Int J CARS (2006) 1:63–73
DOI 10.1007/s11548-006-0044-6

ORIGINAL ARTICLE

Model for defining and reporting reference-based validation protocols in medical image processing

Pierre Jannin · Christophe Grova · Calvin R. Maurer, Jr.

Published online: 18 July 2006

Abstract
Objectives Image processing tools are often embedded in larger systems. Validation of image processing methods is important because the performance of such methods can have an impact on the performance of the larger systems and consequently on decisions and actions based on the use of these systems. Most validation studies compare the direct or indirect results of a method with a reference that is assumed to be very close or equal to the correct solution. In this paper, we propose a model for defining and reporting reference-based validation protocols in medical image processing.
Materials and methods The model was built using an ontological approach. Its components were identified from the analysis of initial publications (mainly reviews) on medical image processing, especially registration and segmentation, and from discussions with experts from the medical imaging community during international conferences and workshops. The model was validated by its instantiation for 38 selected papers that include a validation study, mainly for medical image registration and segmentation.
Results The model includes the main components of a validation procedure and their inter-relationships. A checklist for reporting reference-based validation studies for medical image processing was also developed.
Conclusion The proposed model and associated checklist may be used in formal reference-based validation studies of registration and segmentation and for the complete and accurate reporting of such studies. The model facilitates the standardization of validation terminology and methodology, improves the comparison of validation studies and results, provides insight into the validation process, and, finally, may lead to better quality image management and decision making.

Keywords Reference-based validation · Medical image processing · Image registration · Segmentation · Gold standard · Ground truth · Guidelines

P. Jannin (✉) · C. Grova
Visages, U 746, INSERM-INRIA-CNRS, Medical School, University of Rennes, CS 34317, 35043 Rennes Cedex, France
e-mail: pierre.jannin@irisa.fr
http://www.irisa.fr/visages

C. Grova
Montreal Neurological Institute, McGill University, Montreal, Canada
e-mail: christophe.grova@mail.mcgill.ca

C. R. Maurer
Department of Neurosurgery, Stanford University, Stanford, CA 94305-5327, USA
e-mail: calvin.maurer@gmail.com

Introduction

The role of image processing in medicine is proportional to the increasing importance of medical imaging in the medical workflow. Image processing has an important influence on the medical decision-making process and even on surgical actions. Therefore, high quality and accuracy are expected. Sources of errors are numerous in image processing. Some errors are common to any image processing method, such as the ones related to the limited spatial resolution of the images and the associated partial volume effect, the geometrical distortion in the images, or the intrinsic data variability (e.g., patient movement during tomographic acquisition) [23]. Some others are specific to the type of processing.

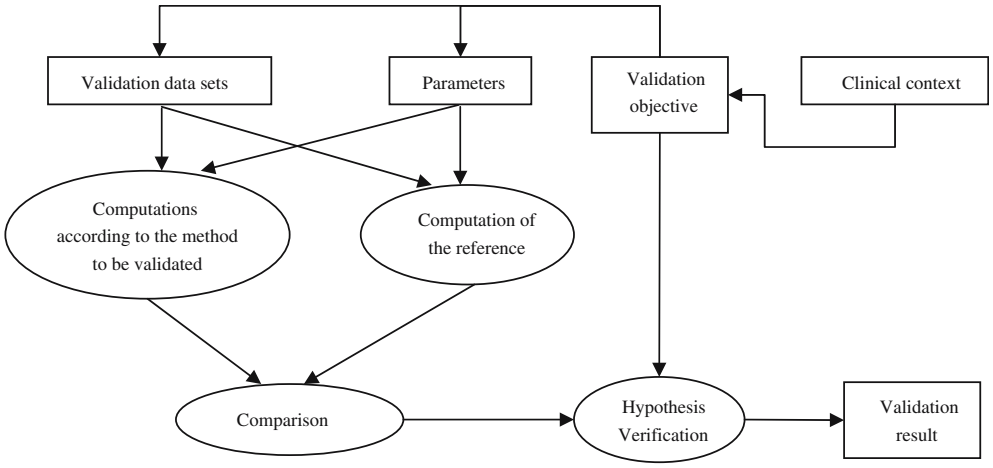

Fig. 1 Main steps of reference-based validation procedures for medical image processing

The process of performance assessment is complex and includes many different aspects. In software engineering, one distinguishes verification, validation and evaluation as follows [14]. *Verification* consists in assessing that the system is built according to its specifications (i.e., assessing that the system is built correctly). *Validation* consists in assessing that the system actually fulfills the purpose for which it was intended (i.e., assessing that the correct system was built). *Evaluation* consists in assessing that the system is accepted by the end-users and is performant for a specific purpose (i.e., assessing that the system is valuable). Verification, validation and evaluation can be performed at all points in the life cycle of the larger system in which the image processing method is embedded: on the conceptual model representing the universe of discourse, on the requirements specification extracted from the model, on the design specification, on the executable software modules, on the integrated application, on its results and finally on the results presented to the end user [2]. Evaluation levels have been previously defined in order to outline the complexity and extent of evaluation studies in medicine and especially for diagnostic imaging [11]. Six evaluation levels were distinguished: (1) technical efficacy, (2) diagnostic accuracy efficacy, (3) diagnostic thinking efficacy, (4) therapeutic efficacy, (5) patient outcome efficacy, and (6) societal efficacy. For example, in image registration, level-1 may correspond to the assessment of a validation criterion characterizing the intrinsic performance of the algorithm (assessed in a simulation stage, for instance), whereas level-2 may be concerned with the assessment of a validation criterion at clinically meaningful anatomical points or structures [12]. In this paper, our main concern is the validation of the last two steps of the system life cycle, namely, the results and the presented results. We are concerned with the validation of the image processing component rather than the larger system in which it is embedded.

