Chapter 18

Assessment of Image-Guided Interventions

Pierre Jannin and Werner Korb

Abstract

Assessment of systems and procedures in image-guided interventions (IGI) is crucial but complex, and addresses diverse aspects. This chapter introduces a frame-work for dealing with this complexity and diversity, and is based on some of the major related concepts in health care. Six assessment levels are distinguished in IGI. The main phases and components of assessment methodology are described with an emphasis on the specification and the reporting phases, and on the clear initial formulation of the assessment objective. The methodology is presented in a systematic order to allow interinstitutional comparison. Finally, we outline the need for standardization in IGI assessment to improve the quality of systems, their acceptance by surgeons, and facilitate their transfer from research to clinical practice.

18.1 Introduction

The use of image-guided interventions (IGI) may have an important influence on the decision and action-making processes before, during, and after surgery. For this reason, it is crucial to assess IGI systems rigorously. Assessment of IGI belongs to the domain of Health Care Technology Assessment (HCTA), which is defined as the "process of examining and reporting properties, effects and/or impacts of a system" [Goodman 2004]. The objective of the assessment is to increase the quality of an IGI system, to reduce risks of malfunctions or misuses, and to enhance customer and user satisfaction. Developers of IGI systems have the responsibility to assess their systems and make the results widely available. This chapter aims to provide a framework for the assessment of IGI systems with more focus on the engineering side rather than on the clinical side. The authors have gathered major assessment concepts in health care, according to their current knowledge and vision of the domain. Emphasis has been placed on the correct formulation of the assessment objective, and on the report of the assessment method and results. The latter is crucial in assessment as it provides users with proof of

T. Peters and K. Cleary (eds.), *Image-Guided Interventions.*

added value, recommendations for optimal uses, and an indication of possible risks.

The following statements regarding assessment and their application to IGI directly outline the *complexity and diversity of assessment*.

18.1.1 General Assessment Definitions

In general assessment methodology, it is usual to differentiate the concepts of *Verification*, *Validation*, and *Evaluation* [Balci 2003]. In product engineering, verification and validation are distinguished in the following way: verification is the confirmation, by the provision of objective evidence, that specified requirements have been fulfilled [ISO9000:2000], and it involves assessing that the system is built according to its specifications. Validation is the confirmation, by provision of objective evidence, that requirements for a specific intended use have been fulfilled [ISO9000:2000]. It is the assessment that the system actually fulfils the purpose for which it was intended. In software engineering, it also is usual to differentiate verification and validation from evaluation. Evaluation involves assessing that the system is accepted by the end user and that it fulfils its specific purpose.

Efficacy and effectiveness both refer to how well a technology performs to improve patient health, usually measured by changes in one or more pertinent health outcomes. A technology that works under carefully controlled conditions, or with carefully selected patients under the super-vision of its developers, does not always work as well in other settings, or as implemented by other practitioners. In HCTA, *efficacy* refers to the benefit of using a technology for a particular problem under ideal conditions, e.g., within the protocol of a carefully managed, randomized controlled trial involving patients, meeting narrowly defined criteria, or conducted at a center of excellence. *Effectiveness* refers to the benefit of using a technology for a particular problem under general or routine conditions, e.g., by a physician in a community hospital for different types of patients [Goodman 2004]. Beside parameters such as efficacy and effectiveness, *efficiency* can also be assessed by HCTA methods to include costs, economic conditions, and other factors.

18.1.2 Complexity of Procedures and Scenarios

IGI is generally used during *complex procedures*, and/or it makes them more complex, because it often integrates various hardware and software components into complex scenarios. Image processing is used intensively for registration, segmentation, and calibration. Each component is a potential source of uncertainties, which may result in errors. Performance of the whole IGI system strongly depends on the performance of each component. Assessment may, therefore, include the whole IGI system or one or more of

its components. For example, performance and validity of each component may be studied, or the performance and validity of the whole system may be seen as a black-box. Further uncertainties propagating inside the system also may be investigated.

18.1.3 Direct and Indirect Impact of IGI Systems

Diagnostic technologies have an *indirect* impact on surgery, whereas therapeutic technologies have a *direct* one. The impact of IGI may be both direct and indirect as (1) it can provide surgeons with diagnostic images and further information in the OR, and (2) it directly guides the surgeon during surgical performance by emphasizing areas to be targeted or avoided, and showing the trajectories of instruments. Both direct and indirect impacts should be studied.

18.1.4 Interdisciplinary Collaborations

Technical, clinical, social, economical, and ethical aspects are crucial in IGI assessment, which requires *interdisciplinarity* involving clinicians, computer scientists, natural scientists, ergonomists, and psychologists. The result of this is that different roles, languages, motivations, and methods come into play during assessment studies.

18.1.5 Human–Machine Interaction

Human–machine interaction in IGI is embedded in the surgeon–patient–machine triangle (Fig. 18.1).

