Statistical Power in Image Segmentation: Relating Sample Size to Reference Standard Quality

Eli Gibson[1,2], Henkjan J. Huisman[1], and Dean C. Barratt[2]

[1] Radboud University Medical Center, Nijmegen, The Netherlands
[2] University College London, London, UK

Abstract. Ideal reference standards for comparing segmentation algorithms balance trade-offs between the data set size, the costs of reference standard creation and the resulting accuracy. As reference standard quality impacts the likelihood of detecting significant improvements (i.e. the statistical power), we derived a sample size formula for segmentation accuracy comparison using an imperfect reference standard. We expressed this formula as a function of algorithm performance and reference standard quality (e.g. measured with a high quality reference standard on pilot data) to reveal the relationship between reference standard quality and statistical power, addressing key study design questions: (1) How many validation images are needed to compare segmentation algorithms? (2) How accurate should the reference standard be? The resulting formula predicted statistical power to within 2% of Monte Carlo simulations across a range of model parameters. A case study, using the PROMISE12 prostate segmentation data set, shows the practical use of the formula.

Keywords: Segmentation accuracy, statistical power, reference standard.

1 Introduction

Segmentation of anatomy and pathology on medical images plays a key role in many clinical scenarios, such as the delineation of the prostate to plan radiotherapy [2]. As a result, many algorithms for supporting or automating segmentation have been developed, and segmentation remains an active area of research [5].

Selecting reference standards (e.g. expert manual segmentations) to evaluate and compare segmentation algorithms involves balancing trade-offs between sample size, quality, and cost. An ideal reference standard would match the anatomy (or pathology) perfectly; however, anatomic/pathologic variation, ambiguous anatomical definitions, clinical constraints, and interobserver variability can introduce errors into the reference standard [8]. The quality and cost of the reference standard may be affected by the time and effort devoted to segmentation accuracy, the number of observers and the expertise of the observer(s). For example, the PROMISE12 prostate segmentation challenge [5] used two reference standards (see Fig. 1), a *high quality* one created by one experienced clinical reader and verified by another independent one, and a *low quality* one created

N. Navab et al. (Eds.): MICCAI 2015, Part III, LNCS 9351, pp. 105–113, 2015.
DOI: 10.1007/978-3-319-24574-4_13

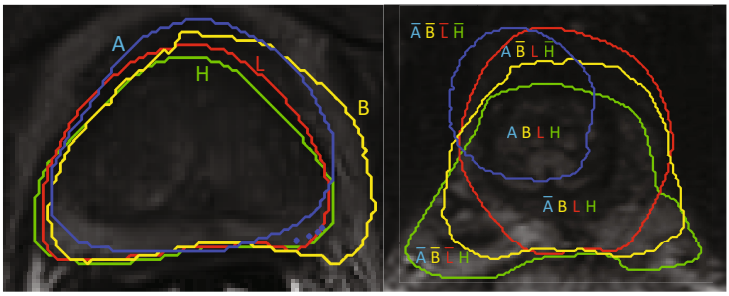

Fig. 1. Left: Prostate MRI segmentations by algorithms A (blue) and B (yellow), and low (L; red) and high (H; green) quality reference standards from the PROMISE12 data [5]. Relative to H, L oversegmented anteriorly, affecting accuracy measurements of A and B using L. Right: Apical segmentations showing regions of different segmentation outcomes ABLH (overbar denotes negative classifications). The statistical power of segmentation evaluation studies are modeled using outcome probability distributions.

by an inexperienced nonclinical observer. Due to the high costs of creating high quality reference standards, affordable lower quality ones are commonly used.

Reference standard errors can introduce uncertainty into performance measures, and impact the probability of detecting a significant difference (i.e. the statistical power) in validation studies [1]. Thus, there are trade-offs between generating large (and expensive) data sets to mitigate the uncertainty from imperfect reference standards, generating highly accurate reference standards (requiring substantial clinician time and expertise), and successfully finding significant differences. To balance these trade-offs, it is important to quantify the impact of reference standard quality on the statistical power of experiments comparing segmentation algorithm performance.

