Published in final edited form as:
IEEE Trans Med Imaging. 2014 July ; 33(7): 1541–1550. doi:10.1109/TMI.2014.2317796.

Application of Tolerance Limits to the Characterization of Image Registration Performance

Andriy Fedorov[*],
Radiology Department, Brigham and Women's Hospital, Boston, MA 02115 USA.

William M. Wells,
Brigham and Women's Hospital, Radiology, Boston, MA 02115 USA, and also with Harvard Medical School, Boston, MA 02115 USA, and also with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (sw@bwh.harvard.edu).

Ron Kikinis,
Brigham and Women's Hospital, Radiology, Boston, MA 02115 USA and also with Harvard Medical School, Boston, MA 02115 USA kikinis@bwh.harvard.edu.

Clare M. Tempany, and
Brigham and Women's Hospital, Radiology, Boston, MA 02115 USA and also with Harvard Medical School, Boston, MA 02115 USA ctempany@bwh.harvard.edu.

Mark G. Vangel
Radiology Department, Massachusetts General Hospital, Boston, MA 02114 USA (vangel@nmr.mgh.harvard.edu).

Abstract

Deformable image registration is used increasingly in image-guided interventions and other applications. However, validation and characterization of registration performance remain areas that require further study. We propose an analysis methodology for deriving tolerance limits on the initial conditions for deformable registration that reliably lead to a successful registration. This approach results in a concise summary of the probability of registration failure, while accounting for the variability in the test data. The (β, γ) tolerance limit can be interpreted as a value of the input parameter that leads to successful registration outcome in at least 100β% of cases with the 100γ% confidence. The utility of the methodology is illustrated by summarizing the performance of a deformable registration algorithm evaluated in three different experimental setups of increasing complexity. Our examples are based on clinical data collected during MRI-guided prostate biopsy registered using publicly available deformable registration tool. The results indicate that the proposed methodology can be used to generate concise graphical summaries of the experiments, as well as a probabilistic estimate of the registration outcome for a future sample. Its use may facilitate improved objective assessment, comparison and retrospective stress-testing of deformable.

[*](fedorov@bwh.harvard.edu).

Keywords

Biomedical image processing; image registration; performance evaluation

I. Introduction

Validation is a critical step in the development of image analysis technology. It is concerned with the assessment of intrinsic characteristics, performance and limitations of a specific tool or algorithm. With the increasing use of image analysis both in research and clinic, robust validation protocols become essential to managing risks and reducing costs of image-guided procedures. Validation of nonrigid image registration aims to establish the ability of an algorithm to recover spatial correspondence in presence of deformation. Extensive work has been done to bound expected clinical accuracy of rigid registration based on phantom and retrospective experiments [1]. Fewer advances have been made for validating nonrigid registration [2]. Approaches to estimation of deformable registration uncertainty at arbitrary locations have been proposed [3], [4]; however, characterization of accuracy and reliability of nonrigid registration based on clinical data virtually always relies on comparison with a reference solution, such as locations of implanted fiducial markers or expert annotations of the images. The results of such reference-based evaluation are typically summarized using statistical measures that describe *average* performance of an algorithm. In practice, average performance is of limited utility. A much more practical measure is one that describes (based on the experimental data) the probability of the method producing a meaningful result the next time it is used, together with the associated uncertainty in this estimate. Here we investigate the use of tolerance limits [5] to provide such estimates. Compared to the commonly used summary statistics that aim to capture average or extreme results observed in the experimental evaluation, tolerance limits establish confidence bounds on a proportion of the experiments, thus characterizing the expected performance on new subjects.

A taxonomy for reference-based validation of image processing tools has been proposed by Jannin *et al.* [6]. Briefly, validation typically involves comparison of the results produced by a method under investigation with that of a reference. A reference can be obtained using a computational method that has been validated earlier, or using knowledge of a domain expert. Given the results produced by these two methods, a comparison function is used to measure the discrepancy, or the “distance” to the reference. In image registration, Target Registration Error (TRE) or Landmark Registration Error (LRE) are the distances commonly used [7]. The errors are computed for the different datasets and parameter values used in the validation, and are summarized by a quality index. The quality index captures statistical properties of the distribution of the local discrepancies at the intrinsic level (input dataset and fixed parameters) or global level (evaluation done using different parameters and validation datasets). The most commonly reported quality index is concerned with summarizing the average error observed in the evaluation. As an example, we examined the manuscripts presented in the Registration I and II sections of the MICCAI 2013 conference [8], and found that most of those articles concerned with the evaluation of a registration methods report mean and standard deviation of the error measure as the summary statistics

in the validation section. Although useful, the characterization of average behavior is not sufficient to describe the performance of an algorithm on a typical pair of images.

Another commonly reported summary statistic is the proportion of successful experiments. In our earlier work, we presented an evaluation of a deformable registration algorithm developed for image-guided prostate biopsy [9]. Success rate (proportion of experiments that were deemed successful based on the defined criteria) was reported individually for each of the datasets used in the evaluation. A similar approach was used in [10] and [11], where the capture range of the method was defined as the starting misalignment that led to a fixed success rate. This approach to reporting results does not directly account for variability observed across the datasets used in the evaluation, does not include uncertainty in the estimate and does not allow inference of the expected performance of the algorithm under similar experimental conditions.

In summary, none of the measures commonly used to summarize results of registration validation studies allow inference of typical (expected) performance of the method. A fundamental distinction between the typical-case and average-behavior scenarios is that behavior in a typical case must be regarded as a random variable, not an unknown constant. For example, the accuracy of a registration will vary from sample to sample, and when applied to a particular pair of images subject to a given failure criterion, an algorithm will either succeed or fail. Two types of intervals useful for inference on random variables are prediction intervals and tolerance intervals. A prediction interval will contain a specified proportion, e.g., 90%, of future values on average. A tolerance interval will contain a specified proportion of future values at specified confidence level, e.g., 90% of future values with 95% confidence.

We propose the use of tolerance intervals, or tolerance limits (one-sided intervals) to summarize the results of experimental evaluation of registration performance. Using this concept, we can derive tolerance limits on the initial conditions that reliably lead to successful registration, while accounting for the variability in the performance of the algorithm observed from the data used in the evaluation. Tolerance limits have been applied in numerous application domains [12], such as clinical chemistry [13], assessment of bioequivalence [14] and characterization of degradation processes in electronic components for aerospace applications [15]. In general, a tolerance limit allows one to ensure a fixed proportion of a population exceeds a bound with specified confidence [5].

When applied to the experimental assessment of registration method reliability, tolerance limits may provide a concise, quantitative and objective summary metric for characterizing the probability of future registration success, given the observed behavior of a registration method on the data used in the evaluation. In this study, tolerance limits are derived for a quantal response (i.e., succeed or fail) scenario. This has been done previously in a simpler and conceptually different application (lead testing) in [16]. In contrast to this earlier work, which employed Bayesian logistic regression and required simulation, the present frequentist approach is closed form, and builds on results in [17], [18].