In this triangle of the surgeon, the patient, and the IGI system, interactions occur between the three components. The surgeon performs surgery on the patient on the basis of a dedicated surgical procedure. The surgeon communicates with the IGI system through the man/machine

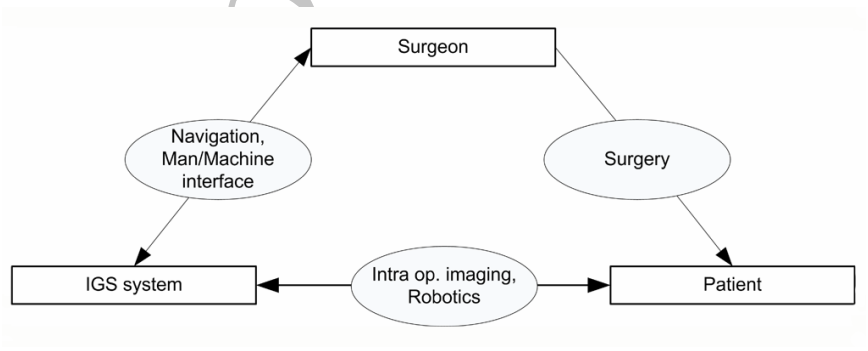

Fig. 18.1 The surgeon–patient–machine triangle in IGI

interface or with the displacement of a surgical tool tracked by a navigation system. In the opposite direction, the IGI system displays images or information to the surgeon. Information about the patient, such as his or her intraoperative images, can be sent to the IGI system. Alternatively, the IGI system may be used to guide a surgical robotic tool.

Table 18.1 The assessment levels for IGI

Levels		Assessed properties	Study conditions	Examples of criteria
Level 1	Technical system properties	Technical parameters	Laboratory	Technical accuracy and precision, latency, noise
Level 2	Diagnostic reliability (indirect assessment)	Reliability in clinical setting	Simulated clinical scenario, laboratory	Sensitivity, specificity, level of quality, level of trust
	Therapeutic reliability (direct assessment)			Target registration error, safety margins, percentage of resection, cognitive workload
Level 3	Surgical strategy (indirect assessment)	Efficacy	Specific clinical scenario, hospital	Change of strategy, time
	Surgical performance (direct assessment)			Cognitive workload, situational awareness, skill acquisition, time, percentage of resection, histological result, pain, usability
Level 4	Patient outcome	Effectiveness	Routine clinical scenario, multisite clinical trials, metaanalysis	Morbidity (recrudescence), pain, cosmetic results
Level 5	Economic aspects	Efficiency	Multisite clinical trials, Metaanalysis	Cost effectiveness, time saving
Level 6	Social, legal, and ethical aspects	Social, legal, and ethical aspects	Metaanalysis, committees, recognized authorities	Quality of life issues

The degree of complexity of human/machine interaction is particularly high as the operation is often performed under extreme (time) pressure, and physical and physiological stress. In many IGI cases, the information flow is considerable, and the relevant data must be processed and condensed by the surgeon's brain. This leads to six categories for the assessment criteria (Table 18.1).

18.2 Assessment Methodology

The complexity and diversity in IGI assessment outlines the importance of using a rigorous methodology for (1) specifying requirements and expected outputs of studies, and (2) precisely reporting objectives, methodology, and output of studies as already mentioned by The Global Harmonization Task Force (GHTF) [GHTF 2004]. This leads directly to the three main steps of assessment of an IGI system (Fig. 18.2).

In the **specification phase**, the *assessment objective* is clearly formulated (phase 1a in Fig. 18.2), the *study conditions* are defined (phase 1b), e.g., setting, criteria, data, as well as relevant assessment methodology, which all fulfil the assessment objective. The specification phase should be able to describe *who* is concerned in the assessment, *what* will be assessed, *what* is expected, in *which* domain or context it is to be assessed, *where* it is assessed, and *which* features will be assessed. Such a specification process results in a better and more manageable design of the assessment study. The *assessment method* is chosen according to the assessment objective and the study conditions. Descriptions of each component of the assessment method are fully included in the specification phase (phase 1c in Fig. 18.2). Good specification of the study phase allows proper measuring and computing as well as statistical analysis of data. Describing and specifying all necessary aspects of the assessment protocol before conducting the study enables correct assessment in accordance with the assessment objective.

In the **implementation phase**, the assessment study is performed (phase 2 in Fig. 18.2). The *assessment method* is strictly applied according to the previously defined *study conditions* for verifying the *assessment objective*. The outcome of phase 2 is the assessment results produced according to the previously defined and properly specified method.

The study documentation is written in the **reporting phase** (phase 3 in Fig. 18.2). The report needs to be well structured and describe assessment objective, study conditions, assessment method, and results to facilitate understanding and comparison of the results between different studies. The specification phase is outlined in the following sections.