Image processing methods can be validated according to specified performance criteria. In most reported validation studies, validation criteria are assessed against a "reference" (also called a "gold standard"), which is assumed to be close or equal to the correct result (also called the "ground truth") (Fig. 1). In medical image registration, the ground truth is the geometrical transformation that correctly maps points in one image to anatomically corresponding points in the other image. In image segmentation, the ground truth may be the correct anatomical labeling of each pixel or voxel of an image data set or the true structure boundaries. In this paper we use the term "reference" rather than "gold standard," which is used in a general meaning. The reference can be an exact or approximate solution based on numerical simulations or physical experiments. It can also be a solution computed using one or several image processing methods. Finally, the reference can be an expert-based solution or one using a priori knowledge about the ground truth.

The importance of validation of medical image processing methods is now well established [4,7,10,13,20, 27,31,32] but standard terminology and methodology are lacking. Standardized terminology and methodology would facilitate the complete and accurate reporting of validation studies and results and the comparison of such studies and results and may be useful in the context of quality management and decision making. A first step towards this standardization is the design of a framework for describing and representing a validation process. In this paper, we propose such a framework. It includes a model describing the main components of a reference-based validation procedure for medical image processing and a checklist designed from this model for

reporting reference-based validation studies. The model was built using an ontological approach; its components were identified from an analysis of the literature and from discussion with experts from the medical imaging community. The model was validated by instantiating it with 38 validation studies reported in the literature. We illustrate the application of this model by using it to describe two reported validation studies in the framework of the model. Finally, we compare our work with similar approaches and draw some perspectives.

Materials and methods

The model of the validation process has been defined using an ontological approach: collection of data, identification of the main concepts and relationships, design of a model, choice of a formalism to represent this model, refinement and validation of the model, and development of tools based on the defined model. The data collection consisted of two parts. First, a limited set of publications (mainly reviews) on medical image processing that include a section on validation was selected [4, 6,7,9,10,13,15–17,19,21–23,27,28,30–33]. Second, several meetings of scientific experts from the medical imaging community allowed the exchange of ideas and issues concerning validation of medical image processing (the meetings were organized at the conferences Computer Assisted Radiology and Surgery 2001, 2002, and 2003 and Medical Image Computing and Computer-Assisted Interventions 2003). Then we identified the main components involved in validation processes from ideas and concepts found in the initial set of publications. These components were used to define a model. Two representations were used for this model: a graphical process diagram and a Structured Query Language (SQL) database. The model was then iteratively refined and validated by its instantiation for 38 papers that included validation studies, mainly about medical image registration and segmentation. The publications were chosen according to the following criteria: the publication was a review of the literature concerning validation of one type of image processing method, it introduced a new image processing algorithm and validated the algorithm, it proposed an original validation methodology, or it was characteristic of a family of validation methods. However, this selection cannot be considered as exhaustive or representative. Finally, a database was created that includes, for each publication, the main characteristics of the validation process described in the publication. Associated Web-based tools were developed for browsing and viewing the database information. Additional tools allow new publications to be added to the database [18].

Results

Figure 1 describes the overall validation process of an image processing method. It starts with the specification of the validation objective, which includes the clinical context in which the validation process has to be performed, and the specification of a hypothesis, relying on expected values required within the considered clinical context. The validation process then proposes an experiment to test the hypothesis. Our model of the validation process begins with the definition of the validation data sets and the parameters that are designed to test some properties of the image processing method being validated. These data sets and parameters are applied to this image processing method, as well as to another method chosen to provide a reference for the validation study. Results computed by the image processing method and the reference method are compared. Finally, comparison results are tested against the validation hypothesis in order to provide the validation result. The following sections describe each component of the model, i.e., the validation objective, the validation data sets and parameters, the reference, estimation of the validation criterion, and finally validation hypothesis testing.

Model of the validation objective

It is essential to describe both the clinical context in which the validation is performed and the clinical objective of the validation study [4,7,20,31,32]. The validation objective may be formulated as a hypothesis. The validation process aims to test this hypothesis.

An initial version of the validation objective model, proposed in [20], is described here in more detail. We consider the clinical context (C) as the task performed by the clinician, which includes an image processing tool. This task corresponds to a clinical decision or action, such as a diagnosis or a surgical or therapeutic procedure. The validation objective relates to an image processing method (F_{M}) to be validated at a specific level of evaluation (L) as presented above and for a specific clinical context (C). This validation objective is defined by the involved validation data sets (D_{I}) and their intrinsic characteristics (e.g., imaging modalities, spatial resolution, tissue contrast), by some clinical assumptions (A) related to the data sets or to the patient (e.g., assumptions regarding anatomy, physiology, and pathology), and by the validation criterion (V_{C}) to be assessed (e.g., accuracy, precision, reliability, robustness). The valida-

tion objective consists in comparing a value of the validation criterion measured by the validation metric (F_C) on information (I) extracted from the validation data sets with an expected value or model (M_{QI}). Such comparison may be performed using a statistical hypothesis test (F_H). Therefore, a validation objective could be defined as follows: " *In a context defined by L and C and knowing A, the method* F_M *applied to the data sets* D_I *is able to provide validation results, by estimating* V_C *provided by* F_C *and computed on I, in accordance with the expected value* M_{QI}, *when compared using the test* F_H." An example of a level-1 (L), i.e., technical efficacy, validation hypothesis may be: "In the clinical context of image-guided surgery to biopsy a cranial lesion (C), a particular registration method (F_M) based on matching skin surface points in the physical space of the patient to a surface model of the skin extracted from a contrast-enhanced CT image (D_I) (assuming that the lesion enhances (A)) is able to perform registration with an accuracy (V_C) (evaluated by computing RMS error (F_C) on points within the brain (I)) that is significantly better than (F_H) the clinically expected accuracy (M_{QI})."

Model of the validation process

In this section we propose a model describing the main components and the main stages that may be involved in reference-based validation of image processing methods (Fig. 2).