The developed methodology is demonstrated in three examples of increasing complexity. Our examples utilize a versatile and commonly used registration tool. The annotated image data we use was collected in a clinical setup of image-guided prostate biopsy motivating the need for reliable deformable registration. Both the registration tool we evaluate and the annotated image data are publicly available.

II. Methodology

Tolerance intervals and limits are statistical concepts that so far have not been applied widely in evaluating image registration and image analysis tools in general. As mentioned above, tolerance intervals are used for inference on random variables. As a simple example from common experience, consider the problem of summarizing the results of the laboratory measurement of a component of a patient's blood. What the patient would like to know is not how his measurement compares to population average (which could be summarized by a mean and confidence interval), but rather whether his measurement is extreme when compared to an appropriate measurement population. Based on preliminary data, one can estimate an interval which contains at least 90% of the population of test results, with 95% confidence. If the patient's result falls outside this interval, then it can be highlighted for consideration. We now illustrate the applicability of one-sided tolerance intervals (tolerance limits) to image registration algorithm validation.

A (β, γ)lower tolerance limit (LTL) is a statistic such that at least a proportion β of the population of the random variable of interest exceeds this limit, with 100γ% confidence. In other words, a (β, γ)lower tolerance limit is a 100γ% lower confidence limit on the $1-\beta$ quantile of the population. Upper tolerance limits are defined similarly. Applied to registration method validation, a tolerance limit can specify, for example, that with 95% confidence at least 90% of the population of error values obtainable from an experimental setup will be less than 5 mm. Such a summary statistic allows one to characterize the expected performance of an algorithm.

The performance characteristic that we aim to describe is the probability of the registration tool producing a discrepancy (TRE or LRE) below an application-specific threshold with respect to the reference solution. In the derivation below, we apply the general taxonomy developed by Jannin *et al.* [6]. We define the validation criterion V_C as the probability of successful registration. Technical efficacy studies typically assume a setup where a certain input parameters (P_I) are varied in a controlled fashion in a series of experiments. As an example, the input parameter could be the magnitude of the initial misalignment of the images in studies investigating the sensitivity of the registration tool to the initial conditions. Each experiment produces a normalized output of the method $\hat{R}_{NM}$ (such as geometric points defined in the image) which is compared to the reference $\hat{R}_{Nref}$ to calculate TRE or LRE values. We define $F_C$as follows:

$$F_C = \begin{cases} 1, & TRE \leq T_{AE} \\ 0, & TRE > T_{AE} \end{cases}$$

where T_{AE} is the application-specific threshold on the acceptable value of the registration error. Note that F_C is a random variable, with distribution parametrized by P_I. Each registration is either successful ($F_C = 1$) or not ($F_C = 0$), and one would like to know a range of values of P_I for which we can guarantee with high confidence that for all P_I in this interval F_C will equal 1 with high probability (a tolerance limit).

For simplicity, we denote the parameter as P_I as x. Denote the probability of failure for the ith subject given a value x of the parameter being explored as

$$p_i(x) = Pr(F_C = 0|x).$$

We use nonparametric logistic regression to model the logit of the registration failure rate as a function of x

$$f_i(x) = logit(p_i(x))$$

where the logit transformation of a probability is defined as

$$logit(t) \equiv ln\left(\frac{t}{(1-t)}\right).$$

We will further assume that x is the magnitude of the initial mis-alignment, since this is an input parameter frequently explored in registration validation studies.

To estimate $f_i(x)$ we utilize local polynomial regression methods. Such methods employ low-order (usually linear or quadratic) polynomials with locally determined coefficients. This class of procedures has proven to be very useful in applied statistics (e.g., “lowess” [19] and, more generally, locally-weighted regression [20]). *locfit* is one such regression procedure [21], which is notable for determining the coefficients using local likelihood approximations. This fitting approach is applicable to general linear models, in particular to logistic regression. Therefore, we have chosen to use *locfit* as a nonparametric logistic regression method for our analysis.

The output of the function *locfit* applied to data on the ith sample includes a function $\hat{f_i}(x)$ which estimates logit ($p_i(x)$). The corresponding standard error function $\hat{\sigma}_i(x)$ is also provided. We assume that, for given x, $\hat{f_i}(x)$ is approximately Gaussian with standard error $\hat{\sigma}_i(x)$ (as shown in [21, p.167]).

We will construct a lower tolerance limit in order to ensure that for sufficiently small x, the algorithm will have an acceptably small probability of failure, with high confidence. Let F_C be the result of a registration simulation experiment E for which the initial misalignment vector norm is x, and for which the probability of failure is

$$Pr_E(F_C = 0|x) = p(x) = 1 - \beta.$$

We will use the data from *s* simulated registrations on each of *n* samples to estimate a bound $U(x)$ for which

$$Pr_U\left(U(x) \leq p(x)\right) \geq \gamma$$

where γ is a confidence level.

The subscripts indicate the random variables corresponding to the probability statements. A statistic $U(x)$ which has the above properties is a (β, γ) lower tolerance limit. We chose $\beta = 0.9$ and $\gamma = 0.95$ for our examples.

In the limit of infinitely many simulated registrations ($s \to \infty$), $\hat{f_i}(x) \to f_i(x)$, where $f_i(x)$ is the logit of the true probability of failure, and consequently $\hat{\sigma}_i(x) \to 0$. In the limit of infinitely many *samples* ($n \to \infty$), we assume that $\sum_i f_i(x)/n$ converges to a smooth function $f(x)$.

For fixed x, we employ a Gaussian random-effects model for $f_i(x)$

$$f_i(x) = f(x) + Z_i$$

where $\{Z_i\}_{i=1}^n$ are independent Gaussian random variables with standard deviation $\sigma(x)$. Of course, the values $Z_i(x)$ are correlated in x, but we will only be concerned with inference for fixed x.

Following Paule and Mandel [22], we estimate $f(x)$ by the weighted mean

$$\tilde{f}(x) \equiv \frac{\sum_{i=1}^n w_i(x)\hat{f}_i(x)}{\sum_{i=1}^n w_i(x)}$$

with weights

$$w_i(x) \equiv \frac{1}{\hat{\sigma}^2(x) + \hat{\sigma}_i^2(x)}.$$

The within-sample variances $\hat{\sigma}_i^2(x)$ are obtained from the logistic regressions. After having substituted the estimates $\hat{\sigma}_i^2(x)$ in the weights, the function

$$Q\left[\sigma^2(x)\right] \equiv \sum_{i=1}^n w_i(x)\left(\hat{f}_i(x) - \hat{f}(x)\right)^2 - (n-1) \quad (1)$$

can be seen to be an implicit function of the between-sample variance $\sigma(x)$, which we estimate as the root of this equation. In the Appendix, we motivate (1), and demonstrate that it is a monotone decreasing function of σ^2. This approach is equivalent to modified restricted maximum likelihood [18].

As shown in [17], an approximate 100γ% confidence interval for $f(x)$ is

$$\tilde{f}(x) \pm z_{\gamma} \frac{\sqrt{\sum_{i=1}^{n} w_i(x)^2 \left[\hat{f}_i(x) - \hat{f}(x)\right]^2}}{\sum_{i=1}^{n} w_i(x)}$$

where z_y is the 100γ percentile of a standard Gaussian distribution.

