Registration Accuracy: How Good Is Good Enough? A Statistical Power Calculation Incorporating Image Registration Uncertainty

Eli Gibson[1], Aaron Fenster[1,2,3,4], and Aaron D. Ward[2,4]

[1] Robarts Research Institute, London, Canada
[2] Lawson Health Research Institute, London, Canada
[3] Department of Oncology, The University of Western Ontario, London, Canada
[4] Department of Medical Biophysics,
The University of Western Ontario, London, Canada

Abstract. Image registration is an important tool for imaging validation studies investigating the effect of underlying focal disease on the imaging signal. The strength of the conclusions drawn from these analyses is limited by statistical power. Based on the observation that in this context, statistical power depends in part on uncertainty arising from registration error, we derive a power calculation formula relating registration error, sample size, and the minimum detectable difference between normal and pathologic regions on imaging. Statistical mappings between target registration error and fractional overlap metrics are also derived, and Monte Carlo simulations are used to evaluate the derived models and test the strength of their assumptions.

Keywords: imaging validation, registration error, statistical power.

1 Introduction

Registration of medical images can enable complex analyses of medical data as well as image-guided diagnosis and treatment, provided the registration is performed with sufficient accuracy. There can be tradeoffs associated with achieving higher accuracy [1], including greater human interaction to guide registration algorithms to correct solutions, higher required image quality, and higher computational cost. Thus, for each study, it is important to identify the maximum acceptable level of registration error. This threshold is application-dependent [2], and establishing application-specific thresholds for maximum acceptable error has been identified as a key challenge in the field [1,2].

In image-guided interventions (IGI), registration can be used to guide a tool tip to a target region. Studies of such systems usually involve quantifying either the distance from the tool tip to the target or the overlap of the tool's treatment volume with the target. In some IGI applications (e.g. aortic aneurysms [3] and prostate cancer biopsy [4]), acceptable registration error thresholds have been identified for specific anatomy and imaging modalities. However, in some IGI

N. Ayache et al. (Eds.): MICCAI 2012, Part II, LNCS 7511, pp. 643–650, 2012.

contexts (e.g. keyhole brain surgery for ablation of epileptogenic foci [5]), it is currently unclear how to localize the targets on preoperative imaging. This motivates the need for studies to address this question.

In studies of the utility of medical imaging for disease localization (henceforth, *imaging validation studies*), registration can be used to align images of the medical imaging modality to be studied (*study images*) with *reference images* wherein ground truth regarding localized disease is defined (e.g. on pathologist-contoured digital pathology images). Such studies involve the measurement of the effect of the presence of disease *features of interest* (e.g. cancerous tissue) on image intensity (or other derived quantities) on the corresponding region of interest on the study image. Each ground truth delineation of disease features of interest on the reference image is mapped by an ideal (0 error) registration to a region of interest on the study image (denoted as R hereafter). We denote as $\tilde{R}$ such a region determined by a non-ideal (> 0 error) registration. Thus, in contrast to the IGI scenario, for imaging validation studies, the fidelity of the $R - \tilde{R}$ *regional overlap* is paramount. Mapping errors that result in smaller overlap may lead to larger required sample sizes to achieve a given minimum detectable difference (MDD) on imaging between pathologic and benign regions. This observation leads to three key questions affecting study design. **(1) What is the maximum acceptable registration error?** For a fixed sample size and desired MDD, what is the maximum acceptable image registration error? **(2) How many subjects are needed?** For a known image registration error and desired MDD, what is the required sample size? **(3) What is the minimum detectable difference?** For a fixed sample size and known image registration error, what is the MDD? To the best of our knowledge, there has been no previous work in the literature addressing these questions in the context of imaging validation studies based on image registration.