Inputs: validation data sets and parameters

The validation process is performed on validation data sets (D_I) and their precise description is of high importance. Three main types of validation data sets can be distinguished: numerical simulations, physical phantoms, and clinical data sets. In image registration, one image modality can be simulated from another one or, in image segmentation, images can be simulated from known anatomical or geometrical structures. By acquiring images of physical phantoms one can control the geometry of the validation data sets and take into account the physical conditions of the image acquisition. Concerning clinical data sets as validation data sets and for image registration, some may include an extrinsic system specifically used to estimate the ground truth or to control acquisition geometry (e.g., stereotactic frame, bone-implanted fiducial markers). Some validation studies on clinical data sets require a specific protocol for validation (e.g., intra-operative identification of anatomical landmarks or fiducial markers using an optically tracked probe, for image registration validation, or multimodal image acquisition, for image segmentation). Differences between validation data sets stand on the trade off between quality of the given reference and clinical realism of the data.

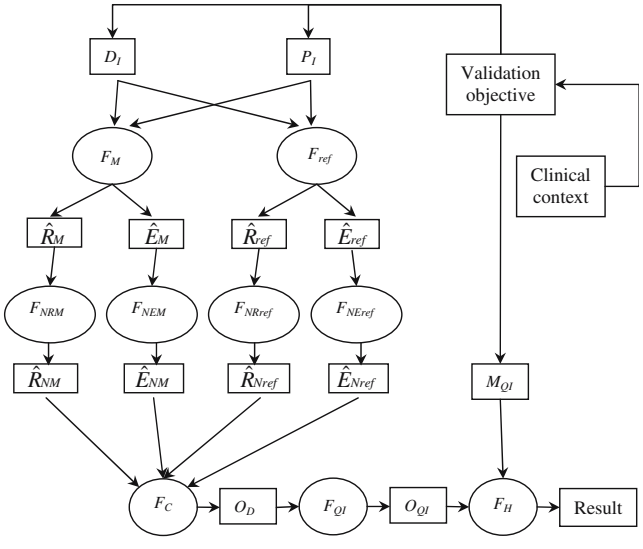

Fig. 2 Model of reference-based validation procedures for medical image processing

Input parameters (P_I) generally refer to parameters we want to study that can influence the performance of the image processing method. Input parameters include parameters related to the validation data sets or to the image processing method itself. Parameters related to the validation data sets are (1) parameters used to generate or disrupt the validation data sets, such as the signal-to-noise ratio in the images or other parameters used for numerical simulations; known misalignment range for image registration or known anatomical structure locations for image segmentation; or (2) parameters related to the clinical assumptions, such as presence or simulation of pathological areas. Parameters related to the image processing method may be, for instance, configuration values of the method (e.g., initialization), image preprocessing (e.g., smoothing, correction of MR geometrical distortions or intensity inhomogeneities), and choice of the components of the image processing method (e.g., choice of an optimization method or optimization strategy, choice of a cost function). Input parameters are discriminating for the characterization of the validation objective and the validation criteria.

Image processing method to be validated and reference

Given the validation data sets (D_I) and input parameters (P_I), let R denote the theoretical ideal result (e.g., exact geometrical transformation for registration, exact tissue boundaries for segmentation) that the image processing method to be validated (F_M) estimates. The theoretical

ideal result R denotes the ground truth of the validation process and is generally not directly available. The image processing method F_M to be validated computes an estimate $\hat{R}_M$ of the ground truth R. A reference-based validation process uses another method F_{ref} to estimate the ground truth more accurately than F_M. The method F_{ref} provides an estimate $\hat{R}_{ref}$ of R, which is considered as the reference against which the result from the method F_M will be assessed.

For image registration, the reference $\hat{R}_{ref}$ may be represented by the parameters of a geometrical transformation or by locations of fiducial markers in the validation data sets. The reference $\hat{R}_{ref}$ may be estimated from the validation data sets using a function F_{ref}. For instance, for image registration when $\hat{R}_{ref}$ is represented by geometrical parameters, F_{ref} can be the geometrical transformation used for simulating misaligned images. In this case, the reference $\hat{R}_{ref}$ can be error free and is considered as an absolute reference ($\hat{R}_{ref} = R$). The method F_{ref} can also be the computation of a geometrical transformation by aligning fiducial markers in a least-squares sense. When $\hat{R}_{ref}$ is represented by locations of fiducial markers, F_{ref} can simply be the identification of these fiducial markers in both data sets. The reference may also be computed using systems that control the location of the patient during acquisition or control displacement of the test bed within the imaging device. In this case, the reference can be either absolute or estimated depending on the acquisition process. For image segmentation, the reference may be represented by a label assigned to some voxels, by the exact tissue boundaries or by the description of the structure(s) to be segmented. The method F_{ref} can be a simple reference to the anatomical map used for simulating images or to the known geometry of an imaged physical phantom. When using clinical data sets as validation data sets, the ground truth may not be available. In such situations, a reference may be provided by expert observers (e.g., control of registration results by visual matching of the data sets, manual delineation of anatomical structures) or by a priori clinical knowledge or clinical assumptions (e.g., overlapping ratios of corresponding anatomical structures, anatomical landmarks). A reference may also consist of the results computed by one or several other independent similar image processing methods or computed from the analysis of a series of experiments, e.g., mean transformations for image registration or "averaged" contours for image segmentation.

The method F_{ref} is generally chosen to be as accurate as possible, but in some situations F_{ref} may have an error (or a bias) E_{ref} ($E_{ref} \neq 0$) that should be taken into account during the validation process. In almost all cases E_{ref} is unknown, as is the ideal result R, but in some situations, an estimate $\hat{E}_{ref}$ may be proposed. For example, in image registration, when the reference $\hat{R}_{ref}$ is computed using fiducial markers, an estimate $\hat{E}_{ref}$ may be obtained using point-based registration error theory, which provides the statistically expected target registration error (TRE) as a function of the number and configuration of the fiducials, the fiducial localization error (FLE), and the position of the target relative to the fiducials [8,26]. Similarly, Bromiley et al. [5] proposed an estimator of the covariance of the geometrical parameters for mutual information-based image registration. Intrinsic errors of the image processing method F_M itself ($\hat{E}_M$) are difficult to estimate, since they are generally not a reliable value for validation. However, the same approaches for $\hat{E}_{ref}$ estimation could be used for $\hat{E}_M$ and taken into account in the validation procedure during the final comparison with reference.