In the first steps towards this goal, we present the derivation of a new segmentation sample size formula that relates the statistical power to reference standard quality and algorithm performance measured with respect to a higher quality reference standard. After estimating the reference standard and algorithm performance (e.g. in a pilot study), this formula can inform key questions affecting study design: **(1) How many validation images are needed to evaluate a segmentation algorithm?** (i.e. given a reference standard with an estimated error rate, what is the sample size needed to show a clinically important improvement? **(2) How accurate does the reference standard need to be?** (i.e. given a data set of a fixed sample size, what level of reference standard accuracy must be attained to show a clinically important improvement)

2 Derivation of the Sample Size Formula

Since sample size formulae are analysis-specific, this paper focuses on one performance metric (differences in the mean segmentation accuracy between a pair of algorithms), and one statistical analysis comparing the performance of two

algorithms using a paired T test on the same data set of images. The sample size formula can then be expressed in a generic form as

$$N = \left(T_{\alpha/2}\sigma_0 + T_\beta \sigma_{Alt}\right)^2 / \delta_R^2, \tag{1}$$

where N is the number of images needed to detect a population difference in accuracy δ_R with respect to the reference standard R, σ_0^2 and σ_{Alt}^2 are the variances of the differences in accuracies under the null hypothesis ($\delta_R = 0$) and alternate hypothesis ($\delta_R \neq 0$), respectively, and $T_{\alpha/2}$ and T_β are $N-1$-degree-of-freedom Student T quantiles controlling type I and type II study error rates, respectively. A segmentation-specific sample size formula is derived in Section 2.1.

If the clinical goal requires true improvements in accuracy, these may be better reflected by specifying the minimal detectable difference δ_R with respect to the high quality reference standard, even if the actual study will use a lower quality reference study. In Section 2.2, this concept is used to relate the impact of reference standard quality on statistical power by expressing the sample size in terms of the performance of the algorithms and a low quality reference standard, measured against a higher quality reference standard (e.g. in a pilot study).

2.1 Sample Size for Segmentation Accuracy

We model the segmentation of an image as a set of binary classifications of n segmentation elements (such as voxels or superpixels). For each element, these classifications are modeled as independent samples from random variables representing the high (H) and low (L) quality reference standards and the algorithms (A and B). The classification outcome from all four is denoted $ABLH$ (see Fig. 1). One image event in a segmentation study can be represented as a scaled contingency table denoting the proportion of each type of classification outcome. If the outcome probabilities were fixed and known, this could be represented as a sample from a 16-element multinomial distribution with n trials, scaled by $\frac{1}{n}$. To model variability in the multinomial probability, the conjugate Dirichlet prior is commonly used [3], parameterized by the mean probability vector p and precision ω [7]. With this prior, the resulting image events are distributed as a 16-element Dirichlet-multinomial (Pólya) distribution P with n trials, scaled by $\frac{1}{n}$, with mean p and covariance $\frac{(n+\omega)}{n(\omega+1)}\left(diag(p) - p^T p\right)$. The differences in accuracy are then distributed as a linear transformation D of P, weighting outcomes where A outperforms B (event $C_A : A = L \neq B$) by 1, outcomes where B outperforms A (event $C_B : A \neq L = B$) by -1, and other outcomes ($A = B$) by 0. This distribution has a mean $\delta_L = p(C_A) - p(C_B)$ and a variance $\sigma_{Alt}^2 = \frac{(n+\omega)}{n(\omega+1)}(\psi - \delta_L^2)$, where $\psi = p(C_A) + p(C_B)$. Under the null hypothesis, $\delta_L = 0$, therefore $\sigma_0^2 = \frac{(n+\omega)}{n(\omega+1)}\psi$. Substituting σ_0 and σ_{Alt} into Eq. 1 and factoring out $\frac{(n+\omega)}{n(\omega+1)}$, the sample size for accuracy difference with respect to reference standard L is

$$N = \frac{(n+\omega)}{n(\omega+1)} \frac{\left(T_{\alpha/2}\sqrt{\psi} + T_\beta \sqrt{\psi - \delta_L^2}\right)^2}{\delta_L^2}. \tag{2}$$