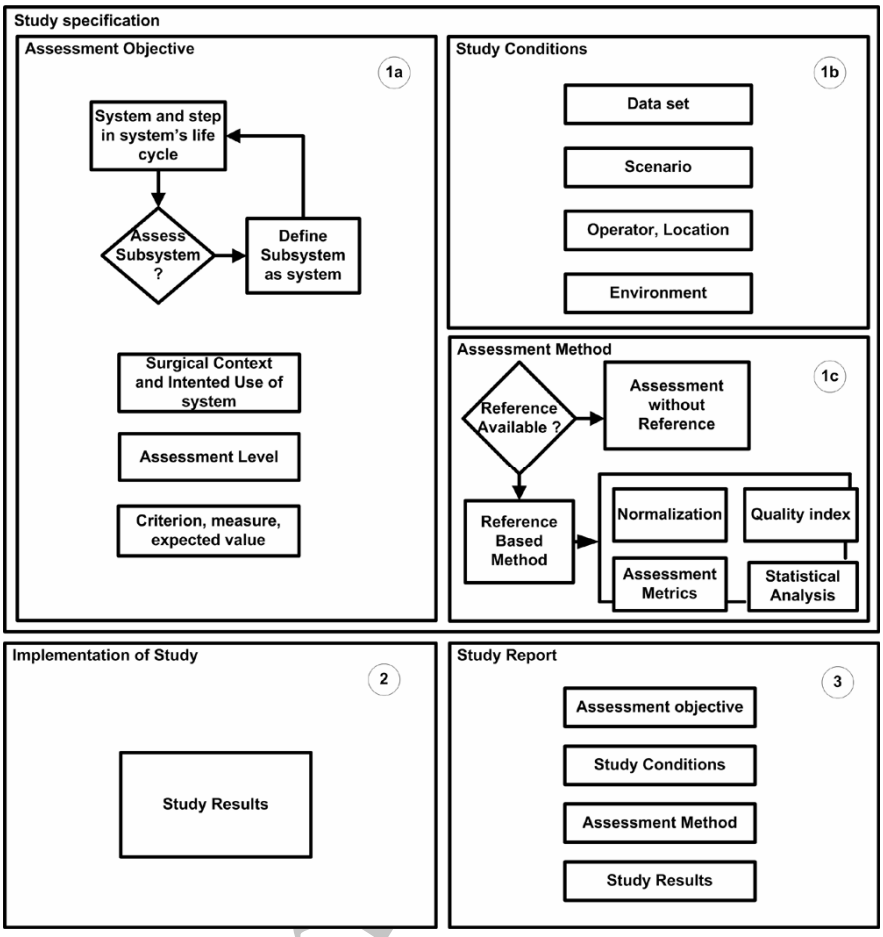

Fig. 18.2. Main phases and components of assessment in IGI

18.2.1 Assessment Objective

Clear and precise design of the *assessment objective* is emphasized as a crucial and required initial step for every assessment study [GHTF 2004; Goodman 2004]. For example, the assessment objective may be formulated as a hypothesis, in which case the assessment study aims to test this hypothesis.

Jannin et al. [2006] proposed a formalization of the assessment objective in medical image processing. Similarly, in IGI, the assessment objective needs to be rigorously formalized and specified before conducting assessment studies. It includes (Fig. 18.2, phase 1a) a precise description of the motivation for assessment, the description of the IGI system to be assessed,

the targeted surgical context and intended use of the system, the corresponding assessment level (see later), criteria to be assessed and corresponding measure, and the expected results or performances. The assessment objective usually consists of comparing measured performances with expected ones for a specific IGI system in specific study conditions, and with dedicated data sets. Performances of an assessment criterion are measured by the assessment metric applied to information extracted from the assessment data sets. Expected performances may correspond to a value or a model. Comparison may be performed using a statistical hypothesis test.

18.2.1.1 Motivation for Assessment

Starting from a high level, it is important to know the targeted consumer of the study, e.g., the developer of the system, the end-user, the manufacturer, an approval body, or a scientific society. Different needs require different methods. At this initial level, the consumer must have expressed his or her motivations and expectations.

18.2.1.2 Description of the System to be Assessed

It is important to clearly specify and describe the system to be assessed. Not only is the performance of the whole IGI system of interest, but also the performance of its components, such as a module for image to patient registration or a tracking cameras as part of a navigation system. Assessment should occur both retrospectively for existing innovative products (e.g., systems or methods), and prospectively during the product life cycle. As mentioned in [GHTF 2004], assessment has to be performed throughout the *product life cycle*. Consequently, the time point along this life cycle and the associated state of the system needs to be described (i.e., whether the system is assessed after specification phase, design phase, or implementation phase).

18.2.1.3 Surgical Context and Intended Use of the System

One essential aspect in the definition of the *assessment objective* is the description of the surgical context in which the IGI system will be used and, therefore, in which the assessment will be performed. The *surgical context* can be considered as a surgical task performed by the surgeon for a targeted population of patients. It is usual to define as many surgical contexts as required and, consequently, to distinguish different corresponding assessment studies. For the chosen surgical context, the system to be assessed is used in a specific manner, which is defined as the *intended use* of the system.

18.2.1.4 Assessment Levels

In HCTA, the complexity and diversity of assessment is usually organised and managed through a hierarchy of levels [Fryback and Thornbury 1991; Goodman 2004; Chow and Liu 2004; Pocock 2004; Korb et al. 2006]. A similar assessment level hierarchy is also relevant and required in IGI (Table 18.1). According to this hierarchy, it is important to decide on the appropriate assessment level during the specification phase of an assessment study.

