Inferring the Performance of Medical Imaging Algorithms

Aura Hernàndez-Sabaté[1], Debora Gil[1], David Roche[2],
Monica M.S. Matsumoto[3], and Sergio S. Furuie[3]

[1] Computer Science Department and Computer Vision Center - UAB, Bellaterra, Spain
[2] Laboratory of Systems Pharmacology and Bioinformatics, UAB, Bellaterra, Spain
[3] Faculdade de Medicina and Escola Politécnica da USP, São Paulo, Brazil
{aura,debora}@cvc.uab.es

Abstract. Evaluation of the performance and limitations of medical imaging algorithms is essential to estimate their impact in social, economic or clinical aspects. However, validation of medical imaging techniques is a challenging task due to the variety of imaging and clinical problems involved, as well as, the difficulties for systematically extracting a reliable solely ground truth. Although specific validation protocols are reported in any medical imaging paper, there are still two major concerns: definition of standardized methodologies transversal to all problems and generalization of conclusions to the whole clinical data set.

We claim that both issues would be fully solved if we had a statistical model relating ground truth and the output of computational imaging techniques. Such a statistical model could conclude to what extent the algorithm behaves like the ground truth from the analysis of a sampling of the validation data set. We present a statistical inference framework reporting the agreement and describing the relationship of two quantities. We show its transversality by applying it to validation of two different tasks: contour segmentation and landmark correspondence.

Keywords: Validation, Statistical Inference, Medical Imaging Algorithms.

1 Introduction

Researchers agree that validation of medical imaging algorithms is essential for supporting their validity and applicability in clinical practice [1]. Although validation is addressed in any medical imaging paper, there is no consensus in the statistical and mathematical tools required for standardized quantitative analysis [2,3,1,4]. Given the diversity of imaging tasks and final clinical applications, techniques are prone to be validated using specific protocols, not easily extendable to a unifying general framework [1].

A validation protocol should face two main challenges: extracting ground truth (GT) and defining a metric quantifying differences between GT and the algorithm output (AO). A main difficulty in medical imaging is that GT might not be always available or might vary across observers [5]. The first case is common in image registration tasks, since the deformation matching images might not be easily extracted from in vivo cases. Current solutions, base validation on either synthetic experiments or correspondence of

A. Berciano et al. (Eds.): CAIP 2011, LNCS 6854, pp. 520–528, 2011.

anatomical landmarks [6]. The realism of synthetic databases might be too low for generalization of conclusions to clinical data [5]. It follows that, in real data, a verification based on structures (landmarks) correspondence is usually required. Variability in GT typically arises in segmentation tasks, due to discrepancies across manual tracers. This implies that an analysis of automated errors might not reflect, by its own, the true accuracy of segmentations, since variations might be caused by a significant difference among expert models. A standard solution [7] is comparing automated errors to the variability among different manual segmentations. Concerning comparison between GT and automatic computations, several metrics can be considered. For landmark correspondence, the difference in positions is the accepted goodness measure [6], while for contour segmentation [8, 3] differences can be measured by means of area overlap or distances between contours. The counterpart of these metrics is that they assess complementary quality scores and, thus, several quality measures need to be considered.

Two major concerns still remain: 1) defining the subset of scores best reflecting accuracy for clinical application and 2) whether the results of validation tests are generalizable to all clinical data. We claim that a validation protocol assessing to what extent an image processing algorithm can substitute the manual interaction would address both issues. In this context, validation should report the agreement between AO and GT, as well as, a model describing the relation between both quantities.

Agreement between observers can be assessed by means of Bland-Altman plots or regression analysis. Bland-Altman [9] measures this agreement by analyzing the variability of their differences. In the case of disagreement, Bland-Altman fails to either describe or report the degree of disagreement [10]. Regression analysis provides a (linear) statistical model of the relation between two quantities. Existing techniques usually only report regression coefficients (slope and intercept) and correlation. Given that correlation only reports the degree of linear dependence between both quantities, a high correlation does not imply that the variables agree [10]. In order to explore agreement, one should consider the slope and intercept of the regression model, since they describe the relation between the two variables. However, even in the case of a perfect relation (identity), the regression coefficients alone are not sufficient to ensure that the quantities can be swapped. The slope and intercept describe the behavior of the specific sample we are analyzing, but they do not allow to generalize conclusions to the whole population. The only way to obtain generalizable conclusions is by means of statistical inference.