We assume that $\hat{f}(x)$ and the confidence limits are monotone increasing, at least for x sufficiently small. We select a value for a probability of failure which we want to ensure is exceeded only rarely, i.e., $p_0 \equiv 1 - \beta = 0.1$. We determine the critical misalignment norm x_0 such that the upper confidence limit at x_0 equals logit (p_0)

$$\tilde{f}(x_0) + z_{\gamma} \frac{\sqrt{\sum_{i=1}^{n} w_i(x_0)^2 \left[\hat{f}_i(x_0) - \tilde{f}(x_0)\right]^2}}{\sum_{i=1}^{n} w_i(x_0)} = logit(p_0). \quad (2)$$

Because of the monotonicity assumption and the fact that this is an upper confidence limit, we see that with 100γ% confidence the probability of failure is less than or equal to logit (p_0), for all misalignment norms $x \leq x_0$.

However, x_0 is of only limited use as a bound in applications, because we obviously need to make a probability statement not just about the data that we have analyzed, but also about the curve $f^*(x)$corresponding to a *future* sample. In order to do this, we merely replace the above confidence interval with a *prediction* interval, and then determine the value x_1 which satisfies the following equation, analogous to (2):

$$U(x) = \tilde{f}(x_1) + + z_{\gamma} \sqrt{\hat{\sigma}^2 + \frac{\sum_{i=1}^{n} w_i(x_1)^2 \left[\hat{f}_i(x_1) - \tilde{f}(x_1)\right]}{\left[\sum_{i=1}^{n} w_i(x_1)\right]^2}} = logit(p_0). \quad (3)$$

Since $U(x)$ is an upper prediction interval, for $x \leq x_1$

$$Pr_U(U(x) \leq p(x_1) = Pr_U(U(x) \leq 1 - \beta) \geq \gamma$$

and hence $U(x)$ is a (β, γ) lower tolerance limit.

The described analysis approach was implemented using the R statistical analysis package [23][1]. A custom Fortran routine was used for solving (1) using the Newton–Raphson algorithm.

Evaluation of the prediction performance was done using leave-one-out cross-validation. A (0.9,0.95) lower tolerance interval was estimated using 9 out of 10 cases, and for the case

[1] http://www.r-project.org

left out, the observed probability of failure for the input parameter values below the estimated LTL value was compared with the expected 10% probability of failure.

III. Examples

In this section, we apply the developed methodology to estimate tolerance limits in three illustrative examples. Each example corresponds to an experimental setup designed to evaluate the performance of a registration tool under various conditions. The goal of these experiments was neither to establish clinical relevance of the derived values of tolerance limits, nor to perform an in-depth technical validation of a specific registration method for a specific clinical application. Rather we aimed to demonstrate how tolerance limits can be applied to summarize the result of an experimental evaluation.

1) Image Data

All of our examples utilize a dataset of 10 annotated prostate MR images presented earlier in [9]. The MR images of the prostate were collected for the purposes of transperineal targeted MRI-guided biopsy [24]. Registration was performed between the T2-weighted (T2W) images obtained before the procedure for planning purposes (pre-procedural images) and the T2W images collected in the beginning of the biopsy (intra-procedural images). Pre-procedural MRI was performed in a 3T GE Signa HDx scanner. The patient was scanned in supine position using a combination of endorectal and pelvic array coils, resulting in T2W images with the 0.3 × 0.3 × 3 mm resolution. Intra-procedural scans were acquired in a 3T Siemens Magnetom Verio scanner using pelvic array coil and no endorectal coil, leading to different configu-ration of the prostate gland as compared to the pre-procedural images. Resolution of the intra-procedural T2W images was 0.4 × 0.4 × 3 mm.

For the purposes of registration and its evaluation, images were annotated with the manual segmentations of the prostate gland prepared using 3D Slicer[2][25] software. The images were also annotated with manually placed fiducial points at corresponding image-specific reference locations to enable calculation of the LRE. Identification of reliable point landmarks in prostate imaging is often challenging due to the lack of point-like anatomical features and poor out-of-plane resolution (as compared to in-plane). When identification of point landmarks was not possible, regions corresponding to the case-specific image features were contoured, and the landmarks were defined as centers of gravity for those regions.

We selected the dataset described above for several reasons. First, modeling of inter-subject variability cannot be done using a single pair of images and necessitates the use of a dataset containing multiple subjects representative of a clinical context. Second, datasets used in the evaluation must be annotated to provide reference for registration evaluation. Third, we were interested primarily in evaluating deformable registration, as it is most relevant in the clinical context of registering soft tissue image data. Few annotated datasets that meet these criteria are publicly available.

[2]Available online: http://slicer.org

2) Registration Methodology

Our experiments utilized BRAINSFit [9], [26], [27] as the registration method under investigation. BRAINSFit is an open source registration tool available as a module within 3D Slicer. Originally developed for intensity-based registration of multi-modal brain MRI data, BRAINSFit has been applied successfully in a variety of single-and multi-modality (MRI, PET/CT) registration problems and a range of organs (liver, brain, bone, abdomen, kidney) in both healthy and disease affected human and animal tissue[3].

Registration using BRAINSFit can be parameterized with different choices of transformation models, similarity metrics and initializations. The specific details of the parameterization we used is described for each of the examples separately. Mutual information was utilized as the similarity metric in all experiments. Optimization of the transformation parameters was achieved using the gradient descent optimizer for up to affine transformations, and using limited memory Broyden–Fletcher–Goldfarb–Shannon optimization with simple bounds for the final B-spline transformation step. We used the standard implementations of the registration framework components available in the Insight Toolkit (ITK) [28]. Registration was performed using identical parameters in all cases. Example 2 utilized the version of BRAINSFit tool that was based on version three of ITK, while in the other two examples we used a newer version of the software based on ITK version four.

3) Experimental Setup Considerations

A natural first approach to an experiment designed to validate a registration algorithm is to apply this algorithm to a sample of image pairs and simply observe the proportion of registration failures. However, any practically useful registration algorithm will have high reliability, and hence the sample size for a naïve validation experiment can be very large.

One way to overcome this problem is to stress the registration algorithm by introducing noise or perturbations to the algorithm configuration in such a way that failures occur sufficiently often to enable validation and comparison of algorithms with a reasonable sample size. This is analogous to accelerated life testing of materials [29], and is commonly applied to software validation [30].

Weillustratetheproposedmethodologywiththreeexamplesof validation experiments. All three examples make use of the same data: 10 pairs of prostate images, each from different subjects. In the first example, we used only the intra-procedural images and transformed these images so as to simulate the displacement due to a rectal coil. An important feature of this example is that the exact solution is known. For a second example, we consider the 10 pairs of images, and we assume that the quality metric is parametrized by misalignment norm alone. As discussed above, we added experimental noise in this example in order to make the registration sufficiently difficult so that useful information on reliability can be obtained. The ground truth registration is not known,andthereforetheresultisevaluatedrelativetothefiducial annotations defined by the

[3]Examples demonstrating the utility of BRAINSFit in a variety of use cases are described in http://na-mic.org/Wiki/index.php/Projects:RegistrationDocumentation:RegLibTable

domain expert. For the third example, we also use the image pairs with experimental noise. In contrast to the second example, for this case we investigate the effect of multiple parameters on the quality metric.

