Validation of Medical Image Processing in Image-Guided Therapy

I. Introduction

CLINICAL use of image-guided therapy (IGT) systems is widespread today and growing rapidly, creating the need for a common and rigorous validation methodology, as reported in recent workshops and conferences [1]–[6]. One key characteristic of IGT systems is that they employ medical image processing methods (e.g., segmentation, registration, visualization, and calibration). As a result of this intrinsic structure, validation of IGT systems should include both individual components, the overall system and evaluation of how uncertainties propagate through the entire IGT process. Today, almost all peer-reviewed publications reporting on the development of new medical image processing methods include a validation section, although this was not always true in the past.

Validation of a medical image processing method allows its intrinsic characteristics to be highlighted, as well as evaluation of its performance and limitations. Moreover, validation clarifies the potential clinical contexts or applications that the method may serve. Validation may also demonstrate a method's clinical added value as well as estimate social or economic impact. However, standardization of validation processes is required in order to compare various IGT systems. Validation tests can facilitate the user's task of determining whether a particular system meets a given set of clinical requirements.

This editorial identifies the principal requirements of IGT system validation and encourages the medical imaging community to develop a common methodology and terminology so we may all share analyses and results in this topic.

II. Validation

IGT system validation is a special case of health care technology assessment (HCTA). Goodman [7] defines the HCTA as the "process of examining or evaluating and reporting properties, effects and/or impact of a medical technology." Goodman divides this process into the following steps: 1) identify assessment topics; 2) clearly specify assessment problem or question (i.e. assessment objective); 3) determine locus of assessment (e.g., who will perform the assessment?); 4) retrieve available evidence; 5) collect new primary data; 6) interpret evidence; 7) synthesize evidence; 8) formulate findings and recommendations; 9) disseminate findings and recommendations; and 10) monitor impact.

The efficacy of diagnostic imaging systems is evaluated at six main levels that span the range from technical performance to societal value [8]. The six levels of efficacy evaluation include: 1) technical capacity; 2) diagnostic accuracy; 3) diagnostic impact (i.e., improvement of diagnosis); 4) therapeutic impact (i.e., influence in the selection and delivery of the treatment); 5) patient outcome (i.e., improvement of the health of the patient); and 6) societal impact (e.g., cost effectiveness). In IGT, level 1 would correspond to the study of technical feasibility and level 2 to the study of system accuracy on relevant anatomical or pathological areas whereas levels 3–6 would remain approximately the same. An evaluation study must consider only one level at a time but a complete evaluation study should theoretically address all these levels separately. An example of validation at both levels 1 and 2 is provided in the development of the MAGI stereo augmented reality system based on a modified binocular operating microscope [9]. Computer simulation of error propagation in the complete system, tests of individual components and phantom tests of the complete system provided the level 1 validation (technical capacity). Level 2 validation (guidance accuracy) was provided by assessment of overlay accuracy in a series of surgical procedures in which three-dimensional (3-D) alignment accuracy was recorded.

A key characteristic of IGT systems is that various medical image processing methods are encountered in all stages of an IGT process, in pre-planning, planning, simulation, treatment delivery and posttreatment control. These methods must be validated separately, as well as the overall system. In this paper, we will primarily focus on the two first validation levels. The other levels apply to IGT systems, and should be addressed, but they are beyond the scope of this paper.

III. Criteria for Validation

Validation requires the application of defined criteria to a device or process. Common examples of validation criteria which may be applicable to IGT include:

Accuracy: Goodman [7] defines accuracy as the "degree to which a measurement is true or correct." For each sample of experimental data local accuracy is defined as the difference between observed values and theoretical values, i.e., known from a ground truth. This difference is generally referred to as local error. Under specific assumptions, a global accuracy value can be computed for the entire data set from a combination of local accuracy values.

Precision and Reproducibility or Reliability: Precision of a process is the resolution at which its results are repeatable, i.e., the value of the random fluctuation in the measurement made by the process. Precision is intrinsic to this process. Goodman defines reliability as "the extent to which an observation that is repeated in the same, stable population yields the same result."

Robustness: The robustness of a method refers to its performance in the presence of disruptive factors such as intrinsic data variability, pathology, or interindividual anatomic or physiologic variability.

Digital Object Identifier 10.1109/TMI.2002.806568

Precision and robustness computations do not necessarily require a ground truth.[1] For instance, repeatability studies may examine the intrinsic distribution error (e.g., mean value and standard deviation).