At **Level 1**, the *technical feasibility and behavior* of the system are checked. Level 1 aims to characterize the intrinsic performance of the system. It is usually the best level at which to verify how technical parameters influence the output of the system. Examples are accuracy or latency investigations, but also may be the impact of noise or the interference of signals or material. It is also the best level for assessing subcomponents of the system independently (e.g., 3D localization, trackers, uninterruptible power supplies, registration components, segmentation modules, or surgical instruments). This level is useful for better understanding possible surgical applications of the system. The clinical realism of the experimental conditions is not crucial, as phantoms and numerical simulations are used for assessment at this level. It often makes sense to perform the technical assessment before doing the assessment studies of Level 2, as Level 1 studies can be considered as verification as defined earlier, and Level 2 studies can be considered as validation as defined earlier.

At **Level 2**, the *diagnostic and therapeutic reliability* is assessed. The technical applicability and reliability of a system for a clinical setting is checked before clinical studies. IGI systems are assessed for their clinical accuracy, patient and user safety, and reliability in a realistic clinical setting context. The methods at Level 2 may include the process of a risk analysis [ISO-14971:2000], as risk analysis is a method to discover the inherent risks of a surgical device or new surgical method. *Diagnostic* and *therapeutic* reliabilities are distinguished as IGI systems may have indirect and direct impact on surgery as discussed earlier. From this level, the realism is increased from the experimental environment to daily clinical environment at Level 3 and above.

At **Level 3**, the efficacy is assessed in clinical trials, including indirect and direct effects. Level 3 studies address assessment criteria dedicated to clinical reality, such as patient outcome, surgical time, or the surgeon's cognitive workload. Level 3 studies and above can be considered as Evaluation. Design and performance of such clinical trials require inter-disciplinary research groups of surgeons, psychologists, ergonomic scientists, and other related scientists.

It may not be straightforward to measure the outcome or added value of the indirect effects of some IGI systems, such as navigation and

intraoperative imaging devices. Indirect effects on the surgical procedure can be assessed either:

1. By performing prospective assessment of therapeutic intent before and after the intervention
2. By asking the surgeons the hypothetical question: "What would you do for the patient, if the information source (navigation or intraoperative imaging) was not available?"
3. By requesting the effect retrospectively from the clinical records
4. By controlled trials in demo-scenarios, i.e., simulations of surgical complications that have to be solved with and without the information provided by the assessed equipment (based on Fryback [1991]).

At **Level 4**, the *effectiveness and comparative effectiveness* are assessed particularly in multisite clinical trials. Facing and addressing clinical assessment of IGI in terms of "large scale multisite randomized clinical trials" (RCT) is difficult for the following reasons: there is a high interpatient and intersurgeon variability that makes large clinical trials difficult; surgical and technological skills are disparate along the population of surgeons; RCTs with randomized assignment of patients mean that some have the opportunity to benefit from a potentially useful technology, which others do not, create a dilemma for surgeons who always want the best possible techniques for all their patients [Paleologos et al. 2000]. Further-more, IGI technologies are still recent and long-term outcome is usually difficult to study. Therefore, the dedicated methodology of "small clinical trials" could be applied in IGI [Evans and Ildstad 2001].

At **Level 5**, the *economic impact* is assessed, on the basis of criteria like cost-effectiveness [Gibbons et al. 2001; Draaisma et al. 2006]. Such evaluations are mainly done by health organizations or cost bearers, based on political requests. For example, one cost analysis in IGI was performed by Gibbons et al. [2001]. This study included the measurement of the costs and benefit of image-guided surgery with an electromagnetic surgical navigation system in sinus surgery.

At **Level 6**, the *social, legal, and ethical impacts* are assessed. The goal of HCTA is also to advise or inform regulatory agencies, such as the US Food and Drug Administration (FDA) or the European Community about whether to permit the commercial use (e.g., marketing) of a system. HCTA also informs standards-setting organizations for health technology and health care delivery about the manufacture, use, and quality of care [Goodman 2004]. HCTA contributes in many ways to the knowledge base for improveing the quality of health care, especially in innovative areas of medicine and surgery, such as image-guided surgery or robot-assisted surgery [Corbillon 2002; OHTAC 2004a,b, National Horizon Scanning Center (NHSC) 2002].

Such reports are usually metaanalyses, based on literature reviews, expert interviews, and local reviews in centers of excellence. These metaanalyses enable governmental organizations to plan future investments, as well as surgeons to consider the use of new technologies in their daily routines.