We present a statistical inference framework for assessing how well two methodologies performing the same task behave equally and can replace one each other. We define a regression model for predicting the performance of an image processing algorithm in clinical data from a subset of validated samples. Our model is applied to two main tasks involved in medical image processing: detection (registration) and segmentation of anatomical structures. Experiments on vessel wall segmentation show the correlation between our model and standard metrics. Meanwhile, experiments on cardiac-phase detection illustrate its versatility for assessing difficult tasks.

2 Inference Model

We note by GT the ground truth we want to substitute and AO, the algorithm output. Their nature is prone to vary depending on the particular problem we are facing:

1. **Segmentation.** Image segmentations produce a (continuous) contour enclosing the area of interest. Therefore, $GT = GT(t) = (GT_1(t), GT_2(t))$, $AO = AO(t) = (AO_1(t), AO_2(t))$ are curves parameterized by a common parameter $t \in [0, 1]$.
2. **Detection.** In detection tasks, the output is a (finite) list storing the positions of k corresponding landmarks. Thus, $GT = \{GT^i\}_{i=1}^k$, $AO = \{AO^i\}_{i=1}^k$, for $GT^i = (GT_j^i)_{j=1}^n$, $AO^i = (AO_j^i)_{j=1}^n$ points in $\mathbb{R}^n$, where n=1,2,3 is the dimension of the data domain.

Our final goal is to control (predict) the values taken by GT from the values taken by the alternative measure AO. In inference statistics, this is achieved by relating both quantities using a regression model.

2.1 Regression Model

The linear regression of a response variable y over an explicative variable x is given by:

$$Y = X\beta + \epsilon \tag{1}$$

for $\beta = (\beta_0, \beta_1)$ the regression parameters, $X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix}$, $Y = (y_1, \cdots, y_N)$ a sampling of x, y and $\epsilon = (\varepsilon_1, \cdots, \varepsilon_N)$ an uncorrelated random error following a multivariate normal distribution, $N(0, \Sigma^2)$ of zero mean and variance $\Sigma^2 = \sigma^2 Id$.

The parameters of the regression model (1) are the regression coefficients $\beta = (\beta_0, \beta_1)$ and the error variance σ^2. The regression coefficients describe the way the two variables relate, while the variance indicates the accuracy of the model and, thus, measures to what extent x can predict y.

Given that, in our case, the inference is over GT, our model is:

$$GT_i = \beta_0 + \beta_1 AO_i + \varepsilon_i \tag{2}$$

for $(GT_i)_{i=1}^N$, $(AO_i)_{i=1}^N$ samplings of GT and AO obtained for each task as:

1. **Segmentation.** In the case of contours, the sampling is given by the coordinates of a uniform sampling of each of the curves:

   $$\begin{array}{l} (GT_i)_{i=1}^{2N} = (GT(t_i))_{i=1}^N = (GT_1(t_i), GT_2(t_i))_{i=1}^N \\ (AO_i)_{i=1}^{2N} = (AO(t_i))_{i=1}^N = (AO_1(t_i), AO_2(t_i))_{i=1}^N \end{array}$$

   for $t_i = i/N, i = 1 : N$. In order to have pair-wise data, samplings are taken using a common origin of coordinates.
2. **Detection.** In this case, the sampling of the two variables is given by:

   $$\begin{array}{l} (GT_i)_{i=1}^{N=nk} = (GT^i)_{i=1}^k = ((GT_j^i)_{j=1}^n)_{i=1}^k \\ (AO_i)_{i=1}^{N=nk} = (AO^i)_{i=1}^k = ((AO_j^i)_{j=1}^n)_{i=1}^k \end{array}$$

   Pair-wise data is obtained by using the same scanning direction in images for sorting the vector of landmarks.