As a first step toward fully answering these questions, in this paper, we provide a derivation that yields the relationship between image registration error, sample size, and MDD, where image registration is used to determine whether the presence of particular anatomy, pathology or other features of interest in the underlying tissue is reflected in a change in the mean intensity of study image voxels containing the features of interest. The derivation of a statistical power calculation that incorporates uncertainty due to registration error yields a set of three equations that can be used to answer the three questions enumerated above. Statistical power is a measure that describes the probability of a study finding a statistically significant result when there is an underlying difference to be found. Thus, for studies to determine whether focal disease affects study image intensity, the acceptable registration error is defined relative to the study's statistical power. The statistical power is a relationship between the size of the study, the acceptable levels of type I error (false positive results from the study) and type II error (false negative results from the study), the intensity distributions in R and in the background, and the registration error. This statistical power is commonly expressed in the form of a sample size calculation that relates how many subjects to recruit for a particular study design or an MDD

calculation that relates how small a difference can be detected for a particular study design. To the best of our knowledge, this work represents the first derivation of a statistical power calculation for medical imaging validation studies that incorporates uncertainty in the overlap of R and $\tilde{R}$ due to registration error.

The remainder of this paper outlines the derivation of the relationship between registration error and statistical power for one study design (Section 2), describes simulations used to validate components of the derivation (Section 3), presents the results of these simulations (Section 4) and discusses the implications of these relationships (Section 5).

2 Derivation of MDD and Sample Size Relative to Registration Error

The statistical power calculation depends in part on the type of statistical analysis used in the study. In this paper, we addressed a specific analysis that statistically compares a pool of samples drawn from multiple identified $\tilde{R}$ to another pool drawn from background regions B using a t-test on the sample intensities. For imaging validation studies, image registration error is preferably measured as the target registration error (TRE), since a post-registration comparison of accurate segmentations is not feasible due to the absence of knowledge of the boundary of R on the study image. To derive the relationship between the TRE and inferential power, we utilize an intermediate metric, fractional overlap (FO), which is the ratio of the volume of the intersection of R and $\tilde{R}$ to the volume of $\tilde{R}$. The following two sections will derive the relationships between the MDD, sample size and the FO, and the relationship between TRE and FO.

2.1 Mapping Registration Error to Fractional Overlap

Fractional overlap ($\frac{|\tilde{R}\cap R|}{|\tilde{R}|}$) of two registered spherical regions R and $\tilde{R}$ can be expressed as a function of the radius of the regions r and the 3D registration error $\boldsymbol{x}$ between their centers: $f = \frac{\pi(4r+ry)(2r-ry)^2/12}{4\pi r^3/3} = (y^3 - 12y + 16)/16$ for $y \leq 2$ otherwise $f = 0$, where $y = ||\boldsymbol{x}||/r$ is the relative error.

The probability density function (PDF) F of FO can be derived as a function of the PDF Y of the relative error under certain assumptions: (1) each R and $\tilde{R}$ are spherical and of a fixed size, and (2) the registration error can be modeled as a 3D Gaussian. For $f = 0$, $p(F = 0) = p(Y >= 2)$. For $f > 0$, we use the relation for functions of random variables, $p(F = f) = p(Y = y(f))/|f'(y(f))|$. The derivative $\frac{df}{dy} = (3y^2 - 12)/16$. As $(y^3 - 12y + 16)/16 - f = 0$ has 3 real roots for $0 \leq f \leq 2$, we can express the inverse using the trigonometric expressions for cubic roots. There is one solution in the range $0 \leq y \leq 2$: $f^{-1}(y) = y(f) = 4cos(\frac{acos(1-f)+\pi}{3})$. Combining these intervals,

$$p(F = f) = \delta(f)p(Y >= 2) + \frac{16p(Y = 4cos(\frac{acos(1-f)+\pi}{3}))}{|3(4cos(\frac{acos(1-f)+\pi}{3}))^2 - 12|}. \quad (1)$$