Estimation of validation criterion: discrepancies and validation metrics

The outputs of the image processing method F_M to be validated and the method F_{ref} to compute the reference are given by $\hat{R}_M$ and $\hat{R}_{ref}$, respectively, and also possibly by $\hat{E}_M$ and $\hat{E}_{ref}$. By comparing $\hat{R}_M$ and $\hat{R}_{ref}$, a validation criterion aims at characterizing different properties of the method F_M, such as its accuracy, precision, robustness or reliability. Three features describe the quantification of a validation criterion from the results of F_M and F_{ref}: (1) the kind of information on which the comparison will be performed, (2) a comparison function F_C, and (3) a function F_{QI} to compute a quality index from the results of F_C. This quality index is chosen to estimate the validation criterion.

The first feature of a validation criterion is the type of information on which the validation criterion will be measured. The comparison metric may be applied directly to the output of F_M and F_{ref} ($\hat{R}_M$ and $\hat{R}_{ref}$, respectively), as for instance the parameters of the geometrical transformation or the volumes of the segmented areas. But it is generally more interesting to provide the clinician with meaningful discrepancy measurements, such as spatial distance between anatomical points or surfaces for image registration or specificity and sensitivity for image segmentation. In that case, points, surfaces, or even volumes may be the information on which the validation criterion will be estimated. For that purpose, a normalization step (F_{NRM} and F_{NRref}) may consist in transforming $\hat{R}_M$ and $\hat{R}_{ref}$ to such meaningful information, respectively, to compute $\hat{R}_{NM}$ and $\hat{R}_{Nref}$. For instance, for image registration, the normalization step may use $\hat{R}_M$ and $\hat{R}_{ref}$ to resample whole data sets in the same coordinate system, by applying them to a list

of points. For image segmentation, the normalization step may consist in converting a contour into a region. Similarly, a normalization step ($F_{\mathrm{NEref}}, F_{\mathrm{NEM}}$) could be performed on the estimation of intrinsic errors $\hat{E}_{\mathrm{ref}}$ and $\hat{E}_{\mathrm{M}}$, when available, in order to compute normalized errors $\hat{E}_{\mathrm{Nref}}$ and $\hat{E}_{\mathrm{NM}}$.

Normalized output from the method ($\hat{R}_{\mathrm{NM}}, \hat{E}_{\mathrm{NM}}$) and from the computation of the reference ($\hat{R}_{\mathrm{Nref}}, \hat{E}_{\mathrm{Nref}}$) must be analyzed using a comparison function F_{C}, which measures a "distance" to the reference. In our model, we call discrepancy the result of the comparison $O_{\mathrm{D}} = F_{\mathrm{C}}(\hat{R}_{\mathrm{NM}}, \hat{R}_{\mathrm{Nref}})$, with a given set of validation data sets and parameters values. Ideally, computation of O_{D} should also take into account $\hat{E}_{\mathrm{Nref}}$ and $\hat{E}_{\mathrm{NM}}$, when available. For image registration, if outputs to be compared are geometrical parameters, differences between parameters may be used such as rotation and translation errors. If the outputs are points (e.g., vertices of a head bounding box, anatomical structures or points of interest), one may use a Euclidean distance, such as TRE. If the outputs are volumes, one may use a similarity measurement computed on intensity values of data sets, previously resampled in the same coordinate system (e.g., least square difference, correlation coefficient, standard deviation of the difference image). Overlapping ratios between $\hat{R}_{\mathrm{NM}}$ and $\hat{R}_{\mathrm{Nref}}$, corresponding to the anatomical structures, may also be computed. For image segmentation, if the outputs are surfaces, the Hausdorff distance between the surfaces may be used. If the outputs are volumes, the discrepancy may be computed on the number of correctly or incorrectly segmented voxels. False negative, false positive, true positive and true negative volume fractions provide measures [27], which can be further used to compute sensitivity, specificity, or Receiver Operating Characteristic (ROC) curves. Kappa statistics or Dice's similarity coefficient may also be used to characterize the discrepancy [34,35]. Positions of the incorrectly segmented voxels can also be taken into account in the computation of the discrepancy. Finally, discrepancy may be computed on general characteristics related to the segmented objects or structures, such as position and number of incorrectly segmented objects or values of segmented object features [33].

Given different validation data sets and parameter values and a method F_{C} to compute discrepancies, a quality index (O_{QI}) may be computed using a function F_{QI}. This function computes a statistical measure of the distribution of local discrepancies by assessing an intrinsic and/or a global discrepancy. The intrinsic discrepancy reflects properties of the distribution of the local discrepancies in a condition when validation data sets and parameters are fixed (e.g., to characterize the spatial distribution of local discrepancies), whereas the global discrepancy corresponds to the study of the variability of intrinsic discrepancies among different test conditions (i.e., when using several validation data sets with the same method or with different methods). Standard statistics are generally used to characterize the properties of discrepancies distributions, such as mean or root-mean-square error, standard deviation of the error, order statistics of the error (e.g., median, maximum, percentiles of the distribution), or false positive rate. ROC curves and area under the curve (AUC) computed from ROC curves may also serve as quality indices.