Consistency or Closed Loops: This criterion is mainly studied in image registration validation [10]–[12], by studying the effects of the composition of n transformations that forms a circuit: $T_{n1} \circ \cdots \circ T_{23} \circ T_{12}$. The consistency is a measure of the difference of the composition from the identity. This criterion does not require any ground truth, but the onus is on the user to convince that there is no bias in the error estimates obtained. Even when the methods used to compute the different transformations are different, there will be significant risk that different methods will produce similar errors.

Other criteria from algorithmic evaluation may be addressed (e.g., fault detection, functional complexity and computation time, code verification, algorithmic proof).

Fault Detection: This is the ability of a method to detect when it succeeds (e.g., result is within a given accuracy) or fails.

Functional Complexity and Computation Time: These are characteristics of method implementation. Functional complexity concerns the steps that are time-consuming or cumbersome for the operator. It deals both with man–computer interaction and integration in the clinical context and has a relationship with physician acceptance of the system or method. The degree of automation of a method is an important aspect of functional complexity (manual, semi automatic, or automatic).

Among the most important validation criteria applied in the U.S. market are those required to receive premarket approval for a medical device from the Food and Drug Administration (FDA). Other nations have similar medical device regulatory agencies. In the European Union, every medical device that is put on the market has to be "CE marked" and the responsibility for CE marking is devolved to the relevant competent authorities within each country (e.g., AFSSAPS in France, TÜV in Germany, and MDA in the United Kingdom). Briefly, the FDA criteria are derived from a legal requirement that the device be shown to be safe and effective. If a predicate device exists, the FDA may grant approval (510 K) based on substantial equivalence in performance. Otherwise a premarket approval (PMA) is required consisting in clinical trials (e.g., human studies) for a specific indication. The gold standard for most PMA evaluations is the randomized and blinded multicenter clinical trial, a costly and time-consuming endeavor. For practical reasons, demonstration of feasibility and comparative performance will suffice for journal publication, but not for widespread dissemination and clinical use.

Other factors may have to be studied but are beyond the scope of this communication such as cost effectiveness, patient acceptance, and outcome factors.

IV. Validation Requirements

The principal technical requirements for validation include: standardization of validation methodology, design of validation data sets, and validation metrics [1]–[6], [13]–[15].

A. Standardization of Validation Methodology

Actual validation methodologies lack standardization. Without standardization it remains difficult to compare the performance of different methods or systems and occasionally to understand the results of a validation process. Standardization is also required to perform meta analysis, i.e., comparison of different systems or different types of systems. Furthermore, the standardization of validation processes may be useful in the context of quality management (e.g., FDA or other regulatory body approval). Standardization of validation methodology can be facilitated by common (i.e., standardized) characterization of image processing methods, of the clinical contexts of validation, and of validation procedures. There is a Global Harmonization Task Force (GHTF) that works toward harmonization in medical device regulation (http://www.ghtf.org/).

1) Characterization of Image Processing Methods: Common characterization of image processing methods is done by describing any method in a generic and standardized fashion from the main characteristics of its process. It begins with a standardized description of the process's components.

2) Clinical Contexts in Validation: The two first stages of an HCTA, as described by Goodman, consist in precisely defining assessment topics (i.e., clinical context of validation) and the assessment objective. Just as the development of new image processing tools in medical imaging requires an accurate study of the clinical context, validation of these new tools has to be performed according to this clinical context. Formalization of the clinical context of validation (also referred as the necessity of "full understanding of the problem domain" [5] or "modeling the clinical settings" [15]) is not a trivial task but is essential with regards to clinical relevance. The assessment objective (i.e., goal of the validation study) may be formulated as a hypothesis. The result of the validation process is to reject or not this hypothesis.

The validation hypothesis can be defined from the specifics of the clinical context. Similarly, this hypothesis should be precisely characterized in a standardized fashion. This hypothesis is related to a specific level of evaluation (as defined in Section II) and is defined by the data sets involved in the clinical context and their intrinsic characteristics (e.g., imaging modalities, spatial resolution, and dimensions), by the clinical assumptions related to the data sets or to the patient (e.g., regarding anatomy, physiology, and pathology), and by the values related to validation metrics representing required or expected results (e.g., accuracy or resolution values). In medical image registration, one example of a level 2 validation hypothesis may be: "In the context of temporal lobe epilepsy, a particular registration method M based on similarity measurements is able to register 3-D T1-weighted magnetic resonance images (with a spatial resolution of 2 mm and without any pathological signal) to ictal single photon emission computed tomography (SPECT) (with a spatial resolution around 12 mm and with hyper and/or hypo perfusion areas) with a root-mean-square error (evaluated on points within the brain) that is significantly smaller than the SPECT spatial resolution" [16].