For a registration error vector $\mathbf{x}$ that is distributed as a 3D Gaussian with component-wise variance a^2, (i.e. $\mathbf{X} \sim N_3(0, a^2 I)$), the registration error $x = ||\mathbf{x}||$ has a Maxwell-Boltzmann distribution. By a change of variables, $p(Y = y) = rp(X = yr)$, and, $p(Y > 2) = p(X > 2r)$, which can be substituted into Equation 1, yielding the FO as a function of registration error

$$p(F = f) = \delta(f)(1 - (erf(\frac{2r}{\sqrt{2}a}) - \sqrt{\frac{2}{\pi}}\frac{2rexp(-\frac{(2r)^2}{2a^2})}{a}))+$$
$$\frac{16r\sqrt{\frac{2}{\pi}}(4cos(\frac{acos(1-f)+\pi}{3})r)^2 exp(-(4cos(\frac{acos(1-f)+\pi}{3})r)^2/(2a^2))/a^3}{|3(4cos(\frac{acos(1-f)+\pi}{3}))^2 - 12|}. \quad (2)$$

For FO, the mean $\mu_F(\frac{a}{r}) = \int_0^1 fp(F = f)df$ and standard deviation $\sigma_F^2(\frac{a}{r}) = \int_0^1 (f - \mu_F(\frac{a}{r}))^2 p(F = f)df$ vary with the ratio of the TRE scaling parameter to the radius of R, and are invariant to specific choices of a and r. Integrating numerically yields the relationships shown in Figure 1.

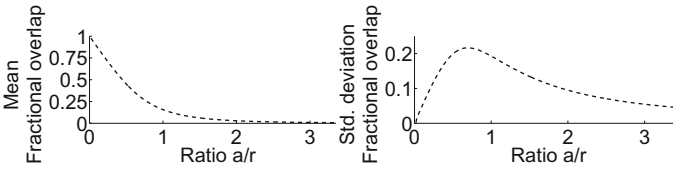

Fig. 1. Mean (left) and std. (right) fractional overlap as a function of the ratio a/r of the target registration error scaling factor to the radius of R.

2.2 Relationship between MDD, Sample Size and Fractional Overlap

The derivation is made under assumptions that (1) intensities of voxels containing the *features of interest* and background are independently distributed as $I_R \sim N(\mu_R, \sigma_R^2)$ and $I_B \sim N(\mu_B, \sigma_B^2)$, respectively; (2) statistical analysis will be performed by an unpaired two-sample heteroscedastic T-test of the null hypothesis that $\mu_R = \mu_B$ against the alternative hypothesis that $\mu_R \neq \mu_B$; (3) the number of measurements from each $\tilde{R}$ is constant across samples; (4) the number n of regions $\tilde{R}$ is large enough that the mean FO approximates a normal distribution (by the central limit theorem); and (5) the number of voxels v in each $\tilde{R}$ is large enough that discretizing error can be ignored.

When there is no registration error, the minimal detectable difference μ_d between μ_R and μ_B using a two sample t-test can be expressed as $\mu_d = T\sqrt{(\sigma_R^2 + \sigma_B^2)/(nv)}$, where T is a statistical threshold $t_{\alpha\{2\},nv} + t_{\beta\{1\},nv}$, where $t_{\alpha\{2\},nv}$ and $t_{\beta\{1\},nv}$ are taken from two- and one-tailed t-tables with nv degrees of freedom, constraining type I error to α and type II error to β.

