Validation of medical image processing in image-guided therapy.

Pierre Jannin, J Michael Fitzpatrick, David J. Hawkes, Xavier Pennec, Ramin Shahidi, Michael W. Vannier

Validation of Medical Image Processing in Image-guided Therapy

Pierre Jannin, J.Michael Fitzpatrick, Dave J. Hawkes, Xavier Pennec, Ramin Shahidi, and Michael W. Vannier

***Abstract*—Do we need an abstract ?.**

***Index Terms*— Do we need these key words ? Validation, performance evaluation, image guided therapy, medical image processing.**

I. Introduction

Clinical use of image-guided therapy (IGT) systems has grown this last decade, creating the need for a common and rigorous validation methodology, as reported in recent workshops and conferences [1,2,3,4,5,6]. One key characteristic of IGT systems is that they employ medical image processing methods (e.g. segmentation, registration, visualization, calibration). As a result of this intrinsic structure, validation of IGT systems should include both individual validation of these components, validation of the overall system and a study of how uncertainties propagate through the entire image guided therapy process. Significant progress has been made on IGT system validation recently. Today almost all peer-reviewed publications reporting on the development of new medical image processing methods include a validation section, but 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 to 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 short editorial identifies the principal requirements of IGT system validation and encourages the medical imaging community to develop a common methodology 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, 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, 10) monitor impact.

In a transversal approach, 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), 6) societal impact (e.g. cost effectiveness). An evaluation study must consider only one level at a time but a whole evaluation study should theoretically address all these levels separately.

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 post treatment control. These methods have to 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 require different skills, and they are beyond the scope of the IGT domain that concerns most engineers and physicists.

Manuscript received November 9, 2002. **Do we have to reference here the white paper in the proceedings of CARS ?**

Corresponding author: P. Jannin is with the IDM Laboratory from the Medical School, University of Rennes, Rennes, 35043 Rennes Cedex, France (phone +33-2234588; fax +33-2234586; e-mail: pierre.jannin@univ-rennes1.fr).

J.M. Fitzpatrick is with the Department of Computer Science, Vanderbilt University, Nashville, TN 37235 USA (email: jmf@vuse.vanderbilt.edu).

D. J. Hawkes is with

X. Pennec is with the Epidaure group from INRIA institute, Sophia-Antipolis, France (e-mail:).

R. Shahidi

M. W. Vannier

III. Criteria of 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 computed 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. This value is generally expressed in the parameter space. 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 inter-individual anatomic or physiologic variability.

Both precision and robustness computations may or may not required a ground truth. 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 [9,10,11], by studying the effects of the composition of n transformations that forms a circuit: $T_{n1} \circ \ldots \circ T_{23} \circ T_{12}$. The consistency is a measure of the difference of the composition from the identity. This criterion does not required any ground truth.

Others criteria from algorithmic evaluation could be addressed (e.g. fault detection, code verification, algorithmic proof).

Fault Detection: This is the ability of a method to detect by itself 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). Briefly, the criteria are derived from a legal requirement that the device be show to be safe and effective. If a predicate device exists, the FDA may grant approval (510K) based on substantial equivalence in performance. Otherwise a Pre Market 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.

Others factors may have to be studied but are beyond the scope of this paper such as cost/effectiveness ratio, patient acceptance, and outcome factors.

IV. Validation requirements

The main categories of requirements concerning validation include: standardization of validation methodology, design of validation data sets and validation metrics [1,2,3,4,5,6,12,13,14].

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 even occasionally to really understand the results of a validation process. Standardization is also required to perform meta analysis. Furthermore, the standardization of validation processes may be useful in the context of quality management (e.g. FDA approval). Standardization of validation methodology can be facilitated by common (i.e. standardized) characterisation of image processing methods, of the clinical contexts of validation, and of validation procedures.

1) Characterization of image processing methods

Common characterization of image processing methods allows 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 problem domain" [5] or "modelling the clinical settings" [14]) 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 confirm or not this hypothesis.