[1] The ground truth may be seen as a conceptual term relative to the knowledge of the truth concerning a specific question. Gold standard may be seen as the concrete realization of the ground truth. The gold standard can be computed or estimated from the validation data sets or from the parameters of the validation procedure. If the gold standard is estimated, it is called bronze standard and provides only an approximated value of the ground truth.

The treatment of some tumors requires highly accurate target localization during a course of fractionated external-beam therapy. Systems that use image-guided localization techniques in the treatment room of the linear accelerator to position patients being treated for cranial tumors using stereotactic radiotherapy, conformal radiotherapy, and intensity-modulated radiation therapy techniques have been developed and validated [17]. Image-guided positioning systems have a critical role in high-precision radiation therapy and therefore special attention must be paid to quality assurance procedures. To ensure accurate treatment delivery, errors in the imaging, localization, and treatment delivery processes must be systematically analyzed. Acceptance tests have been recently developed to validate these systems prior to clinical use [18].

3) Standards for Validation Procedures: The need for protocols for validation was sometimes outlined as the definition of a “unique standardized terminology of validation or evaluation” [5]. The design of models of evaluation processes [13], [14] contributes to this standardization.

We can distinguish the main steps of a gold-standard-based validation procedure as follows. Validation data sets and parameters are used as input by the method to be validated and by the function used to compute the ground truth. Both computations may introduce errors or uncertainties, which have to be taken into account in the comparison. The output of the method is compared to the ground truth for evaluating or validating the method using comparison metrics (i.e., validation metrics). The result of the comparison function provides a quality index also called a “figure of merit” which quantifies distances to the ground truth. The results of the comparison are assessed against the hypothesis of the validation process by means of a simple test on a threshold or statistical analysis. This final result provides the result of the validation (i.e., to reject or not the hypothesis).

Specific statistical approaches have also been investigated to provide validation without recourse to a gold standard (e.g., for studying robustness and internal accuracy of a registration method [19], for comparing quantitative imaging modalities [20]).

B. Validation Data Sets

An important component in any validation is the design of validation data sets, their classification into main families according to the access to the ground truth, and their dissemination through the community. An excellent example of such a dataset is that used by the Vanderbilt group in their retrospective registration evaluation project [21].

Four main types of validation data sets can be distinguished from absolute ground truth to lack of ground truth: numerical simulations; realistic simulations from clinical data sets; physical phantoms; and clinical data sets. The ground truth may be perfectly known, called absolute ground truth (e.g., when using numerical simulations) or may be computed from the data sets (e.g., when using physical phantoms or clinical data sets especially acquired for validation), or finally the ground truth may not be available (e.g., this may be the case when using clinical data sets obtained from clinical routine); in this case the reference for comparison may be given by observers (e.g., manual segmentation versus automatic segmentation) or by some *a priori* clinical knowledge or clinical assumptions. In these last two cases the gold standard is called a bronze standard or fuzzy gold standard. Consequently, the computation of the ground truth may introduce some errors, which must be taken into account in the validation process.

It is quite clear that the different types of data sets provide data for different levels of evaluation. Numerical simulations allow evaluation of the influence associated with various parameters on the method’s performance (e.g., amount and type of noise). But this influence may be overestimated or underestimated. Additionally, it may have functional dependencies between models used to simulate data and models (i.e., assumptions) of the image processing method itself [15]. The realism of the simulated data is difficult to prove and simulated data as well as physical phantoms cannot take into account all the true variability encountered in clinical situations, but they can be very useful when gold standards are difficult, if not impossible, to produce, as is the case in most applications of nonrigid registration. The numerical simulation can also be extremely helpful during system development and integration. By using physical phantoms the whole acquisition set up is taken into consideration. These different types of validation data sets are of complementary nature and study different facets of a method or a system. Therefore, a complete performance evaluation should in principle be performed using each of these different types of data sets whenever this is practical.

Sharing image databases or patient databases helps validation processes and performance comparison, and allows robustness studies. These databases must include “hard” and unusual cases (e.g., pathological cases) and be regularly updated with new imaging protocols, new modalities, and data from new applications. Databases should also include information about images (e.g., characteristics of the subject, such as age and sex, characteristics of the pathology, and clinical history). However, because clinical validation requires clinical image data sets adapted to the local conditions at clinical institutions, the availability of clinical validation data sets will remain difficult until variations among imaging systems can be quantified and normalized [14], [22]. Access to image data bases along with their clinical information could help the PMA applications process but it raises questions about the ownership and credits on the data, about data format and about quality control of this data.