When there is misregistration of the i-th region, the measurements in $\tilde{R}$ may contain samples from the background. Given FO f_i, the sample mean is

$\sum_i(\sum_{j=1}^{f_i v}(I_{j,i,R})+\sum_{j=1}^{(1-f_i)v}(I_{j,i,B}))/(vn)$. As each $I_{j,i,R}$ and $I_{j,i,B}$ is a Gaussian random variable, the distribution of the mean is:

$$N(\sum_i(f_i\mu_{i,R}+(1-f_i)\mu_{i,B})/n,(\sum_i(f_i\sigma_{i,R}^2+(1-f_i)\sigma_{i,B}^2)/(n^2v)), \quad (3)$$

or, by substituting $\mu_d=\mu_R-\mu_B$ and $\sigma_d^2=\sigma_R^2-\sigma_B^2$,

$$N(\mu_d\sum_i(f_i)/n+\mu_B,(\sigma_d^2\sum_i(f_i)/n+\sigma_B^2)/(nv)). \quad (4)$$

Because the FOs f_i are random variables contributing to both the mean and standard deviation of the distribution, the mean distribution is not Gaussian. To simplify the model, we introduce two approximations. First, we approximate $\sum_{i=1}^n(f_i)/n$ with a random variable $\sim N(\mu_F,\sigma_F^2/n)$ in the mean, using the central limit theorem approximation for sufficiently high n. Second, we approximate $\sum_i^n(f_i)/n$ as μ_F in the standard deviation. The resulting distribution $N(\mu_d N(\mu_F,\sigma_F^2/n)+\mu_B,(\mu_F\sigma_d^2+\sigma_B^2)/(nv))$ can be simplified to $N(\mu_d\mu_F+\mu_B,(\mu_F\sigma_d^2+\sigma_B^2)/(nv)+\mu_d^2\sigma_F^2/n)$.

Because this model for the distribution of the mean is Gaussian, as in the errorless case, we can incorporate this into the normal power analysis framework by constructing a hypothetical population that would have the same mean distribution: $N(\mu_d\mu_F+\mu_B,\mu_F\sigma_d^2+\sigma_B^2+\mu_d^2\sigma_F^2 v)$. The pooled variance for this analysis will be $(\sigma_B^2+\mu_F\sigma_d^2+\sigma_B^2+\mu_d^2\sigma_F^2 v)/2$ or, simplified, $\mu_F\sigma_d^2/2+\sigma_B^2+\sigma_F^2\mu_d^2 v/2$. The MDD between the $\tilde{R}$ and background can be expressed in terms of μ_d as

$$\mu_d\mu_F+\mu_B-\mu_B=\sqrt{\frac{\mu_F\sigma_d^2+2\sigma_B^2+\sigma_F^2\mu_d^2 v}{nv}}T. \quad (5)$$

Solving for mean FO μ_F yields

$$\mu_F=\frac{\sigma_d^2T^2\pm T\sqrt{\sigma_d^4T^2+8n\sigma_B^2\mu_d^2 v+4nv^2\mu_d^4\sigma_F^2}}{2\mu_d^2nv}. \quad (6)$$

Solving for the sample size yields

$$n=T^2\left(\frac{2\sigma_B^2+\mu_F\sigma_d^2}{\mu_d^2 v\mu_F^2}+\frac{\sigma_F^2}{\mu_F^2}\right). \quad (7)$$

Solving for the MDD yields

$$\mu_d=\sqrt{\frac{2\sigma_B^2+\mu_F\sigma_d^2}{nv(\mu_F^2-T^2\sigma_F^2/n)}}T. \quad (8)$$

3 Simulations

We performed Monte Carlo simulations to assess the accuracy of the derived statistical model, and the sensitivity of the model to assumption violations.

Table 1. Power simulation parameters. Values were specified as [a,b,c], ranges as [a..b], and values sampled once per $\tilde{R}$ from a normal distribution as N(mean, std.).