18.2.1.5 Criteria, Measures, and Expected Values

Surgery with image-guided systems can be seen as a triangle including the surgeon (plus the surgical staff), the patient, and the IGI system (Fig. 18.1). IGI assessment criteria may be characterized along this triangle into six categories:

1. Patient-related criteria, such as clinical scores, functional outcome, pain [Hanssen et al. 2006], cosmetic results [Hanssen et al. 2006], and resection rate
2. Surgeon-related criteria, such as cognitive stress, skill acquisition, and ergonomic working postures [Matern and Waller 1999; van Veelen et al. 2001]
3. IGI system-related criteria, such as technical accuracy and precision of 3D localization, latency, noise, and interference
4. Criteria related to interactions between surgeon and patient, such as time, complexity of procedure, and process-related criteria (e.g., resources, cost effectiveness, complications [Draaisma et al. 2006], reoperation, and change of strategy or planned surgical management [Solomon et al. 1994])
5. Criteria related to the interaction between surgeon and the IGI system, such as human factors [Goossens and van Veelen 2001] (e.g., usability [Martelli et al. 2003], situation awareness [Strauss et al. 2006], hand-eye coordination [Pichler et al. 1996], perception [Crothers et al. 1999; DeLucia et al. 2006], line-of-sight for tracking devices [Langlotz et al. 2006]), and surgical efficiency criteria (e.g., level of trust in a technical system, level of quality, level of reliance, change of strategy) [Strauss et al. 2006]
6. Criteria that express the interaction between the IGI system and the patient, such as clinical accuracy and precision, target registration error [Fitzpatrick et al. 1998; Fitzpatrick and West 2001], and safety margins.

Within each of these six IGI assessment criteria categories, different validation criteria can be used for assessing the various aspects of an IGI system [Jannin et al. 2002], such as *accuracy, precision, robustness, specificity,* and *sensitivity*. For example, the accuracy can be the *spatial accuracy* of a navigation system or the *accuracy of time-synchronization*. Therefore, these terms cannot be included into one specific category only.

Accuracy is defined as the "degree to which a measurement is true or correct" [Goodman 2004]. For each sample of experimental data, local accuracy is defined as the difference between observed values and theoretical ideal expected values. The *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. Close to precision, *reliability* is defined as "the extent to which an observation that is repeated in the same, stable population yields the same result" [Goodman 2004]. The *robustness* of a system refers to its performance in the presence of disruptive factors such as intrinsic data variability, data artifacts, pathology, or interindividual anatomic or physiologic variability. *Specificity* and *sensitivity* are also useful for IGI assessment by computation of ratios between true positive, true negative, false-positive, and false-negative values.

The measures that give values for each criterion are dependent on the targeted categories of assessment criteria and the chosen assessment level. Some criteria and measures may be used for different assessment levels; others are particularly suited for a dedicated assessment level. The evaluation of each criterion needs to be specified and performed separately. Finally, the evaluation of *expected values* includes the definition of what constitutes nonconformance for both measurable and subjective criteria.

18.2.2 Study Conditions and Data Sets

It is crucial to assess the system in experimental conditions that are as close as possible to the surgical context and to the intended use of a system (as outlined earlier). The *Assessment Study Conditions* (i.e. *Assessment Locus*) [Goodman 2004] describe characteristics related to the actual use of the IGI system during assessment studies. Such characteristics include the assessment scenario, location, environment, and data sets. They must be explicitly specified.

18.2.2.1 The Assessment Surgical Scenario

An IGI system is assessed according to a surgical scenario, which mimics its use and includes a list of surgical steps for which the IGI system is assessed. The IGI system can be assessed throughout the whole surgical procedure, or only during single steps of the procedure. Further temporal conditions of the assessment are important. Clinical timing constraints relative to anaesthesia time, for example, are also a crucial characteristic of the surgical procedure and therefore of the assessment surgical scenario.

18.2.2.2 The Assessment Operator, Location, and Environment

Important questions that have to be clarified during the planning phase of the study are as follows:

1. Who will use the IGI system during the assessment study?
2. Where it will be performed?
3. How much of the full clinical environment will be available when performing the study?

18.2.2.3 Assessment Data Sets

The assessment surgical scenario is then applied in an assessment location and environment by an operator using dedicated *assessment data sets*. It is usual to distinguish families of assessment data sets along a continuum between clinical realism and easy control to the parameters to be studied. Along this continuum, three main categories can be identified: numerical simulations, physical phantoms, and clinical data sets (Fig. 18.3). Additional categories are also located along this continuum, such as data acquired from animals or cadavers. Assessment data sets are described by their location along the continuum, by their intrinsic characteristics, such as imaging modalities, spatial resolution or tissue contrast, and by clinical assumptions related to the data sets or to the patient, such as assumptions regarding anatomy, physiology, and pathology.