For a sample of length N, the regression coefficients, $\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1)$, are estimated by least squares fitting as:

$$\widehat{\beta} = (X^T X)^{-1} X^T Y \tag{3}$$

for X and Y as in eq. (1) and ^{T} denoting the transpose of a matrix.

The difference between the estimated response, $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$, and the observed response y_i, $e_i = y_i - \widehat{y}_i$, are called residuals. Their square sum provides an estimation of the error variance:

$$S_R = \widehat{\sigma}^2 = \frac{\sum e_i^2}{n-2}$$

Previous to any kind of inference, it is mandatory to verify that the estimated parameters make sense. That is, whether it really exists a linear relation between x and y. By the Gauss-Markov theorem, such linear relation can be statistically checked using the following F-test [11]:

$$TM: \; H_0 : \beta_1 = 0 \; , \, H_1 : \beta_1 \neq 0 \tag{4}$$

where a $p - value$ close to zero (below α) ensures the validity of the linear model with a confidence $(1 - \alpha)100\%$.

2.2 Prediction Model

In order to predict the values of GT from the values achieved by AO, we use the regression prediction intervals [11]:

$$PI(x_0) = [L_{PI}(x_0), U_{PI}(x_0)]$$

since, for each $x = x_0$, they provide ranges for y at a given confidence level $1 - \alpha$. That is, given x_0, the values of the response y are within $L_{PI}(x_0) \leq y \leq U_{PI}(x_0)$ in $(1 - \alpha)100\%$ of the cases.

Given $x_0 = AO_0$, the confidence interval at a confidence level (1-α) predicting GT is given by:

$$PI(x_0) = [L_{PI}(x_0), U_{PI}(x_0)] = [\widehat{y_0} + t_{\alpha/2} S_R \sqrt{1 + h_0}, \widehat{y_0} - t_{\alpha/2} S_R \sqrt{1 + h_0}]$$

for $t_{\alpha/2}$ the value of a T-Student distribution with $N - 2$ degrees of freedom having a cumulative probability equal to $\alpha/2$ and $h_0 = (1 \quad x_0)(X^T X)^{-1}(1 \quad x_0)^T = a_0 + a_1 x_0 + a_2 x_0^2$. Prediction intervals achieve their minimum range at the average x and their maximum range at their extreme values xMin, xMax.

A prediction interval within a given precision, $U_{PI}(x) - L_{PI}(x) \leq \epsilon, \forall \, x$, indicates that the regression model predicts GT with high accuracy and, thus, AO is a good candidate for substituting GT. The alternative quantity can substitute GT in the measure that the identity line is within the range given by the prediction interval $PI(x)$. Otherwise, AO presents a systematic bias from the reference, which might be corrected using the regression coefficients. The slope, β_1, is associated to a scaling factor (unit change), while the intercept, β_0, is a constant bias.

The identity line is in the range of the prediction interval $PI(x)$ with a given precision, ϵ, if and only if $(PI(x) - x) \subset (-\epsilon, \epsilon)$, $\forall\ x$. This requirement is fulfilled if the following conditions hold:

$$\begin{array}{l} CP_1:\ max(L_{PI}(x) - x) \leq 0 \leq min(U_{PI}(x) - x) \\ CP_2:\ max(U_{PI}(x) - L_{PI}(x)) \leq 2\epsilon \end{array} \tag{5}$$

The first condition ensures that variables can be swapped with a confidence of $(1 - \alpha)$, while the second assesses the accuracy of the swapping. We note that the above conditions can also be formulated in terms of an identity test for the regression coefficients.