Our overall strategy in devising the example experiments was to choose an input parameter P_I that is expected to cause registration failure when its value exceeds a certain (unknown) threshold. We then identified the range of parameter values that allowed us to test the algorithm in the range of settings where failures change from very rare to very frequent events. We then applied the methodology developed above to establish the tolerance limit for the parameter value that leads to a high probability of success (e.g., 90%) with a high confidence in that estimate (e.g., 95%).

The examples we use are not intended to conduct a comprehensive validation of a specific registration method for a specific clinical use case. Instead, our goal was to illustrate how the methodology we developed for estimating LTL can be applied. Specifically, this requires availability of the experimental data collected from different subjects, and an adjustable parameter that influences the registration success rate. At the same time, we note that the experiments we considered are commonly applied in registration evaluation studies. Numerous examples can be found in the literature that evaluate registration performance using a setup similar to Example 2 [9], [11].

The datasets and analysis scripts used to derive the presented results will be made publicly available to facilitate reproducibility and encourage adoption of the methodology by the community[4].

A. Example 1: Synthetic Deformation

This experiment utilized the intra-procedural T2W images acquired without an endorectal coil to generate synthetic images that emulate deformation of the prostate gland by the endorectal coil. To define the deformation field, we first identified a centroid of the rectum in the central slice of the T2W image. A virtual line was then defined to pass through this point perpendicular to the axial image acquisition plane. Deformation vectors were defined to lie in the image acquisition plane in the directions radial to the point of intersection between the virtual line and the slice plane. The magnitude of the deformation vectors was defined as the input parameter for this experiment and was varied from 0 to 14 mm in increments of 0.2 mm. The resulting deformation fields were applied to each of the 10 intra-procedural T2W datasets to generate synthetic images.

Registration was performed between the original intra-procedural (fixed) and the synthetic deformed (moving) images. The BRAINSFit transformation hierarchy was configured to use only B-spline transformation. The number of samples used for the similarity metric calculation was set to the default value of 100 000. The number of iterations for the optimizer was fixed to 3000. Each registration experiment was repeated 10 times while varying the seed used for the random sampling of image voxels in the similarity metric calculation. Registration error was measured at the image point $\hat{R}_{N\,M}$ approximately

[4]See https://github.com/fedorov/ImageRegistrationToleranceIntervals

corresponding to the center of the prostate gland. Given the synthetic and registration-recovered transformations T_s and T_r, respectively, we defined the error as $\| \hat{R_{NM}} - T_r(T_s(\hat{R_{NM}})) \|$. Registration was considered a failure when this registration error exceeded 0.5 mm.

The probability of registration failure for each of the 10 cases as a function of the magnitude of the simulated displacement is shown in Fig. 1 (left). We can see that registration failures begin to occur when the magnitude of the simulated displacement is around 7 mm, and variability across the datasets in the onset of failures is present. Fig. 1 (right) demonstrates the application of the LTL estimation methodology to this experimental data. The (0.90,0.95) LTL was estimated to be 8.7 mm; i.e., with the 95% confidence we can state that 90% of the future experiments will lead to a successful registration outcome when the magnitude of the simulated displacement is less than or equal to 8.7 mm.

Results of leave-one-out cross-validation showed that for all leave-one-out experiments except one, the observed probability of registration failure was less than 10% for magnitudes of the simulated displacement below the estimated LTL. In one case, the LTL value was 9.2 mm, while observed failure probability exceeded 10% with the initial displacement of 9.0 mm, which was very close to the value of LTL.

B. Example 2: Perturbed Initialization by Randomized Translations

In this example we conducted a secondary analysis of the experimental results collected in an earlier study presented in [9]. Registration experiments were performed between the preand intra-procedural images of the prostate for each of the 10 cases of the prostate MRI dataset discussed earlier. The images were aligned rigidly to bring the images into rough alignment. The experiments involved perturbations of these rigidly aligned images by a known transform, applying the BRAINSFit registration and evaluating the quality of registration using Dice Similarity Coefficient (DSC) [31] between the gland segmentations and LRE [7]. BRAINSFit was configured to perform automatic hierarchical registration parameterized by rigid, affine and B-spline transformations. Manual contours of the prostate gland in the planning and intra-procedural T2w MRI were used to restrict calculation of the similarity measure to the prostate ROI, as this proved to significantly improve robustness of the algorithm.

The total of 500 initial misalignment transformations were defined by sampling the direction uniformly over the unit sphere, and selecting the misalignment magnitude D (i.e., input parameter P_I) uniformly in (0, 10) mm. The comparison function F_C in this example was defined based on an application-specific quality index O_{QI} that included the following criteria: 1) the registration optimization procedure converged; 2) the DSC between the segmentations of the gland in the registered images was above the case-specific threshold; 3) the landmark registration error (LRE) was less than image slice thickness (3 mm for our data). A more detailed description of the registration method and the evaluation results is available in [9]. Dichotomized (success/failure as a function of the initial misalignment) results of the experimental evaluation were used as the input for the statistical summary analysis. An example of the registration result for one of the cases is shown in Fig. 2.

We applied the developed methodology to approximate the registration failure process for each of the 10 samples using local polynomial regression. The resulting curves summarizing failure behavior are shown in Fig. 3 (left). In order to assess the adequacy of the fit of the nonparametric logistic regression model, we compared it to a simple linear logistic regression, both graphically, as summarized in Fig. 4, and quantitatively. For the quantitative comparison, we note that the mean area under the receiver operating characteristic (ROC) curve is 0.9675 for the simple linear logistic model, and 0.9697 for the nonparametric logistic model (P = 0.056, two-tailed paired *t*-test). We note that the nonparametric model provided a better fit by this measure for nine of the ten samples.

Finally, case-specific model fits were used to derive the mean probability of failure, 95% prediction interval and (0.90,0.95) lower tolerance limit on the initial misregistration, as summarized in Fig. 3 (right). The (0.90,0.95) lower tolerance limit estimate was 2.8 mm, i.e., with 95% confidence, the probability of failure for a future experiment is at most 10% for the initial misregistration less than or equal to 2.8 mm.

Cross-validation confirmed that for all leave-one-out experiments the probability of failure was below 10% for values of the initial displacement below the estimated LTL.

C. Example 3: Perturbed Initialization With Multiple Parameters

In this example, we demonstrate the use of LTL in an experimental setting investigating the effect of multiple input parameters on the registration performance. Similarly to the setup used in the previous illustration, the registration was applied to the rigidly aligned images perturbed with a synthetic initial transformation. However, this synthetic transformation utilized controlled sampling of multiple input parameters P_I that included the translation direction and rotation angle in addition to the translation magnitude. The translation was sampled randomly from 14 directions (six directions aligned with the image grid, and eight diagonal directions). The rotation angle was varied in the range from +6° to –6° in 3° increments separately for roll, pitch and yaw. The translation magnitude was initialized between 0 and 10 mm in increments of 2 mm. A registration experiment was considered successful (i.e., $F_C = 1$) when the maximum LRE was less than 3 mm.