Hypothesis testing

Finally, quality indices may be used for statistical analysis of the results. The function F_{H} provides the final result of the validation (i.e., to reject or not the hypothesis expressed in the validation objective). The value of the quality index (O_{QI}) is compared to an expected value or to an a priori model (M_{QI}) defined in the validation objective. This test may be a simple test on a threshold (e.g., fault rate) or a statistical hypothesis test (e.g., paired *t*-test, Wilcoxon rank sum test or sign test, Kolmogorov's test, analysis of variance).

Example applications of our model to reported validation studies

To illustrate our validation model, we describe in the framework of our model the validation studies reported by Maurer et al. [22] for image registration and Aubert-Broche et al. [1] for image detection. These papers are in the set of 38 selected papers used to refine and validate the model.

In Maurer et al. [22], the objective of the paper was to present a new registration algorithm, the weighted geometrical feature (F_{M}) algorithm that uses a weighted combination of multiple geometrical feature shapes (e.g., points and surfaces) for registration. The clinical validation objective was to evaluate the accuracy and precision (V_{C}) of this algorithm for the registration of CT images to the physical space of the patient in the context of cranial image-guided surgery (C). Evaluation levels 1 and 2 were considered (L). The validation data sets (D_{I}) consisted of 12 clinical data sets. Each data set includes a CT head image obtained with four bone-implanted markers, positions of the markers in CT and physical space, positions of a large number of skin and bone surface points in physical space, and triangle set representations of skin and bone surfaces extracted from the CT image.

The studied input parameters (P_{I}) were parameters of the registration method: features used for registration (e.g., one marker, skin surface, and bone surface),

weights of the features, the registration algorithm termination threshold value, and outlier threshold value. The output $\hat{R}_M$ of the registration method (F_M) was the rigid transformation parameters (i.e., rotation matrix and translation vector). No intrinsic error of $F_M(\hat{E}_M)$ was computed ($F_{NEM}, \hat{E}_{NM}$ not applicable). The computation (F_{ref}) of the reference $\hat{R}_{ref}$ was point-based registration using three markers. The reference $\hat{R}_{ref}$ consisted of the rigid transformation parameters (i.e., rotation matrix and translation vector) provided by the three-marker reference registration. The normalization functions F_{NRM}, F_{NRref} mapped every CT voxel inside the brain that was within 75 mm of the center of the craniotomy to physical space. $\hat{R}_{Nref}$ and $\hat{R}_{NM}$ were positions of brain points near the craniotomy mapped from image to physical space. $\hat{E}_{ref}$ was estimated from the computation of TRE obtained from fiducial registration error (FRE) using results from point-based registration error theory, but was not accounted for in F_C. No normalization function F_{NEref} was used ($\hat{E}_{Nref} = \hat{E}_{ref}$). The mean value of TRE (F_C), which was computed as lengths of vector differences between brain points mapped by the evaluated transformation ($\hat{R}_{NM}$) and the normalized reference ($\hat{R}_{Nref}$), averaged over all brain points for each patient, was computed as well as the mean, standard deviation, and 95% TRE values over all patients for six types of registration (F_{QI}, O_{QI}): skin surface only, skin surface plus one marker, bone surface only, bone surface plus one marker, skin plus bone surfaces, and skin plus bone surfaces plus one marker. No statistical evaluation (F_H) was performed.

In Aubert-Broche et al. [1], the objective of the paper was to present a new method (F_M) for detecting interhemispheric asymmetries of brain perfusion in SPECT images. The clinical validation objective was to evaluate detection performances (V_C) on simulated SPECT images in the context of epilepsy surgery (C). Evaluation levels 2 and 3 were considered (L) in two different studies; only one of the studies is described here. The validation data sets (D_I) consisted of realistic analytical SPECT simulations performed with and without any anatomical asymmetry (A). Functional asymmetric zones of various sizes and intensities were introduced (A). A large number of simulations were computed (256 simulations representing all permutations of two anatomical asymmetries, four localizations, four sizes, and eight amplitude values for functional asymmetries). The studied input parameters (P_I) were parameters related to the validation data sets: size, location, and amplitude of possible asymmetric functional areas; parameters related to the detection method were also studied but are not described here. The output $\hat{R}_M$ of the detection method (F_M) was a statistical volume indicating at each voxel the probability of functional asymmetry. No intrinsic error of F_M ($\hat{E}_M$) was computed ($F_{NEM}, \hat{E}_{NM}$ not applicable). The reference $\hat{R}_{ref}$ consisted of the known simulated asymmetric areas stored in volumes, thus F_{ref} = identity, $\hat{E}_{ref} = \hat{E}_{Nref} = 0$. No normalization functions F_{NRM}, F_{NRref} were used, thus $\hat{R}_{Nref} = \hat{R}_{ref}$ and $\hat{R}_{NM} = \hat{R}_M$. The degree of overlap between the actual asymmetric zone ($\hat{R}_{ref}$) and the estimated one ($\hat{R}_M$) was calculated (F_C) by assigning voxels to true positives, true negatives, false positives, and false negatives. Then, ROC curves (F_C) were deduced. The AUC was used as an index characterizing the detection performance of the method (F_{QI}, O_{QI}). The Wilcoxon rank sum test (F_H) was used to test differences in performance between various parameters in the simulations.

Bibliographic application

The results of the instantiation of our model using validation studies from the 38 selected papers have been stored in a SQL database, which can be browsed and viewed on line [18]. Queries on this database are available based on values found from main model components found in the studied publications (Table 1). Queries allow the display of references corresponding to predefined criteria.

Checklist for reporting validation studies of medical image processing

We suggest the use of a checklist in scientific contributions and publications for reporting reference-based validation studies for medical image processing (Table 2).