	$I_R - I_B$	n	v	a/r	α	β	σ_R^2	σ_B^2
S1	MDD	30	30	[0.1,0.5,1]	0.05	0.8	[1..100]	[1..100]
S2	MDD	[1..30]	30	1	0.05	0.8	10	10
S3	MDD	30	[1..30]	1	0.05	0.8	10	10
S4	MDD	30	N(30,[0..10])	1	0.05	0.8	10	10

To assess the model relating MDD, sample size and FO, we sampled N sets of image intensities from the background and $\tilde{R}$ intensity distributions and performed two sample T-tests of the null hypothesis that sample mean intensities were equal. In each simulation, $\mu_R - \mu_B$ was set to the MDD predicted by the model, and N=160,000 samples were taken (to yield a 95% confidence interval of width 0.5% on β). The proportion of positive t-test results from the simulation should match the model's type II error of $1 - \beta$. We assessed (S1) the accuracy of the model under the assumptions, as well as the sensitivity of the model to violations of the assumptions regarding (S2) the number of regions $\tilde{R}$ sampled, (S3) the number of voxels per $\tilde{R}$, and (S4) the constancy of the $\tilde{R}$ volume. The parameters varied in these simulations are described in Table 1.

To assess the model relating FO to registration error, we sampled error vectors $\boldsymbol{x}$ from a 3D Gaussian distribution and calculated the FO of R and $\tilde{R}$ with centers offset by $\boldsymbol{x}$. The resulting empirical PDF was compared to the PDF predicted by our model. We assessed (S5) the accuracy of the model under the given assumptions, with a ranging from $\frac{2}{100}$ to $\frac{350}{100}$ and $r = 10$.

4 Results

Simulation results from S1 through S4 are summarized in Fig. 2(a-d). The y-axes indicate the difference between the power predicted by the model and the simulations. A value of 0% indicates that the model exactly predicted the simulation results. Values of $-x$% and $+x$% indicate that the model under- and overestimated the power, respectively. Fig. 2(a) shows that when registration error is large (i.e. high a/r), the model underestimates power, particularly with small sample sizes. Fig. 2(b) shows that for small sample sizes, the model underestimates power, particularly with large registration errors. Fig. 2(c) shows that the model's estimate of power is reliable except in cases where small numbers (< 5) of point samples (e.g. voxels) are taken from each $\tilde{R}$. Fig. 2(d) shows that the model's estimate of power is accurate and robust to variance in the number of point samples taken from each $\tilde{R}$. In simulation S5, the model predicted the simulated mean and std. FO as a function of a/r (Fig. 1) to within 0.0006.

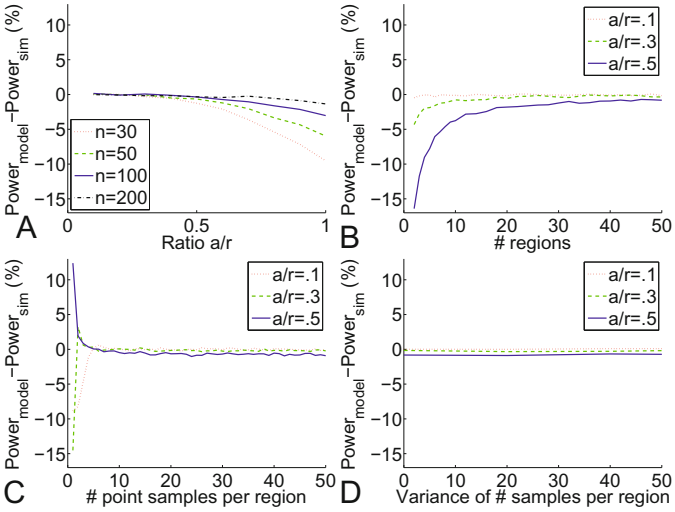

Fig. 2. (a) S1: Power vs. registration error a/r for several sample sizes. (b) S2: Power vs. sample size. (c) S3: Power vs. number of samples (e.g. voxels) / $\tilde{R}$. (d) S4: Power vs. variance in number of samples (e.g. voxels) / $\tilde{R}$.