Our approach does not allow estimation of multi-dimensional tolerance limits (i.e., we can only establish bounds on the most important parameter). Therefore, we considered a single parameter (translation magnitude) and modeled the probability of registration success. To account for the other varying parameters (rotation and translation direction), for each fixed value of the initial misalignment, we collected a random sample with replacement from all the experiments that were parameterized by that misalignment magnitude.

The results of LTL modeling for the 10 cases are shown in Fig. 5 (left). In this example the LTL is not defined, as for any value of the initial misalignment the proportion of the registration experiments with the successful outcome is expected to be less than 0.9. Upon further investigation of the experimental results, we determined that one of the 10 cases demonstrated a particularly high rate of failures. Fig. 5 (right) shows the modeled probability of failure after excluding this case from the analysis, leading to a valid value of the LTL. As can be observed from the plots, the mean success rate of the registration

remains largely unaffected between the modeling results based on 9 and 10 cases, while the LTL changes dramatically. We note that ideally when such behavior is observed in a registration evaluation, more data could potentially be collected, which might lead to a more practical LTL. We provide this example only as an illustration of the effect of the extreme behavior on the reported performance measure.

Leave-one-out cross-validation was not applied in this example. The estimated value of the LTL was less than 2 mm, while the values of the input parameters were sampled at 2 mm increments.

IV. Discussion and Conclusion

Our aim was to develop a methodology of tolerance limit estimation suitable for validation studies in image registration. We considered a setting where the measure to be characterized is the probability of registration success. The methodology we proposed can be used to establish a tolerance limit on the value of an input parameter used in the experimental evaluation. This leads to a measure that guarantees, with specified confidence, the success rate of the registration method for future samples. We applied the developed methodology in various experimental settings to summarize the results of an experimental evaluation for a registration method we are currently using in our MRI-guided prostate biopsy research trial [9]. As opposed to the commonly used statistics that aim to characterize the average performance of a method, the proposed approach estimates the probability of registration success for a future sample, including the effects of inter-sample variability, as a function of the initial misregistration. The ability to account for inter-sample variability in characterizing registration method probability of success, while providing a concise summary of the failure behavior, is the main advantage of the proposed methodology over the alternative approaches used in the literature.

To the best of our knowledge, the concepts of prediction intervals and tolerance limits have found limited use in image processing applications. Hernandez–Sabatè and Gil [32] used regression prediction intervals to compare output of an automated algorithm with ground truth. The methodology they proposed does not account for the inter-sample variability. In radiation oncology applications, tolerance and prediction intervals were used to summarize dose uncertainties and organ motion for optimized treatment delivery [33] and for dose planning [34]. We do not know of any previous application of tolerance limits to summarize experimental evaluation of image registration.

Several related efforts in the assessment and validation of image registration should be considered in the context of our work. Münzing *et al.* [35] developed a methodology that estimates registration quality for a specific registration result based on the features of the specific images being registered using a machine learning approach. Our methodology is suitable to bound the probability of the registration success (not registration error at a specific location) for a future experiment, based purely on the results of a prior experimental assessment. Quantification of registration uncertainty has been investigated by Kybic [3] and Risholm *et al.* [4]. The proposed approaches developed methodologies for estimating registration uncertainty in a specific pair of images being registered. Our approach is

conceptually different in bounding the probability of registration success for a future experiment, based on the accumulated results of the retrospective assessment. The work of Fitzpatrick *et al.* [1] is perhaps the closest to ours in estimating the distribution of TRE for a future experiment. In addition to the differences in the methodology, our approach provides estimate of registration success based on retrospective population studies, while the methodology of Fitzpatrick *et al.* provide TRE bounds for the specific case based on FiducialLocalizationError.Finally,theapproachofFitzpatrickis applicable to rigid registration only, while our approach provides bounds on population quantiles, conceptually being agnostic to the nature of the registration algorithm.

The underlying assumption of the tolerance limit procedure is that the logit-transformed probabilities of failure $f_i(x)$ are reasonably well approximated by a Gaussian random-effects model. If this assumption is satisfied, then the LTL will be a valid descriptor of the behavior of a future sample. Smaller sample size will result in tolerance limit being more sensitive to the assumptions being met exactly. Even if all of the assumptions are satisfied, tolerance limit estimate becomes more conservative as the sample becomes smaller, and may not be practically useful (e.g., may approach zero). The methodology we proposed does not allow the construction of multi-dimensional tolerance limits, i.e., we derive tolerance limit for one input parameter. As we demonstrate in Example 3, the experimental setup can include multiple parameters varied independently. We assumed that the registration failure rate is monotonically increasing with the larger values of the parameter under investigation. This assumption may not hold in all experimental scenarios.

Our analysis led to rather wide prediction intervals and low (i.e., comparable with the voxel resolution of the image data) values for the (0.90,0.95) LTL. These estimates might be overly conservative due to the limited data. For each value of the input parameter used in our experiments, we had only 10 measurements to estimate inter-sample variance. Considering the variability observed, a larger number of samples may be needed to have a more precise estimate. However, the purpose of the present work was to illustrate the application of tolerance limit, and not perform a comprehensive technical validation of the registration method. The small sample we used is suitable for the purposes of this demonstration. We emphasize that the proposed methodology should be considered within the larger context for conducting image technology validation studies [6], as a statistical index that can be used to summarize results of the experimental evaluation.

There are several venues that can be explored to improve validation of image registration tools using the proposed methodology. With a sufficiently large sample size and a clinically relevant experimental setup, practical values of LTL can be used as the ultimate measure of registration tool reliability for the clinical applications. The use of tolerance limits as a summary measure can also facilitate practical aspects of registration validation. Given a common annotated dataset (such as the one used in our evaluation [9], which is available publicly) various registration approaches can be compared using a simple summary measure that incorporates across-subject variability of the registration performance. The tolerance limit value can also be used as a reference for regression testing of the registration software. As an example, modern registration software tools are typically composed of multiple components and libraries. Changes or upgrades to one of the components (e.g., upgrade of

the Insight Toolkit [28] from ver. 3 to ver. 4) often lead to fluctuations in the registration result. The prediction interval curve can serve as a concise visualization summary to help determine whether software changes led to significant degradation of the registration tool performance.

Validation and performance characterization of image registration systems remain major obstacles that hinder translation of the registration tools into the clinic [36], [37]. To address these obstacles, methodologies and procedures need to be established to enable characterization of accuracy and reliability of the registration tools, as well as comparison of different approaches. Towards that goal, we introduced the concept of tolerance limits as a means to concisely summarize the results of the registration tool experimental evaluation. A key advantage of the proposed methodology lies in its ability to infer the expected performance of a registration method on a future sample. We believe there is a potential for this approach to be utilized widely for summarizing experimental evaluation of image analysis technology, leading to the development of improved strategies for validation of image registration tools. We hope that public availability of the scripts implementing the analysis summarized in this paper will facilitate adoption and further development of the proposed methodology.