Discussion

The manual method used in this paper to identify the components of a validation procedure in medical image processing and to design the model is a usual approach [3]. Techniques exist for automatic extraction of information from raw textual data (i.e., text mining), but these were not found adequate in this context since relevant information (i.e., description of the validation procedure) was generally only a small part of the papers and often was described in a non-standard fashion with ad hoc terminology. Standardization of terminology is precisely one objective of our approach. The model was validated by its instantiation on a limited set of papers. One could argue that such a model should be validated on more publications. Our approach is iterative, however: we introduced an initial version of the model that we hope to continuously improve. This model was successfully used for reporting accuracy studies in

Table 1 Values of some model components found in the studied publications for image registration

Symbol	Description	Values
D_I	Validation data sets	Simulations, physical phantoms, clinical data sets
P_I	Input parameters	Data: parameters from numerical simulations, known misalignment range, presence or simulation of pathological areas; Method: registration initialization, image preprocessing (e.g., segmentation, smoothing, and correction of MR geometrical distortions or intensity inhomogeneities), choice of an optimization method or optimization strategy, choice of a cost function
F_{ref}	Function which computed the reference from D_I and P_I	Stereotactic frame or fiducial marker based registration, fiducial marker identification, systems that control the patient location during acquisition (e.g., head holder) or control displacement of the test bed within the imaging device, other registration methods considered as reference, analysis of a series of experiments, e.g., mean transformations, reconciled mean transformation, inconsistencies, or none
$\hat{R}_{ref}$	Reference	Parameters of a reference geometrical transformation or locations of fiducial markers
$\hat{E}_{ref}$	Estimated error relative to the computation of $\hat{R}_{ref}$ by F_{ref}	TRE or none
F_{NRM}	Function which transforms $\hat{R}_M$ for comparison with the reference	Transforming points or surfaces in new coordinate systems, resampling volumes in new coordinate systems, or none
F_{NRref}	Function which transforms $\hat{R}_{ref}$ for comparison	Transforming points or surfaces in new coordinate systems, resampling volumes in new coordinate systems, or none
$\hat{R}_{NM}(\hat{R}_{Nref})$	Normalized output from the method (respectively, from the reference)	Geometrical parameters, points (e.g., vertices of a head bounding box, anatomical structures or points of interest, or points uniformly distributed in the skin or brain)
F_C	Comparison function between $\hat{R}_{NM}$ and $\hat{R}_{Nref}$	Differences between geometrical parameters, Euclidean distance between 2D points, 3D points, or TRE, intensity-based differences (e.g., least square difference, standard deviation of the difference image, correlation coefficient), or overlapping ratios between anatomical structures
O_{QI}	Quality index computed on O_D	Mean or root-mean-square error, standard deviation of the error, order statistics of the error (e.g., median, maximum, percentiles of the distribution), false positive rate
F_H	Function which tests the hypothesis (i.e., comparison of O_{QI} and M_{QI})	Paired t test, Wilcoxon rank sum test or sign test, Kolmogorov test, analysis of variance

augmented reality for image-guided neurosurgery using physical phantoms [25]. It was also used for reporting a detection accuracy study of EEG source localization techniques, in the clinical context of the localization of epileptic spike generators [15]. Additionally, in both examples, our validation model helped us to rigorously design the validation protocols. It even helped in the generation of realistic simulated data appropriate for our validation objective, and in the definition of new validation metrics. We also found it useful for reviewing, classifying, and comparing validation methods in medical image processing. Our model is the union of components and functions we encountered in our review and in our discussions with experts. Thus some components of our model are optional depending on the validation procedure (e.g., normalization functions: F_{NRM}, F_{NRref},

Table 2 Checklist of components to include when reporting a validation study of medical image processing

Component	Symbol	Value
Validation objective		
Clinical context	C	
Evaluation level	L	
Validation criterion	V_C	
Clinical assumptions on patient and data sets	A	
Information extracted from data sets for evaluation	I	
Type, number, and characteristics of validation data sets	D_I	
Studied input parameters	P_I	
Expected result or model	M_{QI}	
Method to be validated	F_M	
Format of the output of the method	$\hat{R}_M$	
Intrinsic error of the method	$\hat{E}_M$	
Normalization function for the output of the method	F_{NRM}	
Format of the normalized output of the method	$\hat{R}_{NM}$	
Normalization function for the intrinsic error	F_{NEM}	
Format of the normalized intrinsic error	$\hat{E}_{NM}$	
Method to compute–estimate the reference	F_{ref}	
Reference type and format	$\hat{R}_{ref}$	
Reference estimated error	$\hat{E}_{ref}$	
Normalization function for the reference	F_{NRref}	
Format of the normalized reference	$\hat{R}_{Nref}$	
Normalization function for the reference error	F_{NEref}	
Format of the normalized reference error	$\hat{E}_{Nref}$	
Comparison function	F_C	
Result of comparison: discrepancy	O_D	
Function(s) to compute quality index(ices)	F_{QI}	
Quality index(ices)	O_{QI}	
Statistical test(s)	F_H	
Result of statistical test(s)		

F_{NEM}, F_{NEref}; error related to the reference: $\hat{E}_{ref}$; and intrinsic error of F_M: $\hat{E}_M$). We proposed a framework to describe and report validation procedures, but not a method to perform this validation.

Standardized description of clinical context and validation objective

As already outlined, the first stages when reporting a validation process are (1) the specification of the clinical context in which the application of the image processing method has to be validated and (2) the exact and precise specification of the objective of the validation procedure. Standard description of both specifications is not obvious.

Udupa et al. [27] proposed a simple characterization of what they called the "application domain." The application domain is described by three letters: A for the application or task (e.g., volume estimation of tumors), B for the imaged body part (e.g., brain), and P for the imaging protocol (e.g., MR imaging with a particular set of image acquisition parameters such as a fluid attenuated inversion recovery (FLAIR) pulse sequence). Buvat et al. [7] defined the "abstract aim" as the evaluation of a method at a specific evaluation level (as defined in Introduction). Then they defined the validation hypothesis as the projection of this abstract aim in the clinical context, which is defined as "the environment in which the method is to be evaluated." In our definition of the validation objective, we included the specification of the clinical context as defined by Udupa and Buvat, but we provided more details, especially regarding assumptions related to the data sets and the performance expectations.