5 Discussion

This work provides a derivation of a statistical power calculation incorporating image registration uncertainty and addressing three central questions in the design of imaging validation studies. **(1) Eq. (6):** What is the maximum acceptable registration error? **(2) Eq. (7):** How many subjects are needed? **(3) Eq. (8):** What is the MDD between normal and pathologic image regions? We derived the relationship between the scaling parameter of a 3D Gaussian TRE and the distribution of FO of spherical tumours. We also derived an approximate relationship between an arbitrary distribution of FO and the statistics of a study design. The combination of these derivations elucidated a relationship between registration error, sample size and statistical power, yielding a set of three equations that are central to the design of imaging validation studies.

These relationships could be used in several applications. During study design, Eq. (7) or (8) could be used to evaluate or control the power, after estimating imaging properties and registration errors, while Eq. (6) could be used to guide the selection of registration algorithms under the constraint of a study design. During algorithm development, Eq. (6) could be used to assess whether an algorithm has sufficient accuracy for a particular application.

We ran Monte Carlo simulations to test the fidelity of our model both when our assumptions were met and when some of them were relaxed. Our results showed that (1) the model predicts statistical power reliably for reasonable registration error (i.e. not larger than 50% of the radius of R) and the sample size > 30 (Fig. 2(a-b)); (2) the model predicts power reliably when > 5 samples (e.g. voxels) are obtained from within each $\tilde{R}$ (Fig. 2(c)); (3) the model predicts

power reliably regardless of the variability in the number of samples obtained from within each $\tilde{R}$ (Fig. 2(d)); and (4) the model accurately predicted the FO as a function of registration error.

The limitations of this work lie mainly in the strong assumptions made by the derivations. Although we have tested the robustness of the model to the relaxation of some of these assumptions, our testing in this regard is not exhaustive. Furthermore, extensions of these models may allow some assumptions to be relaxed (e.g., assumptions of spherical regions, isotropic 3D Gaussian registration error, and no correlation of voxels within each R). Also, our derivation is based on a relatively simple (albeit useful) statistical design; because analysis of statistical power depends on the statistical designs used, it would be valuable to extend the presented derivations to account for paired tests (to account for voxel intensity correlations within subjects), cluster randomization (to allow for intensity correlations within each R), regression (for longitudinal analyses) and multivariate data.

Acknowledgements. This work was supported by the National Sciences and Engineering Research Council of Canada, Cancer Care Ontario, the Ontario Institute for Cancer Research and the Canadian Institutes of Health Research [CTP 87515].

References

1. Simon, D., O'Toole, R.V., Blackwell, M., Morgan, F., Digioia, A.M., Kanade, T.: Accuracy validation in image-guided orthopaedic surgery. In: Proc. Int. Symp. Medical Robotics and Computer Assisted Surgery, pp. 185–192 (1995)
2. Loew, M.H., Rodriguez-Carranza, C.E.: Technical issues in multimodality medical image registration. In: Proc. IEEE Symp. Computer-Based Medical Systems, pp. 2–7 (1998)
3. Penney, G., Varnavas, A., Dastur, N., Carrell, T.: An Image-Guided Surgery System to Aid Endovascular Treatment of Complex Aortic Aneurysms: Description and Initial Clinical Experience. In: Taylor, R.H., Yang, G.-Z. (eds.) IPCAI 2011. LNCS, vol. 6689, pp. 13–24. Springer, Heidelberg (2011)
4. van de Ven, W.J.M., Litjens, G.J.S., Barentsz, J.O., Hambrock, T., Huisman, H.J.: Required Accuracy of MR-US Registration for Prostate Biopsies. In: Madabhushi, A., Dowling, J., Huisman, H.J., Barratt, D. (eds.) Prostate Cancer Imaging 2011. LNCS, vol. 6963, pp. 92–99. Springer, Heidelberg (2011)
5. Eriksson, S.H., Free, S.L., Thom, M., Harkness, W., Sisodiya, S.M., Duncan, J.S.: Reliable registration of preoperative MRI with histopathology after temporal lobe resections. Epilepsia 46(10), 1646–1653 (2005)