Acknowledgments

The authors would like to thank the support team of Linux Cluster and especially J. Jackson of the Enterprise Research Infrastructure & Services (ERIS) group at Partners Healthcare for their in-depth support and for the provision of the high performance computing resources for conducting the registration experiments used in this paper.

This work was supported by the National Institutes of Health under Grant P41 EB015898, Grant R01 CA111288, Grant U01 CA151261, Grant P41 EB015902, and Grant U54 EB005149.

Appendix

In the following, we motivate the use of the root of (1) as the estimator of σ^2 and show that this function is monotone decreasing. Let $y = X\beta + \varepsilon$, where X is a known $n \times p$ regression matrix, β is an unknown vector of coefficients, and ε is a residual vector with mean 0 and covariance matrix V. Assume also that V is of the form

$$V = \sigma^2 I + V_0$$

where V_0 is a known matrix.

We propose estimating σ^2 by the root of the estimating equation

$$Q \equiv (y - \hat{y})^T V^{-1} (y - \hat{y}) - tr\,(I - H)$$

where $\hat{y}$ is the vector of generalized least squares fitted values

$$\hat{y} = X\left(X^T V^{-1} X\right)^{-1} X^T V^{-1} y \equiv Hy.$$

For the special case considered in this paper, $y_i = f_i(x)$, $X = J_n$ is a column vector of ones, $\beta = f(x)$ is a scalar, and

$$V = \sigma(x)^2 I + diag\left[\hat{\sigma}_1^2(x), \ldots, \hat{\sigma}_n^2(x)\right].$$

Hence

$$Hy = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} J_n = \tilde{f}(x) J_n.$$

We first motivate the estimating equation by showing that $E(Q) = 0$ when $\hat{\sigma}^2 = \sigma^2$

$$\begin{aligned}
E\left[Q\left(\sigma^2\right)\right] &= E\left[y^T (I-H)^T V^{-1} (I-H) y\right] - -tr(I-H) \\
&= tr\left[(I-H)^T V^{-1} (I-H) V\right] + +\beta^T X^T (I-H)^T V^{-1} (1-H) X\beta - -tr(I-H) \\
&= tr\left[(I-H)^T V^{-1} (I-H) V\right] + +\beta^T X^T (I-H)^T V^{-1} (X\beta - X\beta) - -tr(I-H) \\
&= tr\left[(I-H)^T \left(I - V^{-1} H V\right)\right] - tr(I-H) \\
&= tr\left[(I-H)^T \left(I - H^T\right)\right] - tr(I-H) \\
&= tr\left\{\left[(I-H)^2\right]\right\} - tr(I-H) \\
&= tr\left[(I-H)^T\right] - tr(I-H) \\
&= tr(I-H) - tr(I-H) = 0.
\end{aligned}$$

In the above calculations, we have made use of the well-known expression for the expectation of a general quadratic form, that is

$$E\left(y^T A y\right) = tr(A\Sigma) + \theta^T A \theta$$

where $E(y) = \theta = X\beta$ and $\mathrm{Var}(y) = \Sigma = V$, and $A = (I - H)^T V^{-1}(I - H)$ (e.g., see [38]). Also, recall that since H is a projection matrix, $H^2 = H$ and $HX = X$.

We now show that $Q(\sigma^2)$ is monotone decreasing. Because of the special form of V

$$\begin{aligned}
\frac{\partial V^{-1}}{\partial \sigma^2} &= \frac{\partial\left(V^{-1} V V^{-1}\right)}{\partial \sigma^2} \\
&= 2\frac{\partial V^{-1}}{\partial \sigma^2} + V^{-2}
\end{aligned}$$

hence

$$\frac{\partial V^{-1}}{\partial \sigma^2} = -V^{-2}.$$

Denote the generalized least squares residual vector by $r \equiv (I - H)y$. The derivative of Q with respect to σ^2 is

$$\frac{\partial Q}{\partial \sigma^2} = -r^T V^{-2} r = -\left\| V^{-1} r \right\|^2 < 0$$

since

$$\frac{\partial (I - H)}{\partial \sigma^2} = H V^{-1} (I - H)$$

and

$$H^T V^{-1} H = V^{-1} H = H^T V^{-1}.$$

REFERENCES

1. Fitzpatrick JM, West JB. The distribution of target registration error in rigid-body point-based registration. IEEE Trans. Med. Imag. Sep; 2001 20(9):917–927.
2. Crum WR, Griffin LD, Hill DLG, Hawkes DJ. Zen and the art of medical image registration: Correspondence, homology, and quality. NeuroImage. Nov; 2003 20(3):1425–1437. [PubMed: 14642457]
3. Kybic J. Bootstrap resampling for image registration uncertainty estimation without ground truth. IEEE Trans. Image Process. Jan; 2010 19(1):64–73. [PubMed: 19709978]
4. Risholm P, Janoos F, Norton I, Golby AJ, Wells WM. Bayesian characterization of uncertainty in intra-subject non-rigid registration. Med. Image Anal. Jul; 2013 17(5):538–555. [PubMed: 23602919]
5. Krishnamoorthy, K.; Mathew, T. Statistical Tolerance Regions: Theory, Applications, and Computation. Wiley; New York: 2009.
6. Jannin P, Grova C, Maurer CR. Model for defining and reporting reference-based validation protocols in medical image processing. Int. J. Comput. Assist. Radiol. Surg. Jul; 2006 1(2):63–73.
7. Hill DLG, Batchelor PG, Holden M, Hawkes DJ. Medical image registration. Phys. Med. Biol. 2001; 46:R1–R45. [PubMed: 11277237]
8. Mori, K.; Sakuma, I.; Sato, Y.; Barillot, C.; Navab, N., editors. Lecture Notes Comput. Sci. Springer; New York: 2013. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013.
9. Fedorov A, Tuncali K, Fennessy FM, Tokuda J, Hata N, Wells WM, Kikinis R, Tempany CM. Image registration for targeted MRI-guided transperineal prostate biopsy. J. Magn. Reson. Imag. May; 2012 36(4):987–992.
10. Shekhar R, Zagrodsky V. Mutual information-based rigid and nonrigid registration of ultrasound volumes. IEEE Trans. Med. Imag. Jan; 2002 21(1):9–22.
11. Ji S, Wu Z, Hartov A, Roberts DW, Paulsen KD. Mutual-information-based image to patient re-registration using intraoperative ultrasound in image-guided neurosurgery. Med. Phys. Oct; 2008 35(10):4612–4624. [PubMed: 18975707]
12. Sharma G, Mathew T. One-Sided and two-sided tolerance intervals in general mixed and random effects models using small-sample asymptotics. J. Am. Stat. Assoc. Mar; 2012 107(497):258–267.
13. Harris, EK.; Boyd, JC. Statistical Bases of Reference Values in Laboratory Medicine. Vol. 146. CRC; Boca Raton, FL: 1995.
14. Brown EB, Iyer HK, Wang CM. Tolerance intervals for assessing individual bioequivalence. Stat. Med. Apr; 1997 16(7):803–20. [PubMed: 9131767]

