Validation in medical image processing.

Pierre Jannin, Elizabeth Krupinski, Simon K. Warfield

1 Introduction

The increasingly important role of image processing in medicine

Medical imaging is one of our most powerful tools for gaining insight into normal and pathological processes that affect health. The role of image processing in medicine is expanding with the increasing importance of finding ways to improve workflow in reading environments where more images are being acquired in more acquisition modalities. Image processing is playing a crucial role in the maturation of quantitative imaging techniques, such as in functional MRI and diffusion tensor MRI, where visualization of the acquired images alone is insufficient. Image processing, embedded in larger systems and applications, is used more and more extensively in medicine from diagnosis to therapy. Image processing has an important influence on the medical decision making process and even on surgical actions [1]. Therefore, high quality and accuracy are expected [2] and some Working Groups have even been formed and workshops held recently to address the topic of validation [3,4] The performance of image processing methods may have an important impact on the performance of the larger systems as well as on the human observer that needs to analyze all of the available image data and render a diagnostic or therapeutic decision. An emerging focus is the development of imaging biomarkers for drug or therapy response [5], and the development and application of sophisticated image analysis methods in order to improve the accuracy of diagnosis, or to better predict outcomes of disease or treatment and intervention strategies.

Sources of errors or uncertainties are numerous in image processing. Some uncertainties are related to biological variability (e.g., organ, tumor and patient variability), including both normal and pathological variability. Some uncertainties are related to the image acquisition process, such as those due to the limited spatial resolution of the images and the associated partial volume effect, those due to geometric distortion in the images, and those related to the intrinsic data variability (e.g., patient movement during tomographic acquisition). Certain types of uncertainties are common to any image processing method whereas others are specific to the type of processing. Others are related to the human observer interpreting the images and the interaction between the human observer and the image data must be understood, especially as we develop more ways to transform the original data into new presentations. In this issue, image processing includes methods computing a new image from an initial one, computing characteristics and measurements from an image (usually named image analysis) or extracting high-level description from an image (usually named image understanding).

The importance of validation in medical image processing

Validation of medical image processing methods is required to understand and highlight the intrinsic characteristics and behaviour of a method, to evaluate performance and limitations, and eventually to compare these performances with different methods. Validation may also examine the clinical efficacy of a procedure, and estimate its social or economic impact. Consequently, validation helps to clarify the potential clinical applications that a method may serve. Validation of a method is important clinically because a method must not impair the interpretation capability of the clinician by rendering an image that contains too many artefacts or present the data in way that is not interpretable by the clinician. Even if a processing method does not improve diagnostic performance, it still may have significant clinical value if it can reduce the time it takes for the clinician to manipulate or process the image to better visualize the data and render a decision. Similarly, results of validation studies help in improving image-processing performances. Validation has a great potential in increasing the role of medical imaging in medicine, from the development of new therapies to the drug discovery process.

Algorithmic advances in medical image processing are often stimulated by the recognition of the need for an image analysis capability that does not yet exist. The characteristics of the need, such as the ultimate requirements for accuracy or for speed, and the type of images under consideration, provide constraints on the algorithm and on the implementation of the algorithm. Validation strategies then provide the essential assessment by which any particular algorithm and its implementation will be judged as acceptable or unacceptable, given the constraints of the particular image analysis challenge to be addressed. Although algorithm development alone is often the contribution of research in this area, it is not possible to create algorithms that will have a significant impact in clinical practice, without simultaneously considering the validation of the proposed algorithm in the context of the problem constraints.

Challenges in Validation

Further research is needed in validation for medical image-processing as issues concerning validation are numerous. Clinically relevant validation criteria need to be developed. Mathematical and statistical tools are required for quantitative evaluation or for estimating performances in the absence of a suitable "ground truth'', "gold standard'' or other reference standard. Comparison of the performance of different methods requires the use of standardized or at least rigorous terminology and common methodology for the validation process. IEEE has a long history of developing standards in various applications, but to date there is very little in terms of standardizing the validation process in medical image processing. The diversity of problems and approaches in medical imaging contributes significantly to this. However, we are convinced that general frameworks or validation guidelines could be established to improve validation in medical image processing. "*The development of standardized methods to physically characterize sources of uncertainty in the use of imaging as a biomarker would stimulate the development of improved imaging methods and software tools",* says Dr. Laurence Clarke, Branch Chief, Imaging Technology Development, Cancer Imaging Program, National Cancer Institute, Rockville, MD (USA). *"The first step being addressed by NCI and NIBIB is to benchmark the performance of change analysis tools against a validated and standardized reference database, a public resource that contains both image and meta-data collected across different imaging platforms and from both NCI and privately funded clinical trials [6]. However there is a need to engage both the broader scientific and industry community to develop standards for benchmarking image processing and data integration tools for clinical decision making as an international effort to ensure a broad dissemination of software tools within the clinical trials and clinical research communities [5-8]*". Validation data sets with available Ground Truth are required. Comprehension of clinical issues is also required. "*Improving operative image guidance and its safety in minimal invasive surgery requires automatic, fast and accurate data acquisition, processing and modelling in the operation theatre",* says Professor Meixensberger, head of Neurosurgery Department and of ICCAS research institute, University of Leipzig (Germany). *"Therefore, validation and performance evaluation in the clinical context is crucial. Taking into account the clinical context must be part of any evaluation, risk analysis, or error prediction methodology.*"