2.2 Sample Size in Terms of the High Quality Reference Standard

Eq. 2 with δ_L measured with respect to the low quality reference standard can be expressed in terms of the performance of the algorithms and the low quality reference standard with respect to the high quality reference. As ψ can be expressed as $p(A\bar{B} \cup \bar{A}B)$ which is independent of the reference standard, only δ_L needs to be rewritten. For tractability, it is furthermore assumed that A, B and L are conditionally independent given H. For brevity, for $X \in \{A, B, L\}$, we denote conditional probabilities $p(X|H)$ using an overbar for $X = 0$, and an underline for $H = 0$: sensitivity $x = p(X = 1|H = 1)$, false negative rate $\bar{x} = p(X = 0|H = 1)$, false positive rate $\underline{x} = p(X = 1|H = 0)$, and specificity $\underline{\bar{x}} = p(X = 0|H = 0)$. Additionally, we use the following notation: $h = p(H = 1)$; and $\bar{h} = p(H = 0)$. Since outcomes where $A = B$ do not affect the *difference* in accuracy, δ_L is the probability of classification outcomes where $A = L$ and $B \neq L$ minus the probability of those where $A \neq L$ and $B = L$ (Eq. 3). By assuming conditional independence (Eq. 4), this can be rearranged algebraically (Eq. 5) to express δ_L in terms of the difference in accuracy ($\delta_a = (a-b)h + (\underline{\bar{a}} - \underline{\bar{b}})\bar{h}$) and sensitivity ($\delta_s = a - b$) with respect to the high quality reference, the sensitivity (l) and specificity ($\underline{\bar{l}}$) of the low quality reference standard, and the probability of positive outcomes (h) according to the high quality reference standard (Eq. 6). Placing this term in Eq. 2 yields the sample size formula in Eq. 7.

$$
\begin{aligned}
\delta_L = \; & p(A\bar{B}LH) + p(A\bar{B}L\bar{H}) + p(\bar{A}B\bar{L}H) + p(\bar{A}B\bar{L}\bar{H}) \\
& - p(\bar{A}BLH) - p(\bar{A}BL\bar{H}) - p(A\bar{B}\bar{L}H) - p(A\bar{B}\bar{L}\bar{H}) && (3) \\
= \; & a\bar{b}lh + \underline{a\bar{b}l}\bar{h} + \bar{a}b\bar{l}h + \underline{\bar{a}b\bar{l}}\bar{h} - \bar{a}blh - \underline{\bar{a}bl}\bar{h} - a\bar{b}\bar{l}h - \underline{a\bar{b}\bar{l}}\bar{h} && (4) \\
= \; & \left((a-b)h + (\underline{\bar{a}} - \underline{\bar{b}})\bar{h}\right)(2\underline{\bar{l}} - 1) - 2\,(a-b)\,(\underline{\bar{l}} - l)h && (5) \\
= \; & \delta_a(2\underline{\bar{l}} - 1) - 2\delta_s(\underline{\bar{l}} - l)h. && (6)
\end{aligned}
$$

$$
N = \frac{(n+\omega)}{n(\omega+1)} \frac{\left(T_{\alpha/2}\sqrt{\psi} + T_\beta \sqrt{\psi - \left(\delta_a(2\underline{\bar{l}} - 1) - 2\delta_s(\underline{\bar{l}} - l)h\right)^2}\right)^2}{\left(\delta_a(2\underline{\bar{l}} - 1) - 2\delta_s(\underline{\bar{l}} - l)h\right)^2}. \qquad (7)
$$

3 Simulations

We performed Monte Carlo simulations to assess the accuracy of the sample size formula. In each simulation, we instantiated a parametric model (described below) representing a segmentation validation experiment with an underlying difference in accuracy, and repeatedly simulated the experiment to estimate the simulated power (i.e. the proportion of simulations yielding true positive outcomes) and compared it to the specified power. To exclude error due to using $\lceil N \rceil$ instead of N (because N must be a natural number), we determined $\lceil N \rceil$ using a specified power $1-\beta = 0.8$, but compared the resulting power to $1-\beta_{\lceil N \rceil}$