The validation hypothesis can be defined from the specificities of the clinical context of validation. Similarly this hypothesis should be precisely characterized in a standardized fashion. This hypothesis is related to a specific level of evaluation (as defined in paragraph 2.) and is notably defined by the data sets involved in the clinical context and their

intrinsic characteristics (e.g. imaging modalities, spatial resolution, 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 1 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 MR images (with a spatial resolution around 2 mm and without any pathological signal) to ictal SPECT (with a spatial resolution around 12 mm and with hyper and/or hypo perfusion areas) with a RMS error (evaluated on points within the brain) that is significantly smaller than the SPECT spatial resolution" [15].

3) Standards for validation procedure

The need for protocols for validation was sometimes outlined as definition of a "unique standardized terminology of validation or evaluation" [5]. The design of models of evaluation processes [12,13] 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)[A]. The result of the comparison function provides a quality index also called "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 threshold or a statistical analysis. This final result provides the result of the validation (i.e. to accept or to reject the hypothesis).

Specific statistical approaches have also investigated validation without gold standard (e.g. for studying robustness and internal accuracy of a registration method [16], for comparing quantitative imaging modalities [17]). These approaches may provide an interesting framework for theoretical validation.

B. Validation data sets

Some of the most commonly mentioned requirements about validation concern 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 [18].

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 vs. 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 uncertainties, which have to be taken into account in the validation process. As it can be noticed, there is a trade-off between clinical realism of the data sets and easy access to ground truth.

It is also quite clear that the different types of data sets provide data for different levels of evaluation. Numerical simulations allow to study the influence of various parameters on the performances of the method (e.g. amount and type of noise). But this influence may be over or under estimated. Additionally, it may have functional dependencies between models used to simulate data and models (i.e. assumptions) of the image processing method itself [14]. Finally the realism of the simulated data is rarely proven and simulated data as well as physical phantoms do not take often into account the variability encountered in clinical situations. By using physical phantoms the whole acquisition set up is taken into consideration but few of them are multimodal by simulating different physical properties. Anyway, these different types of validation data sets are of complementary nature and study different facets of a method or a system. Therefore, a whole performance evaluation should be theoretically performed using each of these different types of data sets.

Sharing image databases or patient databases helps validation processes and comparison of performances, 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. Data bases 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 will not be quantified and normalized [13,19]. 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 defining the validation data sets allow distinguishing effectiveness studies (i.e. benefit of using a technology for a particular problem under general or routine conditions) from efficacy studies (i.e. benefit of using a technology for a particular problem under ideal conditions) [7].

[A] These validation metrics are chosen according to the validation criterion used in the study.

C. *Validation metrics*

The "assessment objective" generally refers to a validation criterion to be studied. Validation metrics and the corresponding mathematical or statistical tools have to be defined according to the validation criterion. Consequently validation metrics have to be chosen or defined according to their suitability to assess the clinical assessment objective. They have to be "clinically useful indicators of outcome" [13]. For instance, for accuracy studies in registration, it is now well established that computing or estimating the Target Registration Error (TRE) [20] 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 uncertainty at each stage of the image guided therapy process, and by modeling how uncertainties propagate through the entire image guided therapy process [21]. This allows to study the influence of each medical image processing component within the overall process.

V. Conclusion

Medical image processing sub-systems are key components of image-guided therapy systems, and their intrinsic performances are key factors of the overall IGT system performance. However, their validation still remains driven through a "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 world wide accepted validation protocols with associated guidelines should also facilitate the comparison of new IGT systems and their acceptance and transfer from research to industry. Nevertheless this standardization should not restrained the creativity of researchers but rather allows better sharing of data, results and methods.

Acknowledgment

References

[1] Loew M H, "Medical Imaging Registration Study Project", Report of NASA Image Registration Workshop November 1997. http://www.seas.gwu.edu/~medimage/report97.htm

[2] Shtern F, Winfield D et al., "Report of the Joint Working Group on Image-Guided Diagnosis and Treatment", April 12-14, 1999 Washington, D.C. http://www.nci.nih.gov/bip/IGDT_final_report.PDF

[3] Cleary K, Anderson J, Brazaitis M, et al., "Final report of the Technical Requirements for Image-Guided Spine Procedures Workshop", April 17-20, 1999, Ellicott City, Maryland, USA. *Comp Aid Surg,* Volume 5, Issue 3180-215, 2000

[4] Shahidi R, Clarke L, Bucholz R D, et al., "White paper: Challenges and opportunities in computer-assisted interventions January 2001", *Comp Aid Surg* Volume 6, Issue 3, 176-181, 2001