The experimental conditions under which the validation data sets were acquired will determine whether they can be used for more general “effectiveness” studies (i.e., benefit of using a technology for a particular problem under general or routine conditions) or more restricted “efficacy” studies (i.e., benefit of using a technology for a particular problem under ideal conditions) [7].

C. Validation Metrics

The “assessment objective” generally refers to a validation criterion to be studied. Validation metrics and the corresponding mathematical or statistical tools are defined according to the validation criterion. Consequently validation metrics should be chosen or defined according to their suitability to assess the clinical assessment objective. They must be “clinically useful indicators of outcome” [14]. For instance, for accuracy studies in registration, it is now well established that computing or estimating the target registration error (TRE) [23] provides more meaningful information than the fiducial registration error (FRE). The requirement of an overall validation of image-guided surgery systems [1], [2], [4] (i.e., including all its components) should also be taken into account by estimating errors at each stage of the IGT process, and by modeling how errors propagate through the entire IGT process [24]. This allows one to study the influence of each medical image processing component within the overall process.

V. Conclusion

Medical image processing subsystems are key components of IGT systems, and their intrinsic performances are key factors contributing to the overall IGT system. However, their validation is still largely dependent on *ad-hoc* "home made" methodology. As said above, validation of medical image processing methods for IGT should benefit from the definition of common validation data sets and their corresponding ground truth, from the definition of validation metrics adapted to clinical requirements, and finally from the design of common terminology and methodology for validation procedures. Standardized and worldwide-accepted validation protocols with associated guidelines should also facilitate the comparison of new IGT systems and their acceptance and transfer from research to industry. Finally, in the drive towards acceptance of effective and standardized validation methodologies, we must not constrain the creativity of researchers. Existing validation methodologies may well not be suitable to new techniques and if so these methodologies must adapt. The validation process should encourage both innovation and greater sharing of data, results, and methods.

Pierre Jannin, *Guest Editor*
Université de Rennes 1
35000 Rennes, France
Correspondence: (email : pierre.jannin@univ-rennes1.fr)

J. Michael Fitzpatrick, *Guest Editor*
Department of Electrical Engineering and Computer Science
Vanderbilt University
Nashville, TN 37235-1679

David J. Hawkes, *Guest Editor*
Guy's Hospital
King's College London
London SE1 9RT, U.K.

Xavier Pennec, *Guest Editor*
INRIA
06902 Sophia-Antipolis, France

Ramin Shahidi, *Guest Editor*
Stanford University
Stanford, CA 94305-5327

Michael W. Vannier, *Editor-In-Chief*
University of Iowa
Iowa City, IA 52242

References

[1] M. H. Loew, "Medical imaging registration study project," in *Rep. NASA Image Registration Workshop Nov. 1997*, [Online]. Available: http://www.seas.gwu.edu/~medimage/report97.htm.

[2] F. Shtern *et al.*. (1999, April 12–14) Report of the joint working group on image-guided diagnosis and treatment, Washington, DC. [Online]. Available: http://www.nci.nih.gov/bip/IGDT_final_report.PDF.

[3] K. Cleary *et al.*, "Final report of the technical requirements for image-guided spine procedures workshop," *Comput.-Aided Surg.*, vol. 5, no. 3, pp. 180–215, 2000.

[4] R. Shahidi *et al.*, "White paper: Challenges and opportunities in computer-assisted interventions January 2001," *Comput.-Aided Surg.*, vol. 6, no. 3, pp. 176–181, 2001.

[5] K. W. Bowyer, M. H. Loew, H. S. Stiehl, and M. A. Viergever, "Methodology of evaluation in medical image computing," in *Rep. Dagstuhl Workshop*, Mar. 2001, [Online]. Available: http://www.dagstuhl.de/DATA/Reports/01111/.

[6] J. Gee, "Performance evaluation of medical image processing algorithms," in *Proc. SPIE, Image Processing*, K. Hanson, Ed., 2000, vol. 3979, pp. 19–27.

[7] C. S. Goodman. (1998) Introduction to health care technology assessment. Nat. Library Medicine/NICHSR. [Online]. Available: http://www.nlm.nih.gov/nichsr/ta101/ta101.pdf.