Table 1. Parameters used to estimate the accuracy of the model

	n	h	A_a	δ_a	a/A_a	δ_s	l	$\underline{l}$	ω
Baseline	10000	0.4	0.8	0.05	1	0.05	0.8	0.8	100
Minimum	100	0.1	0.6	0.01	0.75	0.01	0.6	0.6	16
Maximum	100000	0.9	0.99	0.25	1.25	0 .30	1	1	1024

computed by solving Eq. 7 for $N = \lceil N \rceil$. We used 25,000 repetitions, yielding a 1% wide 95% confidence interval on the error in predicted power.

To assess the accuracy of Eq. 7, we used a parametric model with random variables A, B, L and H, for the two algorithms and the low and high quality reference standards. A, B and L could be defined by mean sensitivities and specificities with respect to H; however, to independently manipulate δ_a and δ_s in Eq. 7, A and B were redefined in terms of δ_a, δ_s, the accuracy of A (A_a) and a sensitivity factor a/A_a. H was parameterized by a mean probability of positive outcomes, and parameter n specified the number of segmentation elements. Type I and II error rates were fixed to be 0.05 and 0.2, respectively. A precision parameter ω was used to model inter-image variability and variability in positive outcomes. The segmentation outcomes were sampled from a Dirichlet-Multinomial distribution parameterized by ω and $p = p(ABLH) = p(A|H)p(B|H)p(L|H)p(H)$. Model parameters were initialized with baseline values given in the first row of Table 1, and were varied independently through the ranges given in rows 2-3.

4 Results

The error in the power predicted by the model over a range of parameters are shown in Fig. 4; the 95% confidence interval bounds on prediction error were within 2% throughout. To ensure high sensitivity to all parameter values where the model has prediction error, multiple comparison correction, which would widen the intervals and hide model errors, was not used. Thus, the 95% confidence intervals for a perfect model would include 0% error for 95% of parameter sets; for our model, the confidence intervals included 0% error for 89% of the parameter sets. Three regions showed notable deviation from the simulations: high accuracy ($A_a > 0.975$), large differences in accuracy ($\delta_a \geq 0.2$) and high precision ($\omega = 1024$), although these errors did not exceed 2% (95% confidence).

5 Case Study

Using data from the PROMISE12 prostate MRI segmentation challenge [5], this case study demonstrates how to apply Eq. 7. In this data set, two experienced clinicians generated a high quality reference standard, and a graduate student generated a low quality reference standard. Although, in the challenge, algorithms were compared to the high quality reference standard, this case study considers comparing segmentation algorithms using the graduate-student reference

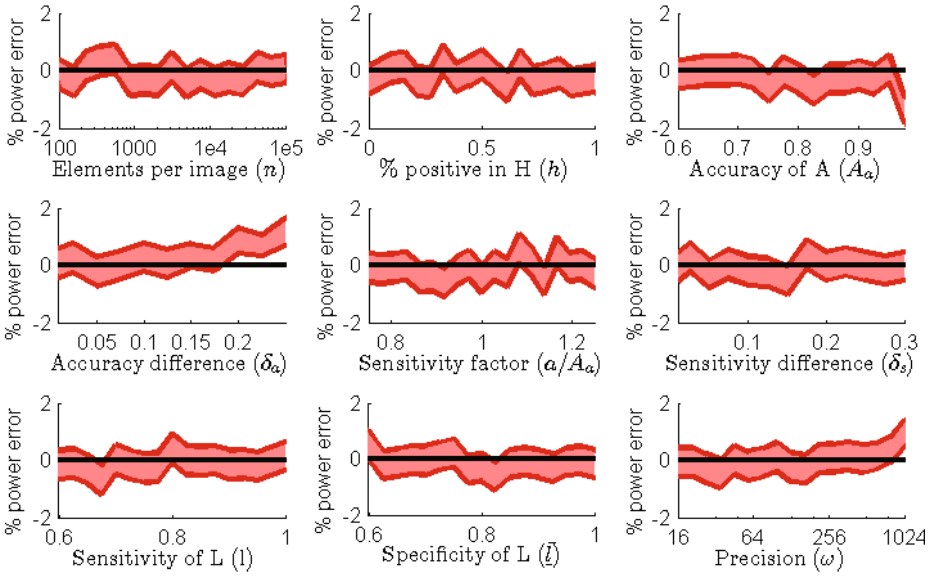

Fig. 2. Model accuracy: 95% confidence regions on the power prediction error (%). Values where the region does not contain the 0% error line suggest prediction error.