Validation is rarely the main objective of traditional papers in medical image processing. Innovation usually stands in the image processing method itself and validation is usually addressed only as a section in the paper. However, validation is by itself a research topic where methodological innovation and research are required.

2 Issue content

The editors and reviewers gave particular attention in this Special Issue to contributions describing advances and innovation in the context of validation in medical image processing with close relevance to clinical objectives. It is our hope that readers will consider the importance of validation in their future submissions to TMI and possibly incorporate some of the techniques proposed in the papers in this issue. Properly validated image-processing techniques are more likely to receive clinical acceptance than those that have not been validated, and it is the eventual clinical use of these techniques that provide the final test of their ability to impact patient treatment and care.

Validation is a multi-faceted process as the list of topics noted above indicates, and all of the papers included in this Special Issue touch upon more than one of these topics. In organizing this Special Issue, we started with the basics – creation of image sets for use in validation – and proceeded through the papers with respect to how much reader (i.e., radiologist, surgeon) involvement there is both during validation and eventual clinical implementation. Thus, the first three papers deal with the creation of image sets for various clinical applications. The next three papers are on image registration, which may or may not involve input from the human user. Segmentation of images and image structures may or may not involve input from the human user, but generally does and good segmentation often depends on having validated registration methods in place. Two segmentation papers, one in MRI and one in ultrasound, and one calibration paper follow the registration papers. Finally, there are two papers that deal with validation in the context of image quality, where the human user is most affected and involved the most since image quality affects directly diagnostic performance.

Creation of validation image databases

Use of realistic simulated images for validation is highly relevant. It helps deeply understanding the behaviour of a method in settings close to the clinical reality. Studying clinical realism of these simulated images as well as taking into account inter-patient variability are both crucial. Making such validation images freely available to the community strongly contributes to make easier performances comparison across different processing methods. All these aspects are addressed in the following papers. The paper by Aubert-Broche et al. entitled “20 New Digital Brain Phantoms for Creation of Validation Image Data Bases” introduces the extension of the well-known “BrainWeb” database in order to take into account inter-subject variability. They introduce the method they used for building such images database that includes 20 simulated digital phantoms. Each digital phantom includes 11 fuzzy volumes corresponding to anatomical classes within the brain. These phantoms are publicly available. They can be used for simulation of different modalities including MR (as presented in this paper), PET and SPECT, and for validation of various image-processing methods. The authors demonstrate the clinical realism of their MR simulated images by voxel-wise comparison or by comparing intensities distributed inside anatomical classes with real MR images. Computation of cerebral atrophy from images can be an useful tool for studying and early detecting neurodegenerative diseases. Complexity and high variability of atrophy make it difficult o assess image processing methods aiming at automatic detection. The paper by

Camara-Rey at al. entitled “Phenomenological model of diffuse global and regional atrophy using Finite-Element methods” presents a methodology for realistically simulating brain atrophy in MR images. This method consists in computing a realistic deformation model from both measurements on clinical data and biomechanical models. This method can be applied on different patient images in order to generate a database of realistic images exhibiting cerebral atrophy. New tracers are continuously being introduced for use in PET imaging, but they must be rigorously validated and characterized. In PET ground truth is not generally available so simulated databases are often used for validation of performance and processing algorithms. Reilhac et al. present in “Creation and Application of a Simulated Database of Dynamic [^{18}F]MPPF-PET Acquisitions Incorporating Inter-Individual Anatomical and Biological Variability” the methods they used to create a database of simulated dynamic [^{18}F]MPPF-PET data that included inter-individual anatomical and biological variability. The database has been rigorously evaluated and can be used for validating PET data correction and processing methods. These 3 papers are important contributions for providing publicly available image databases with known ground truth for validation of medical image processing.