3 Results

We have chosen the following applications for each task:

1. **Vessel Wall Segmentation in Intravascular Ultrasound Sequences.** We have applied our model to the validation of the adventitia wall detection reported in [12] in order to compare the regression-prediction assessment to standard metrics (mean distance, noted by MeanD). We have considered two sequences of 300 frames each manually segmented every 20 frames (15 samples). One case (C1) has a low error and the other one (C2) a poor performance of the automatic method.
2. **Cardiac Phase Detection.** We have applied our model to assess replacing ECG signal sampling by manual sampling of longitudinal cuts of IntraVascular UltraSound sequences [13]. Comparison of cardiac phase samplings is a difficult task because it should not penalize constant shifts associated to a sampling of a different fraction of the cardiac phase. We have considered 3 sequences between 378 and 1675 frames long and acquisition rate between 10 and 30 fps. The first case (C1) is a short segment (378 frames) acquired without pullback. The other two are a (visually) good and bad acquisitions (C2 and C3, respectively).

Our goal is by no means validating the performance of alternative methods, but to show the benefits of regression-prediction models for performance evaluation. To such end, we have assessed the validity of the linear model (given by S_R and TM test), as well as, its prediction value (given by CP_i, $i = 1, 2$). Positions are given in mm.

Tables 1 and 2 report regression parameters and predictive value for each task and, in the case of segmentation (table 1), we also report the range (computed for 300 frames) of the metric MeanD in the last column. For the regression model, we report the $p-value$ for TM test, confidence intervals for β_0, β_1 and S_R. For the prediction model, we give the interval for the interchangeability condition, CP_1, and the accuracy ϵ in mm, CP_2. In the case of detection, CP_1 has been computed for $x + \beta_0$ instead of x in order to account for constant shift in samplings.

For all cases and tasks, there is a clear linear relation between GT and the image processing AO (with p close to the working precision). For the segmentation task (table 1), C1 has an accurate regression model close to the identity line. The squared root of the model accuracy ($\sqrt{S_R} = 0.1224$) agrees with MeanD ranges computed for the 15 manually segmented samples. Concerning predictive value, manual and automatic

Table 1. Regression-Prediction Model for Vessel Wall Segmentation

	Regression Model				Prediction Model		Distance
	TM	β_1	β_0	S_R	CP_1	CP_2	MeanD
C1	$\leq 10^{-308}$	1.009 ± 0.002	0.063 ± 0.001	0.015	(-0.072,0.280)	0.200	0.105 ± 0.018
C2	$\leq 10^{-308}$	1.001 ± 0.006	0.229 ± 0.029	0.221	(-0.561,0.899)	0.822	0.365 ± 0.171

Table 2. Regression-Prediction Model Scores for Cardiac Phase Detection

	Regression Model				Prediction Model	
	TM	β_1	β_0	S_R	CP_1	CP_2
C1	$\leq 10^{-308}$	0.998 ± 0.002	0.018 ± 0.038	0.004	(-0.113 ,0.047)	0.113
C2	$\leq 10^{-308}$	0.997 $\pm 7.5e^{-4}$	-0.284± 0.026	0.003	(-0.096,0.001)	0.094
C3	$\leq 10^{-308}$	0.966 ±0.004	0.792 ± 0.255	0.662	(-1.454,-2.449)	1.362