15. Myhre J, Jeske DR, Rennie M, Bi Y. Tolerance intervals in a heteroscedastic linear regression context with applications to aerospace equipment surveillance. Int. J. Quality, Stat., Reliabil. Feb. 2009 2009:1–8.
16. Rossiter WJ, Vangel MG, McKnight ME, Dewalt G. Spot test kits for detecting lead in household paint: A laboratory evaluation NIST. Tech. Rep. NISTIR. 6398:2000.
17. Rukhin AL, Vangel MG. Estimation of a common mean and weighted means statistics. J. Am. Stat. Assoc. Mar; 1998 93(441):303–308.
18. Rukhin AL, Biggerstaff BJ, Vangel MG. Restricted maximum likelihood estimation of a common mean and the Mandel-Paule algorithm. J. Stat. Plan. Inference. 2000; 83(2):319–330.
19. Cleveland WS. Robust locally weighted regression and smoothing scatterplots. J. Am. Stat. Assoc. Dec; 1979 74(368):829–829.
20. Fan, J.; Gijbels, I. Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability. Vol. 66. Taylor Francis; New York: 1996.
21. Loader, C. Local Regression and Likelihood. Springer; New York: 1999.
22. Paule RC, Mandel J. Consensus values and weighting factors. J. Res. Nat. Bureau Standards. 1982; 87(5):377–385.
23. R: A language and environment for statistical computing R Development Core Team. 2011
24. Tokuda J, Tuncali K, Iordachita I, Song S-E, Fedorov A, Oguro S, Lasso A, Fennessy FM, Tempany CM, Hata N. In-bore setup and software for 3T MRI-guided transperineal prostate biopsy. Phys. Med. Biol. Sep; 2012 57(18):5823–40. [PubMed: 22951350]
25. Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin J-CC, Pujol S, Bauer C, Jennings D, Fennessy F, Sonka M, Buatti J, Aylward S, Miller JV, Pieper S, Kikinis R. 3D Slicer as an image computing platform for the Quantitative Imaging Network. Magn. Reson. Imag. 2012; 30(9):1323–1341.
26. Johnson HJ, Harris G, Williams K. BRAINSFit: Mutual information registrations of whole-brain 3D images, using the Insight Toolkit. Insight J. Jul-Dec;2007
27. Pierson R, Johnson H, Harris G, Keefe H, Paulsen JS, Andreasen NC, Magnotta VA. Fully automated analysis using BRAINS: AutoWorkup. NeuroImage. Jan; 2011 54(1):328–36. [PubMed: 20600977]
28. Ibanez, L.; Schroeder, WJ. The ITK Software Guide. Kitware; Clifton Park, NY: 2003.
29. Nelson, WB. Accelerated Testing: Statistical Models, Test Plans, and Data Analysis, ser. Probabil. Stat. Wiley-Interscience; New York: 2004.
30. Littlewood B, Strigini L. Validation of ultrahigh dependability for software-based systems. Commun. ACM. Nov; 1993 36(11):69–80.
31. Zou KH, Warfield SK, Bharatha A, Tempany CMC, Kaus MR, Haker SJ, Wells WM, Jolesz FA, Kikinis R. Statistical validation of image segmentation quality based on a spatial overlap index. Acad. Radiol. Feb; 2004 11(2):178–89. [PubMed: 14974593]
32. Hernandez-Sabaté A, Gil D. Inferring the performance of medical imaging algorithms. Proc. 14th Int. Conf. Comput. Anal. Images Patterns. 2011:520–528.
33. Maleike D, Unkelbach J, Oelfke U. Simulation and visualization of dose uncertainties due to interfractional organ motion. Phys. Med. Biol. May; 2006 51(9):2237–52. [PubMed: 16625039]
34. Seidensticker M, Wust P, Rühl R, Mohnike K, Pech M, Wieners G, Gademann G, Ricke J. Safety margin in irradiation of colorectal liver metastases: Assessment of the control dose of micrometastases. Radiat. Oncol., London, U.K. Jan.2010 5:24–24.
35. Münzing SE, van Ginneken B, Murphy K, Pluim JPW. Supervised quality assessment of medical image registration: Application to intra-patient CT lung registration. Med. Image Analy. Dec; 2012 16(8):1521–31.
36. Slomka PJ, Baum RP. Multimodality image registration with software: State-of-the-art. Eur. J. Nucl. Med. Molecular Imag. Mar; 2009 36(Suppl. 1):S44–55.
37. Crum WR, Hartkens T, Hill DLG. Non-rigid image registration: Theory and practice. Br. J. Radiol. Dec; 2004 77(2):S140–S153. [PubMed: 15677356]
38. Seber, GAF.; Lee, AJ. Linear Regression Analysis, ser. Probabil. Stat. Wiley; New York: 2003.

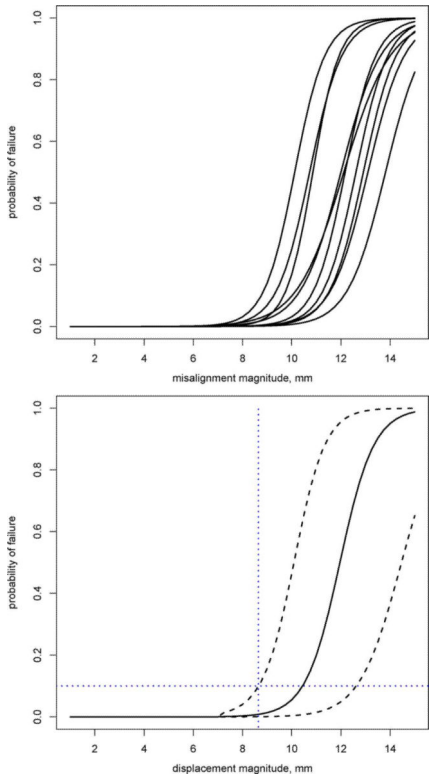

Fig. 1.

Estimation of tolerance limit for the synthetic deformation example in Example 1. Top: Result of the model fitting for the registration failure rate as a function of the initial misalignment. Each line corresponds to the modeled registration failure rate for an individual pair of images. Bottom: Mean failure rate with the 95% prediction interval. The (0.90,0.95) LTL corresponds to the intersection of the vertical blue line with the upper prediction interval and is equal to 8.7 mm.