standard. To apply Eq. 7, one must estimate, from literature or pilot data, the algorithm and reference performance with respect to the high quality reference standard (a, b, l, $\underline{a}$, $\underline{b}$, and $\underline{l}$), the probability of positive outcomes (h), the precision ω, the probability of disagreement between A and B (ψ), and the desired (or observed) performance differences (δ_L, δ_a and δ_s). In this case study, image events (i.e. 16-element contingency tables) were computed for 30 cases from the segmentations of the high and low quality reference standards and two algorithms submitted for the challenge. If such a data set is not available, a high quality reference standard on a small pilot data set could be generated to make these estimates. Precision was estimated by setting $\frac{n+\omega}{n(\omega+1)}$ to $s^2/(D^T(diag(\tilde{p})-\tilde{p}^T\tilde{p})D)$, where s^2 was the observed variance in accuracy difference and $\tilde{p}$ was the vector of observed probabilities: $\omega_L = 1600$ (low quality reference standard), and $\omega_H = 1900$ (high quality). Other parameters were estimated by combining the counts of classification outcomes (e.g. $\underline{a} = p(\bar{A}|\bar{H}) = \frac{||\bar{A}\cap\bar{H}||}{||\bar{H}||}$) and averaging over the images: $a = 0.892$, $b = 0.899$, $l = 0.892$, $\underline{a} = 0.998$, $\underline{b} = 0.997$, and $\underline{l} = 0.999$, $h = 0.171$, $\delta_L = 0.0016$, $\delta_a = 0.0016$, $\delta_s = -0.0066$, $\psi = 0.0047$. Substituting ψ, δ_a and ω_T into Eq. 2 yielded a sample size of $N = 7.2$ to detect a difference of δ_a using the high quality reference standard. Substituting the parameters (except δ_L) into Eq. 7 yielded $N = 8.4$ to detect a difference of δ_a (as measured with a high quality reference standard) using the low quality reference standard. For comparison, $N = 8.5$ when substituting ψ, δ_L and ω_T into Eq. 2 directly, suggesting that assuming conditional independence introduced minimal error.

In this case study, the low quality reference standard required only a slightly larger sample size, and could be a suitable approach.

6 Discussion

This paper derived a sample size formula for comparisons of segmentation accuracy between two algorithms, and expressed it as a function of performance of the algorithms and a low quality reference standard performance measured against a higher quality reference standard. This relationship can be used to address central questions in the design of segmentation validation experiments: **(1) How many validation images are needed to evaluate a segmentation algorithm? (2) How accurate does my reference standard need to be?** The relationship can be used in several contexts. For researchers designing novel segmentation algorithms, the relationship can inform the selection of validation data sets and reference standards needed to evaluate the algorithms. For researchers involved in creating validation data sets and reference standard segmentations, the relationship can guide the trade-offs between the costs of generating large data sets and those of generating highly accurate reference standards.

While this paper considers the common approach of using low quality reference standards directly, other approaches have been proposed to leverage lower quality reference standards. Most notably, label fusion algorithms, such as STAPLE [8] aim to infer a high quality reference standard from multiple low quality reference standards. This has even been extended to use *crowd-sourced* segmentations by minimally trained users [4]. These methods may be preferable, when feasible, to using low quality reference standards; however, the need to create multiple reference standards may increase the cost/complexity of such studies.