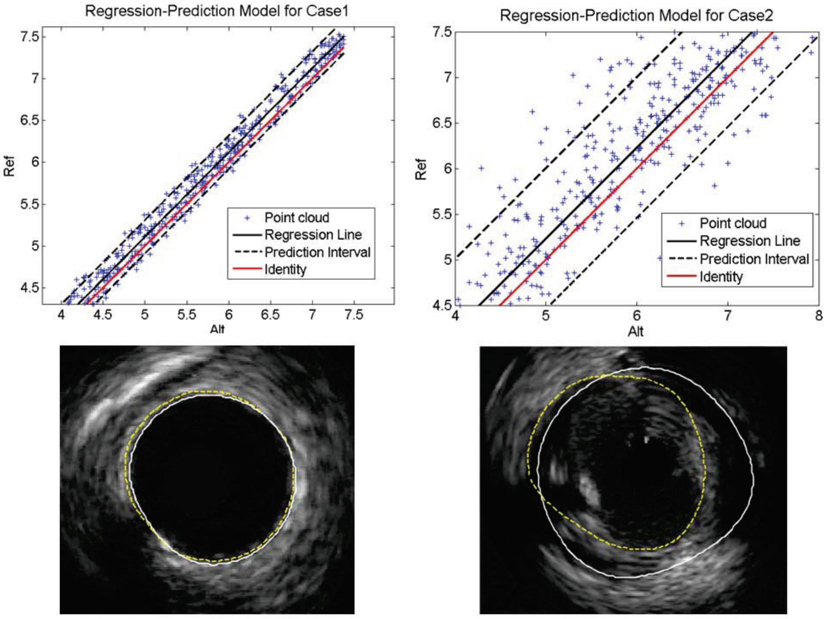

Fig. 1. Regression Model and Prediction Intervals for Vessel Wall Segmentation

contours can be swapped for the whole sequence with high accuracy. For C2, the model is a translation of the identity and the fitting is worse. This indicates severe contour misalignment (see right image in fig. 1). We observe that in this case the squared root of the fitting error ($\sqrt{S_R} = 0.4701$) also agrees with MeanD ranges. Although the two variables can be swapped (both measure the same [10]), the low accuracy of the

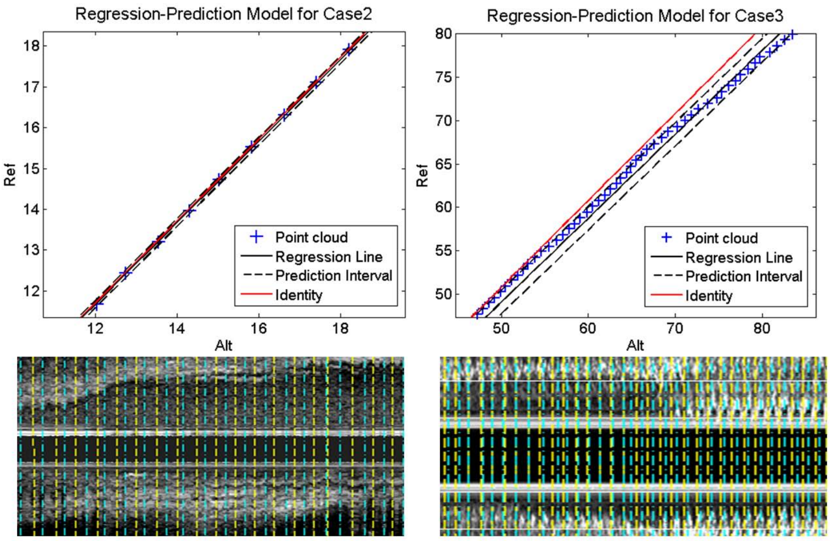

Fig. 2. Regression Model and Prediction Intervals for Cardiac Phase Detection

prediction advises against swapping them. For the detection task (table 2), C1 and C2 present an accurate regression model close to the identity line and a good predictive value. The non-zero intercept of C2 is due to a constant shift in manual samplings (see left image in fig. 2). Concerning C3, both, the regression model and the predictive value, present very bad scores and the samplings cannot be swapped (CP_1 does not hold).

Figures 1 and 2 show the regression-prediction plots for vessel wall segmentation and cardiac phase sampling, respectively. Each plot shows the point cloud AO (x-axis) versus GT (y-axis), the regression line in black solid, PI limits in dashed black and the identity line $AO = GT$ in red. In the case of vessel wall segmentation (fig. 1), we show a representative frame with manual (solid white) and automated (dashed yellow) contours, while for cardiac phase sampling (fig. 2) we show a longitudinal cut with ECG (yellow lines) and manual (cyan lines) samplings. For segmentation cases, the deviation of the identity line from the regression model is similar in both cases, though the range of the prediction interval is substantially larger for C2. This increase in error is reflected in the visual quality of the segmentation shown at the right bottom image. Regarding detection plots, visual inspection of the longitudinal sampling for C2 reasserts the agreement up to a constant shift reflected by the thin prediction interval in top left plots. For C3, the identity line traverses prediction interval upper bound as suggested by CP_1 interval. The prediction model coincides with the erratic relation between manual and ECG samplings observed in the left bottom image.