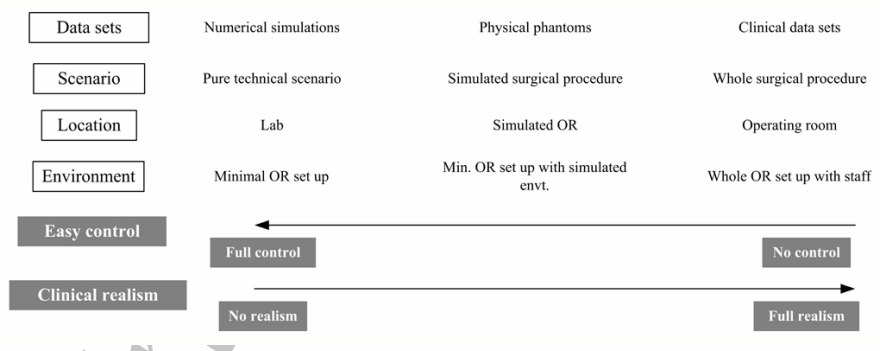

Fig. 18.3. Continuum for assessment data sets, scenario, location, and environment from full control of parameters to full clinical realism

A similar continuum can be used for describing the other study conditions: assessment surgical scenario, location, environment, operator, and temporal conditions. As shown in Fig. 18.3, there is usually a trade-off between control of parameters and clinical realism. Choice of a solution along these continua for all the categories are strongly related to the *assessment levels* as defined above.

18.2.3 Assessment Methods

The main prerequisite of a study is the proper definition of the *assessment objective* and the *study conditions*. This means that for each component as defined earlier (Fig. 18.2), a dedicated *value* (e.g., level, criterion, measure, data sets) is selected. Then a method is chosen to fulfil the assessment objective. The selection of an appropriate method will primarily depend on the assessment objective and secondarily on the study conditions. For some studies, it is possible to choose a reference for assessment (a gold standard), but for others it is impossible. This leads to a categorisation into (1) reference-based assessment and (2) assessment without reference.

18.2.3.1 Reference-Based Assessment

Reference-based assessment compares the direct or indirect results of a system with a reference (also called a *gold standard*)[1] that is assumed to be very close or equal to the ideal expected solution (also called the *ground truth*). Jannin et al. [2006] proposed a model for describing and reporting such reference-based assessment methods for medical image processing. This model is also useful for IGI when assessed criteria require such comparison with a reference (e.g., accuracy).

The main components and the main stages of this model are described as follows. The IGI system to be assessed is used according to the study conditions (surgical scenario, temporal conditions, operator, location, environment, and data sets). Characteristics of the system output are compared with ideal expected results. As ideal expected results are usually not directly available, another method can be used according to the study conditions to provide results that are closer to the ideal expected results than those computed by the system itself. These results are considered to be the *reference* against which results from the system will be compared. Assessment is a comparison of characteristics of both results, with the characteristics being chosen according to the targeted assessed criteria. The reference is generally chosen to be as accurate as possible, but in some situations it may have an error that should be taken into account during the assessment process or at least in the assessment results. The quantitative comparison may require converting those results in a similar format, which may be seen as a *normalization step*.

[1] In this chapter, we use the term *reference* rather than *gold standard*. *Reference* can be used in a wider meaning including terms of *gold*, *bronze*, or *fuzzy* standards.

This normalization aims at transforming the measurements of the assessed criteria in a clinically meaningful format. Normalized results, computed by the system to be assessed, and the reference method are compared using a *comparison function,* also called the *Assessment Metric*. Statistics of the comparison results distribution may serve as a *quality index*, also called a *figure of merit*. Finally, values of the quality indices are compared with expected values or models using a *statistical hypothesis test*. This consists of testing quality indices computed on comparison results against the assessment hypothesis to provide the assessment result.

To avoid mistakes or bias in reference-based assessment methods, the following aspects have to be carefully checked. The relevance of the data sets used in assessment studies must be verified according to two aspects: the realism of the data sets, and the coherence between the data sets and the assessment objective. The reference usually comes from one of the following methods:

1. It can be an exact and perfect solution computed from numerical simulations
2. It can be an estimated solution from the results of one or several reference methods
3. It can be another estimated solution from the results of the same assessed system but used with different data sets or conditions
4. It can be an expert-based solution relying on assumptions, or on a priori knowledge of the results.

In all cases, the quality of the reference has to be checked according to two aspects: correctness and realism. As the reference is also computed from the reference method, it is an approximated value of the ideal expected results only. A study of the error associated with the reference is crucial.

The comparison function can be considered as an assessment metric, as it quantifies criteria of assessment. This function has to be chosen or defined according to its suitability to fulfil the assessment objective.

Training and testing methods are also usually used for validation, such as the leave-one-out method. They could be considered as reference-based validation methods, since the result of the system when using the training set is compared to the result of the system when using the testing set. However, in such methods there is no evidence that the reference is close to the ideal expected results (ground truth). These methods can show the independence of the results according to the training set only. Robustness aspects can be checked with training and testing methods, but not accuracy.

18.2.3.2 Assessment Without Reference

Some assessment criteria or some assessment objectives do not require any reference. For example, we can mention the consistency criterion for image

registration and the Bland and Altman [1986] method for measurements. However, it is hard to define assessment objectives in clinical terms when using such assessment methods, as they mainly characterize intrinsic behavior of a method only. They allow performance evaluation only, as no comparison with ideal expected results is performed. Finally, some assessment without reference methods relies on strong assumptions on the data sets or the assessed methods, which cannot always be verified easily.