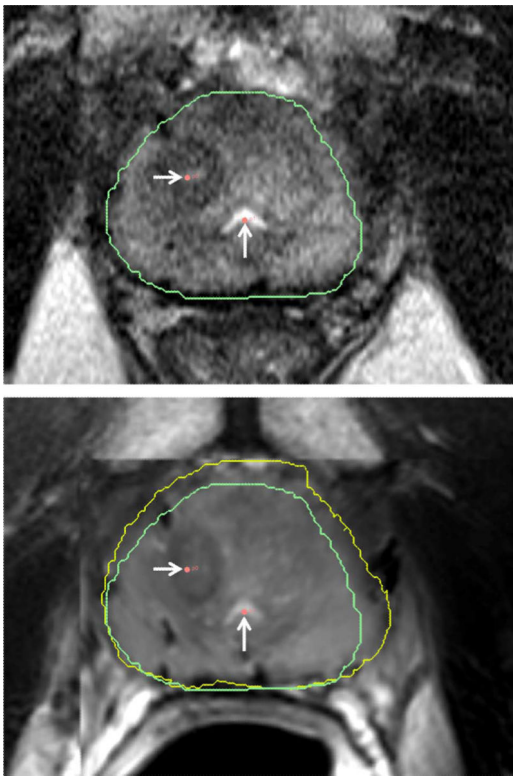

Fig. 2.

Example of the registration results for one of the cases used in the Example 2. Top: Axial slice of the intra-procedural T2w MRI with the contour of the capsule (green outline), and two of the fiducial points used in the evaluation (white arrows). The first fiducial (on the left) corresponds to the center of gravity for the segmentation of the dark round area. The second fiducial is at the corner formed by the ejuculatory ducts and the urethra. Bottom: Registered image, arrows point to the locations of the landmarks in the fixed image, which are close to the anatomical locations corresponding to the landmarks in the registered image.

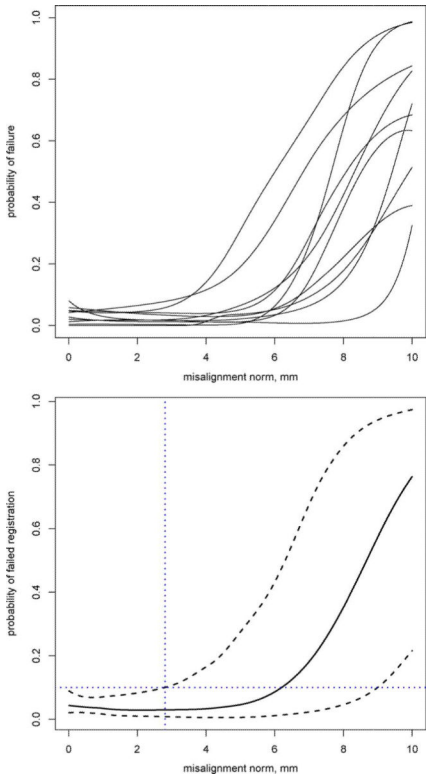

Fig. 3.

Estimation of tolerance limit for the registration experiment from Example 3. Top: Result of modeling for the estimated probability of failure as functions of misalignment norm. Each curve corresponds to a sample (pair of images being registered). As a result of this modeling, both intra- (estimated by the *locfit* procedure) and inter-sample variability can be estimated, as needed for the calculation of tolerance limits. Bottom: Average probability of failed registration with 95% prediction limit (dotted line) and (0.90,0.95) lower tolerance limit 2.8 mm (blue cross hairs).

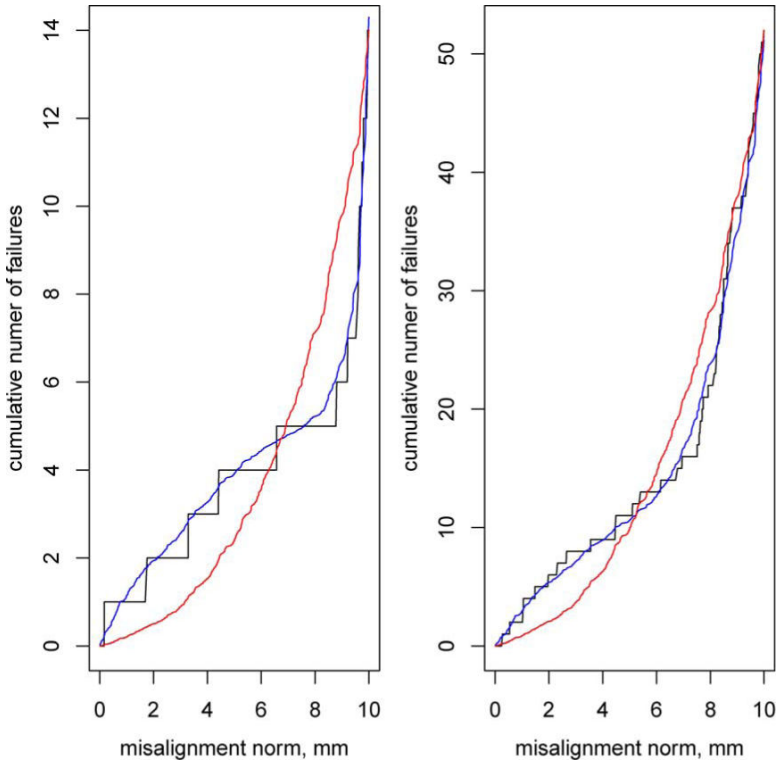

Fig. 4.

Cumulative number of failures versus misalignment norm for two representative samples from Example 2. The step function in black indicates the data, the simple linear logistic model fit is in red, and the nonparametric logistic regression fit is in blue. Plots correspond to sample 1 (left) and 8 (right). Improvement in the quality of fit relative to the linear logistic model fit was concluded based on the visual analysis and ROC quantitative assessment.

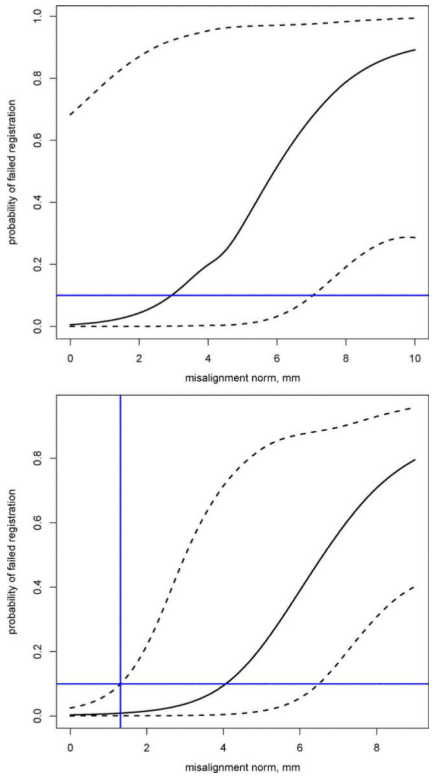

Fig. 5.

Modeling of the tolerance limit for the registration experiment with multiple input parameters being varied (Example 3). Mean rate of failure (solid line) with the 95% prediction interval (dashed line) and LTL corresponding to the intersection of the blue vertical line with the upper prediction interval line. Top: Results of modeling that take into account all 10 cases, LTL is undefined. Bottom: Modeling results after excluding the one case that exhibited very frequent failures. Inclusion of the case with frequent failures has dramatic effect on LTL, while the average failure rate remains largely unchanged (misalignment norm corresponding to the 10% average probability of success changes from 3 to 4 mm).