The predicted power was within 2% of the simulations over the tested parameters. Three conditions showed measurable deviations from 0% error. With high accuracies, the algorithms disagreed on few classifications and accuracy differences were not normally distributed as assumed by the T-test; this was compounded by the low sample size ($N = 11$) where the T-test is more sensitive to assumption violations [6]. Thus, statistical comparisons of highly accurate algorithms may be challenging. Large or very consistent (high ω) accuracy differences, yielded even lower sample sizes ($N \leq 6$). Low predicted sample sizes may have higher error, although, even in these cases, it did not exceed 2%.

Sample size formulae are inherently specific to the statistical analysis being performed. The presented formula is specific to studies comparing the accuracy of two algorithms using one reference standard. As illustrated by the PROMISE12 challenge, many segmentation evaluation studies compare multiple measures (e.g. Dice coefficients and boundary distances) between >2 algorithms using multiple reference standards. Deriving analogous sample size formulae for these studies would be a valuable direction for future work.

Two key derivational assumptions may constrain the use of the formula. First, we assumed that given the high quality reference standard outcome, the low quality reference standard and algorithm segmentations are conditionally independent (i.e. do not make the same error more than predicted by chance). In

practice, segmentation elements with confounding image features (e.g. low contrast or artifacts) may induce similar errors in the segmentations, potentially violating conditional independence. In the case study, any such correlation did not substantially impact the calculation; however, other data sets may be more prone to violations of this assumption. Additionally, segmentation algorithms trained using the low quality reference standard may make the same types of error as the reference standard potentially violating conditional independence. This was not a factor in the PROMISE12 data set, as algorithms were trained using the high quality reference standard. Using pilot data to test for conditional independence or to evaluate the impact of such correlation (as in the case study) may identify such situations. Second, we modelled segmentation as a set of independent decisions on segmentation elements, such as voxels or superpixels. In practice, regularization (e.g. enforcing smooth segmentations), clinical knowledge (e.g. anatomical constraints) or image features (e.g. artifacts) may cause correlated segmentation outcomes. It is unclear to what extent the aggregation of these outcomes in the multinomial and the variance in the Dirichlet prior mitigate violations of this assumption. Characterizing the sensitivity of the model to violations of these assumptions would be a valuable direction for future work.

In conclusion, this paper derived a sample size formula for comparing the accuracy of two segmentation algorithms using an imperfect reference standard, expressed as a function of algorithm and reference standard performance (measured against a higher quality reference standard). The model was accurate to within 2% across the tested range of model parameters, although it began to deviate measurably from simulations when N was low. We also showed a case study where using a low quality reference standard could cause little increase in sample size, and where assuming conditional independence for the algorithms and low quality reference standard introduced little error. The Medical Research Council and the Canadian Institutes of Health Research supported this work.

References

1. Beiden, S.V., Campbell, G., Meier, K.L., Wagner, R.F.: The problem of ROC analysis without truth: The EM algorithm and the information matrix. In: SPIE Medical Imaging, pp. 126–134 (2000)
2. Boehmer, D., Maingon, P., Poortmans, P., Baron, M.H., Miralbell, R., Remouchamps, V., Scrase, C., Bossi, A., Bolla, M.: Guidelines for primary radiotherapy of patients with prostate cancer. Radiother. Oncol. 79(3), 259–269 (2006)
3. Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B.: Bayesian data analysis, vol. 2. Taylor & Francis (2014)
4. Landman, B.A., Asman, A.J., Scoggins, A.G., Bogovic, J.A., Stein, J.A., Prince, J.L.: Foibles, follies, and fusion: Web-based collaboration for medical image labeling. NeuroImage 59(1), 530–539 (2012)

5. Litjens, G., Toth, R., van de Ven, W., Hoeks, C., Kerkstra, S., van Ginneken, B., Vincent, G., Guillard, G., et al.: Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge. Med. Image Anal. 18(2), 359–373 (2014)
6. Lumley, T., Diehr, P., Emerson, S., Chen, L.: The importance of the normality assumption in public health data sets. Ann. Rev. Pub. Health 23(1), 151–169 (2002)
7. Minka, T.P.: Estimating a Dirichlet distribution. Tech. rep., M.I.T. (2000)
8. Warfield, S.K., Zou, K.H., Wells, W.M.: Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans. Med. Imag. 23(7), 903–921 (2004)