18.2.3.3 Statistics

Statistics are crucial in assessment methodology, as assessment is usually performed with multiple studies, multiple data sets, or multiple sites. Finding relevant and appropriate statistical tests is not always straightforward. Statistical hypothesis tests usually rely on strong assumptions about the data that have to be carefully checked. Because of the importance of correct statistical analysis, the assessment team should incorporate statistical skills with knowledgeable partners.

18.3 Discussion

The complexity and diversity of assessment for IGI have been emphasized in this chapter. However, different tools and models are available to deal with this complexity. On the one hand, the effort that is performed in the specification phase of an assessment study, together with a clear and precise definition of the *assessment objective,* also helps to deal with the complexity. The diversity, on the other hand, can be managed with the concept of *assessment levels* in IGI.

The suggested tools, phases, components for each phase, and classifications are not exhaustive. There is no strict sequential structure inside each phase, and there are some dependencies and relationships between components in Fig. 18.2. Also, the assessment levels organization in Table 18.1 may not always be strictly hierarchical.

Obviously, the use of the presented tools does not guarantee the quality of the IGI system; rather it should guarantee the quality of the assessment study and provide a correct understanding and analysis of the assessment results. For the latter, the same rigorous methods that were presented should be used for *assessing* the applied *assessment method* itself. This should include the assessment of its associated components (e.g., criteria, statistical tests). Furthermore, different types of validity [Nelson 1980] (such as face validity, content validity, criterion validity, and construct validity) need to be assessed.

Other aspects close to assessment are not covered in this chapter but are also of great interest for IGI. Risk analysis is the assessment of risk according to a defined methodology. The surgical context and intended use are as important as in assessment studies. On the basis of the specification of these characteristics of an IGI system, the possible hazards are identified and estimated. The estimation is based on the criteria level of severity, occurrence probability, and detection probability. On the basis of these values that are estimated in *risk analysis,* different methods for *risk management* can be performed [Korb et al. 2005].

In this important, diverse, and complex landscape, there is much room for innovation and research. Some directions will now be mentioned. There is still a need for new assessment metrics adapted and relevant to a dedicated surgical context. There is a need for realistic and controllable study conditions, from data sets, surgical scenario, environment, and location. Another important aspect in the assessment methodology for correct dissemination and reproducibility of studies and results is the availability of open source data and tools. Such an open source environment will further facilitate assessment.

Finally, for all assessment aspects, there is a great need for standardization, both for specification, implementation, and reporting assessment. In health care technology assessment, some standards have been recently introduced for reporting of clinical trials, e.g., the *Standard for Reporting of Diagnostic Accuracy* (STARD), The *Grading of Recommendations Assessment, Development and Evaluation* (GRADE) Working Group, and the Current Controlled Trials Ltd. (a part of the Science Navigation Group of companies, which hosts the *International Standard Randomized Controlled Trial Number* (ISRCTN) Register). There is also a standard available for the specification and implementation of clinical patient trials for medical devices [ISO 14155:2003]. Some standards can be directly applied to IGI, some need to be adapted, and some additional ones are required. Standardization in IGI assessment will improve the quality of systems, their acceptance by surgeons, and facilitate their transfer from research to clinic practice.

As mentioned at the beginning of this chapter, each person involved in the development or use of an IGI system should be involved in assessment, at least in a specific assessment level. However, assessment studies can be tedious and difficult, requiring time, energy, and motivation. Results are not always as expected and biases are numerous. Such biases cover the spectrum from the specification of the study to the analysis of the results and can be hidden traps, sometimes requiring restarting the study from the beginning. But rigorous assessment is the only way to develop useful, relevant, and valuable tools for the patient and for society. As we cannot escape assessment, the best way forward is to make it as easy and efficient as possible.

Acknowledgments

The authors thank B. Gibaud, C. Grova, and P. Paul from Visages and M. Audette, O. Burgert, A. Dietz, E. Dittrich, V. Falk, R. Grunert, M. Hofer, A. Klarmann, S. Jacobs, H. Lemke, J. Meixensberger, E. Nowatius, G. Strauss, and C. Trantakis from ICCAS (The Innovation Center for Computer Assisted Surgery) for fruitful discussions. They also thank participants and speakers of dedicated special sessions and workshops on this topic during the Computer Assisted Radiology and Surgery (CARS) conferences. Finally the authors thank S. Duchesne and M. Melke for their help during the preparation of the manuscript.

References

Balci O (2003). "Verification, validation and certification of modeling and simulation applications". In: *Proceedings of the 35th conference on Winter Simulation: Driving innovation*, New Orleans, Louisiana, 150–158.

Buxton MJ (1987). "Problems in the economic appraisal of new health technology: The evaluation of heart transplants in the UK". In *Economic Appraisal of Health Technology in the European Community*. Drummond MF, ed, Oxford Medical Publications, Oxford England.

Bland JM and Altman DG (1986). "Statistical methods for assessing agreement between two methods of clinical measurement". *Lancet*, 1, 307–310.