[8] D. G. Fryback and J. R. Thornbury, "The efficacy of diagnostic imaging," *Med. Decis. Making*, vol. 11, pp. 88–94, 1991.

[9] P. J. Edwards and A. P. King *et al.*, "Design and evaluation of a system for microscope-assisted guided interventions (MAGI)," *IEEE Trans. Med. Imag.*, vol. 19, pp. 1082–1093, Nov. 2000.

[10] M. Holden, D. L. G. Hill, E. R. E. Denton, J. M. Jarosz, T. C. S. Cox, T. Rohlfing, J. Goodey, and D. J. Hawkes, "Voxel similarity measures for 3-D serial MR brain image registration," *IEEE Trans. Med. Imag.*, vol. 19, pp. 94–102, Feb. 2000.

[11] X. Pennec, C. R. G. Guttmann, and J. P. Thirion, "Feature-based registration of medical images: Estimation and validation of the pose accuracy," in *Lecture Notes in Computer Science*. Berlin, Germany: Springer-Verlag, 1998, vol. 1496, Proceeding of the First International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI'98), pp. 1107–1114.

[12] J. M. Fitzpatrick, "Detecting failure, assessing success," in *Medical Image Registration*, J. V. Hajnal, D. L. G. Hill, and D. J. Hawkes, Eds. Boca Raton, FL: CRC, June 2001.

[13] I. Buvat *et al.*, "The need to develop guidelines for evaluations of medical image processing procedures," *SPIE Med. Imag.*, vol. 3661, pp. 1466–1477, 1999.

[14] T. S. Yoo, M. J. Ackerman, and M. Vannier, "Toward a common validation methodology for segmentation and registration algorithms," in *Lecture Notes in Computer Science*. Berlin, Germany: Springer-Verlag, 2000, vol. 1935, Proc. of Medical Image Computing and Computer-Assisted Intervention (MICCAI 2000), pp. 422–431.

[15] R. P. Woods, "Validation of registration accuracy," in *Handbook of Medical Imaging, Processing and Analysis*, I. N. Bankman, Ed. New York: Academic, 2000, vol. 30, pp. 491–497.

[16] C. Grova , P. Jannin , and A. Biraben *et al.*, "Validation of MRI/SPECT registration methods using realistic simulations of normal and pathological SPECT data," in *Proceedings of Computer Assisted Radiology and Surgery (CARS 2002)*, H. U. Lemke, M. W. Vannier, K. Inamura, A. G. Farman, K. Doi, and J. H. C. Reiber, Eds. Berlin, Germany, 2002, pp. 450–455.

[17] M. H. Phillips, K. Singer, E. Miller, and K. Stelzer, "Commissioning an image-guided localization system for radiotherapy," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 48, no. 1, pp. 267–276, Aug. 1, 2000.

[18] W. A. Tome, S. L. Meeks, N. P. Orton, L. G. Bouchet, and F. J. Bova, "Commissioning and quality assurance of an optically guided three-dimensional ultrasound target localization system for radiotherapy," *Med. Phys.*, vol. 29, no. 8, pp. 1781–1788, Aug. 2002.

[19] S. Granger, X. Pennec, and A. Roche, "Rigid point-surface registration using an EM variant of ICP for computer guided oral implantology," in *Lecture Notes in Computer Science*. Berlin, Germany: Springer-Verlag, 2001, vol. 2208, Proc. of Medical Image Computing and Computer-Assisted Intervention (MICCAI 2001), pp. 752–761.

[20] J. Hoppin *et al.*, "Objective comparison of quantitative imaging modalities without the use of a gold standard," *IEEE Trans. Med Imag.*, vol. 21, pp. 441–449, May 2002.

[21] J. B. West *et al.*, "Comparison and evaluation of retrospective intermodality image registration techniques," *J. Comput.-Assist. Tomogr.*, vol. 21, no. 4, pp. 554–566, 1997.

[22] K. Van Laere *et al.*, "Transfer of normal 99mTc-ECD brain SPET databases between different gamma cameras," *Eur. J. Nucl. Med.*, vol. 28, no. 4, pp. 435–449, 2001.

[23] J. M. Fitzpatrick, J. B. West, and C. R. Maurer, Jr., "Predicting error in rigid-body, point-based registration," *IEEE Trans. Med. Imag.*, vol. 17, pp. 694–702, May 1998.

[24] W. J. Viant, "The development of an evaluation framework for the quantitative assessment of computer-assisted surgery and augmented reality accuracy performance," *Stud. Health Technol. Inform.*, vol. 81, pp. 534–540, 2001.