4 Conclusions and Future Work

Standardized validation of medical imaging algorithms allowing generalization of conclusions to clinical data is a challenging task not fully solved. We have approached validation from the point of view of statistical inference. In this context, we use a regression model for assessing to what extent GT and AO can be swapped and a prediction model for inferring conclusions to the whole population. Experiments on a segmentation task are a good proof of concept of the capability of the framework for assessing performance, while experiments on a detection task illustrate its versatility.

The framework presented in this paper can be applied to explore the performance from a relatively small test set. In order to fully generalize results to the whole clinical data involved in each task, we should consider a general regression model with random effects in order to account for variability across acquisitions. Also, in order to fully validate their capability for assessing performance, we are running our methods on the whole data set and metrics used in [12].

Acknowledgements. We thank InCor from USP, São Paulo, for their support with the IVUS images database. This work was supported by the Spanish projects PI071188, TIN2009-13618, TRA2010-21371-C03-01, CSD2007-00018 and the 2nd author by The Ramon y Cajal Program.

References

1. Jannin, P., Krupinski, E., Warfield, S.: Guest editorial validation in medical imaging processing. IEEE Trans. on Med. Imag. 25(11), 1405–1409 (2006)
2. Wiest-Daesslé, N., Prima, S., Morrissey, S.P., Barillot, C.: Validation of a new optimisation algorithm for registration tasks in medical imaging. In: ISBI 2007, pp. 41–44 (2007)
3. Lee, S., Abràmoff, M.D., Reinhardt, J.M.: Validation of retinal image registration algorithms by a projective imaging distortion model. In: Conf. Proc. IEEE Eng. Med. Biol. Soc. 2007, pp. 6472–6475 (2007)
4. Jannin, P., Fitzpatrick, J., Hawkes, D., Pennec, X., Shahidi, R., Vannier, M.: Validation of medical image processing in image-guided therapy. IEEE Trans. Med. Imag. 21(12), 1445–1449 (2002)
5. Gee, J.: Performance evaluation of medical image processing algorithms. In: Proc. SPIE, vol. 3979, pp. 19–27 (2000)
6. Castro, F., Pollo, C., Meuli, R., Maeder, P., Cuisenaire, O., Cuadra, M., Villemure, J.G., Thiran, J.P.: A cross validation study of deep brain stimulation targeting: From experts to atlas-based, segmentation-based and automatic registration algorithms. IEEE Trans. Med. Imag. 25(11), 1440–1450 (2006)
7. Landis, J., Koch, G.: The measurement of observer agreement for categorical data. Biometrics 33, 159–174 (1977)
8. Gerig, G., Jomier, M., Chakos, M.: Valmet: A new validation tool for assessing and improving 3D object segmentation. In: Niessen, W.J., Viergever, M.A. (eds.) MICCAI 2001. LNCS, vol. 2208, pp. 516–528. Springer, Heidelberg (2001)
9. Bland, J.M., Altman, D.G.: Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1(8476), 307–310 (1986)
10. Hoppin, J., Kupinski, M., Kastis, G., Clarkson, E., Barrett, H.: Objective comparison of quantitative imaging modalities without the use of a gold standard. IEEE Trans. Med. Imag. 21(5), 441–449 (2002)

11. Newbold, P., Carlson, W.L., Thorne, B.: Statistics for Business and Economics, 6th edn. Pearson Education, London (2007)
12. Gil, D., Hernàndez, A., Rodriguez, O., Mauri, J., Radeva, P.: Statistical strategy for anisotropic adventitia modelling in IVUS. IEEE Trans. Med. Imag. 25(6), 768–778 (2006)
13. Hernàndez-Sabaté, A., Gil, D., Garcia-Barnés, J., Martí, E.: Image-based cardiac phase retrieval in Intravascular Ultrasound sequences. IEEE Trans. Ultr., Ferr., Freq. Ctr. 58(1), 60–72 (2011)