Comparison with literature

Even if the standardization of validation terminology and methodology has already been outlined by our community, there are few published papers concerning the modeling of validation methods for medical image processing. The model proposed by Yoo et al. [32] restrained input data or test data to the visible human project data (our D_I). It clearly distinguished steps concerning computation of a reference (F_{ref}) and F_M. An automated scoring task included our functions F_C and F_{QI} in one step, whereas we distinguish the computation of a discrepancy between results and reference (F_C) from the computation of a quality index (F_{QI}). These computations represent, in our opinion, two fundamentally different components, even if they are often embedded in a single step. A notion of quality indices (O_{QI}) appeared as "figures of merit." Finally the statistical analysis ending the validation process is analogous to our hypothesis test (F_H). It appears that our model is more generic concerning input data, computation of the reference, and the last evaluation step called statistical analysis. In our model, we take into account errors in the whole process during the comparison step. We also add a normalization step (F_{NRM} and F_{NRref}) in order to provide discrepancy measurements that are more meaningful for the clinician (e.g., spatial distances). These steps did not appear in Yoo's model.

The model proposed by Buvat et al. [7] is of interest because it distinguished a step transforming the output data of the method into a result adequate for comparison. This step corresponds to our normalization step (F_{NRM}). But errors were not modeled and access to the reference (F_{ref}) was not explicitly modeled. As is the case for Yoo's model, Buvat did not distinguish computation of a discrepancy between results and reference from the computation of a quality index.

A standard was proposed [3] to describe and report diagnostic accuracy studies within clinical trials:

Table 3 Validation standardization levels in medical image processing

Standardization level	Requisite	Objective
1	Reporting validation procedures with standard terminology and the proposed validation model	Improve the understanding of the validation procedure and its results
2	Using suggested values for the model's components (e.g., data sets, comparison function, statistical tests)	Assess the quality of the validation method
3	Verifying suggested validation objectives	Assess the quality of the image processing method

STARD (Standard for Reporting of Diagnostic Accuracy). The method used to define their standard is similar to our approach. Based on a review of the literature, they proposed an ontology formalized by a flowchart and a checklist. Both were designed to facilitate the description of a validation study by the authors or by the peer reviewers. Similar approaches have been introduced such as the REMARK guideline (Reporting recommendations for tumour MARKer prognostic studies). Results consist of guidelines to provide relevant information about the study design, the data, and the method. As mentioned, the goal of such approaches is to "encourage transparent and complete reporting" to make the "relevant information available to others to help them to judge the usefulness of the data and understand the context in which the conclusions apply" [24]. Our approach has similar objectives with main applications in reporting technical evaluation and validation studies from levels 1 and 2 of medical image processing.

From formalization to standardization

Our long-term objective is to help standardize the terminology and methodology of validation. Standardization is not possible without a model. Characterization and formalization of validation procedures using a model is a required step for standardization of these procedures. Added values of such standardization rely on three main levels (Table 3). This paper concerns the first one. By describing a validation method with the proposed scheme, researchers should improve the understanding of their validation procedure and their results. No default values or functions should be required for this first level. It cannot assess the quality of the validation method but merely provide an easier understanding of the validation process. The second and third levels are of a prospective nature and should be accomplished with standardization committees. The second level requires the use of functions, parameters, and data sets defined or selected by such committees, including widely accepted validation metrics such as TRE for fiducial-based registration, or shared validation data sets such as the Vanderbilt data sets [29]. It can assess the quality of the validation method and allow comparison with literature and optionally meta-analysis, but not assess the quality of the image processing method. This second level obviously requires the first one. The third level requires the second one and aims at assessing the quality of the image processing method according to a clinical objective. It requires several validation objectives assessing different features of the method and/or different evaluation levels. The validation process should address hypotheses relevant to the clinical context. The first two levels are approvals of the validation method. The third one may concern approval of the validated method itself.

Conclusion

In this paper we propose a model of reference-based validation procedures for medical image processing. This model aims to facilitate the description of a validation procedure and its results. As with similar approaches [3], use of the proposed model should enable clearer and more complete reporting of validation procedures, as well as better understanding of the validity and generalizability of validation results. We suggest the use of this model and the associated checklist in scientific contributions and publications for reporting reference-based validation studies for medical image processing. By improving validation methodology of medical image processing components, our approach could enhance the clinical acceptance of many applications using medical image processing, from diagnosis to therapy, and facilitate technology transfer from the lab to bedside.

Acknowledgements The authors acknowledge the participants of the dedicated meetings at the conferences Computer Assisted Radiology and Surgery (CARS) and Medical Image Computing and Computer-Assisted Interventions (MICCAI)—Profs. J.M. Fitzpatrick, M. Vannier, X. Pennec, R. Shahidi, D. Hawkes, and R. Bucholz—for helpful discussions and for providing useful comments and suggestions. The authors also thank Prof. H. Lemke for helping organize such discussions at the CARS conferences.

