routledge, New York, 2 edn. (1988)
7. Danilchenko, A., Fitzpatrick, J.M.: General approach to first-order error prediction in rigid point registration. IEEE Transactions on Medical Imaging **30**(3), 679–693 (2011). https://doi.org/10.1109/TMI.2010.2091513
8. Du, X., Ma, S., Liu, M., Zhang, Z., Min, Z.: Point set registration metrics reloaded for computer-assisted surgery. In: Shape in Medical Imaging. Lecture Notes in Computer Science, vol. 16171, pp. 319–334. Springer (2026)
9. Evans, J.D.: Straightforward Statistics for the Behavioral Sciences. Brooks/Cole, Pacific Grove (1996)
10. Fitzpatrick, J.M., West, J.B., Maurer, C.R.: Predicting error in rigid-body point-based registration. IEEE Transactions on Medical Imaging **17**(5), 694–702 (1998)

11. Fitzpatrick, J.M.: The role of registration in accurate surgical guidance. Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine **224**(5), 607–622 (2010). https://doi.org/10.1243/09544119JEIM589
12. Hamid, S., Donaldson, I.A., Hu, Y., Rodell, R., Villarini, B., Bonmati, E., Tranter, P., Punwani, S., Sidhu, H.S., Willis, S., van der Meulen, J., Hawkes, D., McCartan, N., Potyka, I., Williams, N.R., Brew-Graves, C., Freeman, A., Moore, C.M., Barratt, D., Emberton, M., Ahmed, H.U.: The smarttarget biopsy trial: A prospective, within-person randomised, blinded trial comparing the accuracy of visual-registration and magnetic resonance imaging/ultrasound image-fusion targeted biopsies for prostate cancer risk stratification. European Urology **75**(5), 733–740 (2019). https://doi.org/https://doi.org/10.1016/j.eururo.2018.08.007
13. Hu, Y., Ahmed, H.U., Taylor, Z., et al.: MR to ultrasound registration for image-guided prostate interventions. Medical Image Analysis **16**(3), 687–703 (2012)
14. Hu, Y., Modat, M., Gibson, E., et al.: Weakly-supervised convolutional neural networks for multimodal image registration. Medical Image Analysis **49**, 1–13 (2018)
15. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**(1), 159–174 (1977)
16. Maier-Hein, L., Reinke, A., Godau, P., et al.: Metrics reloaded: Recommendations for image analysis validation. Nature Methods **21**, 195–212 (2024)
17. Min, Z., Ren, H., Meng, M.Q.H.: TTRE: A new type of error to evaluate the accuracy of a paired-point rigid registration. In: IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 1166–1171 (2017)
18. Mukaka, M.M.: A guide to appropriate use of correlation coefficient in medical research. Malawi Medical Journal **24**(3), 69–71 (2012)
19. Pluim, J.P.W., Muenzing, S.E.A., Eppenhof, K.A.J., Murphy, K.: The truth is hard to make: Validation of medical image registration. In: International Conference on Pattern Recognition. pp. 2294–2299 (2016)
20. Qiu, H., Qin, C., Schuh, A., et al.: Learning diffeomorphic and modality-invariant registration using B-splines. In: Proceedings of the Fourth Conference on Medical Imaging with Deep Learning. Proceedings of Machine Learning Research, vol. 143, pp. 645–664 (2021)
21. Reinke, A., Tizabi, M.D., Sudre, C.H., et al.: Understanding metric-related pitfalls in image analysis validation. Nature Methods **21**, 182–194 (2024)
22. Rusu, M., Shao, W., Kunder, C., Wang, J., Soerensen, S., Teslovich, N., Sood, R., Chen, L., Fan, R., Ghanouni, P., Brooks, J., Sonn, G.: Registration of presurgical mri and histopathology images from radical prostatectomy via rapsodi. Medical Physics **47** (07 2020). https://doi.org/10.1002/mp.14337
23. Sancha, F., Rivera, V., Georgiev, G., Botsevski, A., Kotsev, J., Herrmann, T.: Common trend: move to enucleation—is there a case for greenlight enucleation? development and description of the technique. World journal of urology **33** (06 2014). https://doi.org/10.1007/s00345-014-1339-9
24. Wang, H., Ni, D., Wang, Y.: ModeT: Learning deformable image registration via motion decomposition transformer. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. pp. 740–749. Lecture Notes in Computer Science, Springer, Cham (2023)
25. Wang, H., Ni, D., Wang, Y.: Recursive deformable pyramid network for unsupervised medical image registration. IEEE Transactions on Medical Imaging **43**(6), 2229–2240 (2024)
26. Wang, Y., Guo, T., Yuan, W., et al.: Mamba-based deformable medical image registration with an annotated brain MR-CT dataset. Computerized Medical Imaging and Graphics p. 102566 (2025)

27. Yan, W., Yang, Q., Syer, T., Min, Z., Punwani, S., Emberton, M., Barratt, D., Chiu, B., Hu, Y.: The impact of using voxel-level segmentation metrics on evaluating multifocal prostate cancer localisation. In: Wu, S., Shabestari, B., Xing, L. (eds.) Applications of Medical Artificial Intelligence. pp. 128–138. Springer Nature Switzerland, Cham (2022)