Validation of image processing methods

The use of experts in identifying landmarks is common in medical image processing, but due to intra-expert and inter-expert variability, it is often desirable to find an automatic method. The paper by Sanchez Castro et al. entitled “A Cross Validation Study of Deep Brain Stimulation Targeting: From Experts to Atlas-Based, Segmentation-Based and Automatic Registration Algorithms’’ presents a validation study comparing expert performance with non-rigid registration in the task of identifying the subthalamic nuclei. Since the subthalamic nuclei are usually not clearly identifiable in clinical MRI, the issue of an appropriate reference standard is raised. In this work, landmark localization performance is assessed in both a limited test data set where the subthalamic nuclei are clearly visible, and by examination of the influence of alignment of the surrounding anatomy upon the accuracy of localization of the subthalamic nuclei. The validation study carried out by the authors enables them to conclude that automatic localization of the subthalamic nuclei can be achieved with an accuracy not different from that of interactive localization by experts. In “Generalised Overlap Measures for Evaluation and Validation in Medical Image Analysis”, Crum et al. present a framework in which a single figure-of-merit and a complementary measure of error (the Overlap Distance) can be used to capture the extent of non-overlapping parts when registering MR brain images. The process is demonstrated by constructing ground truth for a set of brain atlas images that can then be used to evaluate various segmentation algorithms that others may wish to use for algorithm performance comparisons. Deligianni et al. also deal with registration issues in their paper “Non-Rigid 2D/3D Registration for Patient Specific Bronchoscopy Simulation with Statistical Shape Modelling: Phantom Validation”. This paper proposes and validates a practical 2D/3D registration framework that incorporates patient-specific deformations captured by 3D tomographic imaging and catheter tip electromagnetic tracking. The incorporation of data from the catheter tip tracking reduces the number of parameters that control airway deformation (modelled by an Active Shape Model), significantly simplifying the optimization problem.

It is a well-known problem in medical imaging that human observers are quite variable during manual segmentation of image structures. Automatic contour propagation methods are developed to help overcome the variability of the time-consuming manual process but they need to be validated. Hautvast et al. present in “Automatic Contour Propagation in Cine

Cardiac Magnetic Resonance Images" a unique approach for contour propagation in cine cardiac MR images as well as a validation method based on parameter optimization. They show very nicely that the optimized method can trace contours within the range of the manual drawings. Segmentation of cardiac ultrasound images requires an understanding of speckle statistics at the level of both the transducer and the image. In "Evaluation of Four Probability Distribution Models for Speckle in Clinical Cardiac Ultrasound Images", Tao et al. evaluate a variety of empirical models for first-order statistics for the distribution of grey levels in speckle using real clinical images. The paper provides a realistic method for comparing probability models of ultrasound image speckle and nicely points out some of the problems that arise when using real clinical images to compare and validate models, and provides a few techniques to overcome some of the more common problems.

Geometric calibration is crucial in freehand 3D ultrasound imaging in order to establish the relationship between positions in space and pixels in ultrasound images. Rousseau et al. in "Quantitative Evaluation of Three Calibration Methods for 3D Freehand Ultrasound'' examine strategies for accurately obtaining the rigid transform that maps image coordinates to the coordinate system of a probe-tracking sensor. The paper investigates three calibration methods, with four different quality criteria that reflect different important aspects of imaging accuracy. Phantoms are used to provide a reference standard, the accuracy with which the phantom is known is examined, and the impact of the phantom structure and quality criteria upon the resulting assessment of the calibration methods is investigated. The paper provides important practical guidance as to how best to obtain calibrated 3D freehand ultrasound images.

Validation in the context of image quality

Telemedicine offers valuable specialty diagnostic services to underserved patients in rural areas, but often requires significant image compression to transmit medical images using limited bandwidth networks. The problem is that compressing images such as those for tele-echography can introduce artifacts and reduce diagnostic image quality. Delgorge et al. in their paper "Towards a New Tool for the Evaluation of the Quality of Ultrasound Compressed Images" provide an elegant statistical approach for combining a variety of mathematical criteria based on image features to assess the effects of compressing ultrasound images and utilize an absolute similarity metric to compare performance to the medical expert. It is not always feasible or practical to use human observers in image evaluation studies, and to overcome this there has been significant work in the development and validation of model observers for a number of years. Tisdell and Atkins extend the use of model observers to MRI in their paper "Using Human and Model Performance to Compare MRI Reconstructions", and demonstrate very nicely high correspondence between both types of observers as well as some surprising findings on SNR and lesion detection.

Together, these papers highlight several important characteristics of validation studies:

- Image databases freely available to the image processing community are critical in enabling standardization and objective and unbiased validation;
- Artefacts and human anatomical variability pose an important challenge for medical image analysis and must be accounted for and utilized in validation;
- Application specific validation processes are important and no general purpose validation approach is yet sufficient;

- Using human observers to register images, segment them and to judge image quality once images have been processed in some way poses varying degrees of difficulty. Although input from the human user will likely always be needed to some extent, providing validated tools to reduce the degree of input required is a common theme expressed throughout this Special Issue.