References

1. Aubert-Broche B, Grova C, Jannin P, Buvat I, Benali H, Gibaud B (2003) Detection of inter- hemispheric asymmetries of brain perfusion in SPECT. Phys Med Biol 48(11):1505–1517
2. Balci O (2003) Verification, validation and certification of modeling and simulation applications. In: Proccedings of the 2003 winter simulation conference, pp 150–158
3. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Lijmer JG, Moher D, Rennie D, de Vet HCW (2003) Toward complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Acad Radiol 10(6):664–669
4. Bowyer KW, Loew MH, Stiehl HS, Viergever MA (2001) Methodology of evaluation in medical image computing. Report of Dagstuhl workshop, March 2001. http://www.dagstuhl.de/DATA/Reports/01111/ (Accessed in January 2006)
5. Bromiley PA, Pokric M, Thacker NA (2004) Empirical evaluation of covariance estimates for mutual information coregistration. In:Barillot C, Haynor DR, Hellier P, (eds) MICCAI 2004—Part I. Lecture notes in computer sciences, vol LNCS-3216 Springer,Berlin Heidelberg New York pp 607–614
6. Brown LG (1992) A survey of image registration techniques. ACM Comput Surv 24(4):325–376
7. Buvat I, Chameroy V, Aubry F et al (1999) The need to develop guidelines for evaluations of medical image processing procedures. In: Proccedings of SPIE medical imaging, Vol 3661, pp 1466–1477
8. Fitzpatrick JM, West JB, Maurer CR Jr. (1998) Predicting error in rigid-body, point-based registration. IEEE Trans Med Imaging 17(5):694–702
9. Fitzpatrick JM, Hill DLG, Maurer CR Jr (2000) Image registration. In: Sonka M, Fitzpatrick JM (eds) Handbook of medical imaging.Medical image processing and analysis. vol 2 SPIE Press, Bellingham, pp 447–513
10. Fitzpatrick JM (2001) Detecting failure, assessing success. In: Hajnal JV, Hill DLG, and Hawkes DJ (eds) Medical image registration. CRC Press, Boca Raten, pp 117–139
11. Fryback DG, Thornbury JR (1991) The efficacy of diagnostic imaging. Med Decis Mak 11(2):88–94
12. Gazelle GS, Seltzer SE, Judy PF (2003) Assessment and validation of imaging methods and technologies. Acad Radiol 10(8):894–896
13. Gee J (2000) Performance evaluation of medical image processing algorithms. In: Hanson K (eds) Proccedings. of SPIE medical imaging, image processing, vol 3979, pp 19–27
14. General principles of software validation; Final guidance for industry and FDA staff v2.0 (2002) http://www.fda.gov/cdrh/comp/guidance/938.html (Accessed in January 2006)
15. Grova C, Daunizeau J, Lina JM, Benar CG, Benali H, Gotman J (2005) Evaluation of EEG localization methods using realistic simulations of interictal spikes. Neuroimage 29(3):734–753
16. Hawkes DJ (1998) Algorithms for radiological image registration and their clinical application. J Anat 193:347–361
17. Hill DLG, Batchelor PG, Holden M, Hawkes DJ (2001) Medical image registration. Phys Med Biol 46(3):1–45
18. http://idm.univ-rennes1.fr/VMIP/model (Accessed in May 2006)
19. Jannin P, Grova C, Gibaud B (2001) Medical applications of NDT data fusion. In: Gros XE (ed), Applications of NDT data fusion. Kluwer Dordrecht, pp 227–267
20. Jannin P, Fitzpatrick JM, Hawkes DJ, Pennec X, Shahidi R, Vannier MW (2002) Validation of medical image processing in image-guided therapy. IEEE Trans Med Imaging 21(12):1445–1449
21. Maintz JBA, Viergever MA (1998) A survey of medical image registration. Med Image Anal 2:1–36
22. Maurer CR Jr, Maciunas RJ, Fitzpatrick JM (1998) Registration of head CT images to physical space using a weighted combination of points and surfaces. IEEE Trans Med Imaging 17:753–761
23. Maurer CR Jr, Rohlfing T, Dean D, West JB, Rueckert D, Mori K, Shahidi R, Martin DP, Heilbrun MP, Maciunas RJ (2002) Sources of error in image registration for cranial image-guided surgery. In: Germano, IM(ed). Advanced techniques in image-guided brain and spine surgery. Thieme,New York, pp 10–36
24. McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM et al (2005) Reporting recommendations for tumour MARKer prognostic studies (REMARK). Br J Cancer 93(4):387–91
25. Paul P, Fleig O, Jannin P (2005) Augmented virtuality based on stereoscopic reconstruction in multimodal image-guided neurosurgery: methods and performance evaluation. IEEE Trans Med Imaging 24(11):1500–1511
26. Pennec X, Thirion JP (1997) A framework for uncertainty and validation of 3D registration methods based on points and frames. Int J Comput Vis 25(3):203–229
27. Udupa J, Leblanc V, Schmidt H, Imielinska C, Saha P, Grevera G, Zhuge Y, Currie L, Molholt P, Jin Y (2002) Methodology for evaluating image-segmentation algorithms. In: Proccedings of SPIE medical imaging, vol 4684, pp 266–277
28. Van DenElsen PA, Pol EJD, Viergever MA (1993) Medical image matching — a review with classification. IEEE Eng Med Biol Mag 12(1):26–39
29. West J, Fitzpatrick JM, Wang MY et al (1997) Comparison and evaluation of retrospective intermodality brain image registration techniques. J Comput Assist Tomogr 21(4):554–566
30. Woods RP, Grafton ST, Holmes CJ et al (1998) Automated image registration: I. General methods and intrasubject, intramodality validation. J Comput Assist Tomogr 22(1):139–152
31. Woods RP (2000) Validation of registration accuracy.In: Bankman IN (ed), Handbook of medical imaging, processing and analysis,vol 30. Academic, pp 491–497
32. Yoo TS, Ackerman MJ, Vannier M (2000) Toward a common validation methodology for segmentation and registration algorithms. In: Delp LD, DiGioia AM, Jaramaz B (eds) MICCAI 2000. Lecture notes in computer sciences, vol. LNCS-1935, Springer,Berlin Heidelberg New York, pp 422–431
33. Zhang YJ (1996) A survey on evaluation methods for image segmentation. Pattern Recognit 29(8):1335–1346
34. Zijdenbos A, Dawant B, Marjolin R (1994) Morphometric analysis of white matter lesions in mr images: methods and validation. IEEE Trans Med Imaging 13(4):716–724
35. Zou KH, Warfield SK, Bharatha A, Tempany CMC, Kaus MR, Haker SJ, Wells III WM, Jolesz FA, Kikinis R (2004) Statistical validation of image segmentation quality based on a spatial overlap index. Acad Radiol 11(2):178–189