[5] Bowyer KW, Loew MH, Stiehl HS and Viergever MA, "Methodology of evaluation in medical image computing", Report of Dagstuhl workshop, March 2001, http://www.dagstuhl.de/DATA/Reports/01111/

[6] Gee J, "Performance evaluation of medical image processing algorithms", Proc. Of SPIE, Image Processing, K. Hanson (Eds), Vol. 3979, 19-27, 2000

[7] Goodman CS, "Introduction to Health Care Technology Assessment", Nat. Library of Medicine/NICHSR, 1998 http://www.nlm.nih.gov/nichsr/ta101/ta101.pdf

[8] Fryback DG and Thornbury JR, "The efficacy of diagnostic imaging", *Med. Decis. Making*, 11, 88-94, 1991

[9] Holden M, Hill DLG, Denton ERE, Jarosz JM, Cox TCS, Rohlfing T, Goodey J, Hawkes DJ, "Voxel similarity measures for 3D serial MR brain image registration", *IEEE Trans Med Imag,* 19(2), 94-102, 2000

[10] Pennec X, Guttmann CRG, and Thirion JP, "Feature-based Registration of Medical Images: Estimation and Validation of the Pose Accuracy", Proc. of First Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (MICCAI'98), Cambridge, USA, *Lecture Notes in Computer Science,* Springer Verlag, Vol. 1496, 1107-1114, 1998

[11] Fitzpatrick JM, "Detecting failure, assessing success", *Medical Image Registration*, Hajnal JV, Hill DLG, and Hawkes DJ ed., CRC Press, June 2001

[12] Buvat I, Chameroy V, Aubry F, et al., "The need to develop guidelines for evaluations of medical image processing procedures", *SPIE Medical Imaging*, 3661, 1466-1477, 1999

[13] Yoo TS, Ackerman MJ, Vannier M, "Toward a common validation methodology for segmentation and registration algorithms", Proc. Of Medical Image Computing and Computer-Assisted Intervention (MICCAI 2000), Pittsburgh, USA, *Lecture Notes in Computer Science*, Springer Verlag, Vol. 1935, 422-431, 2000

[14] Woods RP, "Validation of registration accuracy*", Handbook of Medical Imaging, processing and analysis*, Bankman IN (Eds), Academic Press, 30:491-497, 2000

[15] Grova C, Jannin P, Biraben A, et al., "Validation of MRI/SPECT registration methods using realistic simulations of normal and pathological SPECT data", *Proc. of CARS 2002*, Paris, France, 2002

[16] Granger S, Pennec X, and Roche A, "Rigid Point-Surface Registration Using an EM variant of ICP for Computer Guided Oral Implantology", Proc. Of Medical Image Computing and Computer-Assisted Intervention (MICCAI 2001), Utrecht, The Netherlands, *Lecture Notes in Computer Science*, Springer Verlag, Vol. 2208, 752-761, 2001

[17] Hoppin J, Kupinski M, Kastis G, et al., "Objective Comparison of Quantitative Imaging Modalities Without the Use of a Gold Standard", 17th International Conference Information Processing in Medical Imaging, IPMI 2001, Davis, CA, USA, Insana MF, Leahy RM (Eds.), *Lecture Notes in Computer Science*, Springer Verlag, Vol. 2082, 12-23, 2001

[18] West JB, Fitzpatrick JM, Wang MY, et al., "Comparison and Evaluation of Retrospective Intermodality Image Registration Techniques", *Journal of Computer Assisted Tomography,* 21(4):554-566, 1997

[19] Van Laere K, Koole M, Versijpt, et al.,"Transfer of normal 99mTc-ECD brain SPET databases between different gamma cameras*", European Journal of Nuclear Medicine*, 18(4):435-449, 2001

[20] Fitzpatrick JM, West JB and Maurer CR Jr, "Predicting Error in Rigid-Body, Point-Based Registration", *IEEE Trans Med Imag,* 17(5):694-702, 1998

[21] Viant WJ, "The development of an evaluation framework for the quantitative assessment of computer-assisted surgery and augmented reality accuracy performance", *Stud Health Technol Inform*;81:534-40, 2001