3 Conclusion

Through the creation of this special issue, we hope that IEEE TMI has further contributed to the increasing attention paid to quality in computer-assisted health care, especially when dealing with images. We also hope that this dedicated issue stimulates many more manuscripts, purely dedicated to validation and assessment in their different facets, to be submitted to the journal.

We suggest important areas for future research are:

- The development of standards for terminology, methodology and data sets used in evaluation.
- The ability to create test data sets and evaluation metrics that capture the critical features of important classes of image analysis problems, and so enable generalizable conclusions to be drawn about the efficacy of particular analysis methods.
- The study of cumulative performance and error propagation along complex image processing workflows. Quite often a processing technique is developed and validated for essentially a single point in time, but images are often used in other ways once processed. For example, the performance of computer-aided detection and diagnosis (CAD) schemes generally depends critically on the state of the image data being input to them.
- Extension of validation techniques to other lesion categories and other types of images and/or modalities. Many image processing techniques and thus the approaches used to validate them are often designed for specific lesion types in specific types of images. Ways to generalize these techniques need to be explored.

ACKNOWLEDGMENT

The guest editors would like to thank the editor in chief for giving us the opportunity to outline the importance of validation in medical image processing. They would also like to thank the authors of both accepted and non-accepted papers for their effort in this area and to thank the reviewers who strongly contributed to the level of quality raised by the papers included in this issue. Finally the guest editors would like to thank the invited experts who provided advice and personal feelings about this highly important issue.

REFERENCES

1. Jannin P, Fitzpatrick JM, Hawkes DJ, Pennec X, Shahidi R and Vannier MW. Editorial: Validation of Medical Image Processing in Image-Guided Therapy, IEEE Transactions on Medical Imaging, 2002;21(11):1445-1449.
2. Sonka M, Fitzpatrick JM (Eds). Handbook of Medical Imaging Volume 2. Medical Image Processing and Analysis. SPIE Press, Bellingham, WA, 2000.
3. TTEC 2006, MIE 2006, CARS 2005 European Federation for Medical Informatics, Working Group on Medical Image Processing http://www.efmi-wg-mip.net/ Last accessed July 27, 2006.
4. CARS 2004, CARS 2003, MICCAI 2003, SPIE 2006 http://www.vmip.org Last accessed July 27, 2006.

5. DHHS “New Federal Health Initiative to Improve Cancer Therapy, http://www.fda.gov/oc/mous/domestic/FDA-NCI-CMS.htm Last accessed July 27, 2006.
6. Reference Image Database Resource (RIDER) White Paper: https://imaging.nci.nih.gov/ncia/ Last accessed July 27, 2006.
7. NIST 2006 workshop http://usms.nist.gov/workshops/bioimaging.htm Last accessed July 27, 2006.
8. Trans NIH BECON BISTI report July 2004: http://www.becon.nih.gov/symposium2004.htm Last accessed July 27, 2006.

Dr. Pierre Jannin is a Senior Researcher at the INSERM research institute at the Medical School of the University of Rennes (France). He has more than 15 years experience in designing and developing image-guided surgery systems for neurosurgery. His research topics include image-guided surgery, multimodal imaging data fusion, augmented reality, modelling of surgical procedures, and validation in medical image processing. He has organized many workshops, tutorials and special sessions for international conferences on this last topic. He is the General Secretary of the International Society of Computer Aided Surgery (ISCAS). He has acted as associate editor and reviewer for several journals, and serves on the Editorial Boards of a number of journals in computer assisted surgery.

Dr. Elizabeth Krupinski is a Research Professor at the University of Arizona in the Departments of Radiology and Psychology. She received her undergraduate degree from Cornell and PhD from Temple, both in Experimental Psychology. Her main interests are in medical image perception, assessment of observer performance, and human factors issues. She also is Associate Director of Evaluation for the Arizona Telemedicine Program and carries out a number of studies in this area as well. She is President of the Medical Image Perception Society and serves on the Editorial Boards of a number of journals in both radiology and telemedicine.

Dr. Simon Warfield is an Associate Professor of Radiology at Harvard Medical School, and the Director of the Computational Radiology Laboratory at Children’s Hospital and Brigham and Women’s Hospital in Boston. His research in the field of medical image analysis has focused on methods for quantitative image analysis through novel segmentation and registration approaches, and in real-time image analysis, enabled by high performance computing technology, in support of image guided surgery. The development of algorithms and technology for validation in medical image processing are a focus of his research activity because of their critical role in enabling sophisticated image analysis methods to be brought into practical application.