Chow S-C and Liu JP (2004). *Design and Analysis of Clinical Trials: Concepts and Methodologies*, Wiley, Hoboken, New Jersey, ISBN 0-471-24985-8.

Corbillon E (2002). "Computer-assisted surgery progress report." *ANAES*, Saint-Denis La Plaine.

Crothers IR, Gallagher AG, McClure N, James DTD, McGuigan (1999). "Experienced laparoscopic surgeons are automated to the "fulcrum effect": An ergonomic demonstration." *Endoscopy*, 31(5), 365–369.

DeLucia PR, Mather RD, Griswold JA, Mitra S (2006). "Toward the improvement of image-guided interventions for minimally invasive surgery: Three factors that affect performance." *Hum Factors*, 48(1), 23–38.

Draaisma WA, Buskens E, Bais JE, Simmermacher RKJ, Rijnhart-de Jong HG, Broeders IAMJ, Gosszen HG (2006). "Randomized clinical trial and follow-up study of cost-effectiveness of laparoscopic versus conventional Nissen fundoplication." *Br J Surg*, 93, 690–697.

Evans CH and Ildstad ST (2001). *Small Clinical Trials: Issues and Challenges*. Institute of Medicine, National Academy Press, Washington DC.

Fitzpatrick JM, West JB, Maurer CR, Jr (1998). "Predicting error in rigid-body, point-based registration." *IEEE Trans Med Imag*, 17, 694–702.

Fitzpatrick JM and West JB (2001). "The distribution of target registration error in rigid-body point-based registration." *IEEE Trans Med Imag* 20(9), 917–927.

Fryback DG and Thornbury JR (1991). "The efficacy of diagnostic imaging." *Med Decis Making*, 11, 88–94.

AQ: This reference is not cited anywhere in the text. Kindly insert its citation or delete it from the list of references.

Gibbons MD, Gunn CG, Niwas S, Sillers MJ (2001). "Cost analysis of computer-aided endoscopic sinus surgery." *Am J Rhinol*, 15(2), 71–75.

Global Harmonization Task Force, Quality Management Systems (2004). "Process validation guidance" *GHTF*/SG3/N99-10: http://www.ghtf.org/sg3/inventorysg3/sg3_fd_n99-10_edition2.pdf [Accessed September 2007].

Goodman CS (2004). "Introduction to health care technology assessment." *Nat. Library of Medicine/NICHSR*: http://www.nlm.nih.gov/nichsr/hta101/ta101_c1.html [Accessed September 2007].

Goossens RHM and van Veelen MA (2001). "Assessment of ergonomics in laparoscopic surgery." *Min Invax Ther Allied Technol*, 10(3), 175–179.

Hanssen WEJ, Kuhry E, Casseres, Herder WW de, Steyerberg EW, Bonjer HJ (2006). "Safety and efficacy of endoscopic retroperitoneal adenalectomy." *Br J Surg*, 93, 715–719.

ISO 14155-1+2 (2003). "Clinical investigation of medical devices for human subjects. Part 1+2."

ISO 14971 (2001). "Medical devices – Application of risk management to medical devices."

ISO 9000 (2000). "Quality management systems – Fundamentals and vocabulary. International organization for standardization."

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 Imag*, 21(11), 1445–1449.

Jannin P, Grova C, Maurer C (2006). "Model for designing and reporting reference based validation procedures in medical image processing." *Int J Comput Assist Radiol Surg*, 1(2)2, 1001–1115.

Korb W, Kornfeld M, Birkfellner W, Boesecke R, Figl M, Fuerst M, Kettenbach J, Vogler A, Hassfeld S, Kronreif G (2005). "Risk analysis and safety assessment in surgical robotics: A case study on a biopsy robot." *Minim Invasive Ther*, 14(1), 23–31.

Korb W, Grunert R, Burgert O, Dietz A, Jacobs S, Falk V, Meixensberger J, Strauss G, Trantakis C, Lemke HU, Jannin P (2006). "An assessment model of the efficacy of image-guided therapy." *Int J Comp Assist Radiol Surg*, 1, 515–516.

Langlotz F, Kereliuk CM, Anderegg C (2006). "Augmenting the effective field of view of optical tracking cameras – A way to overcome difficulties during intraoperative camera alignment." *Comput Aided Surg*, 11(1), 31–36.

Martelli S, Nofrini L, Vendruscolo P, Visani A (2003). "Criteria of interface evaluation for computer assisted surgery systems." *Int J Med Informat*, 72, 35–45.

Matern U and Waller P (1999). "Instrument for minimally invasive surgery: Principles of ergonomic handles." *Surg Endosc*, 13, 174–182.

Medical Device Directive, Council Directive 93/42/EEC20 of 14 June 1993 concerning medical devices. *European Community, Official Journal* L 169, 1–43.

National Horizon Scanning Centre (NHSC), The University of Birmingham: Surgical Robots.Update (2002) http://www.pcpoh.bham.ac.uk/publichealth/horizon/PDF_files/2002reports/Robots Update.pdf and http://www.pcpoh.bham.ac.uk/publichealth/horizon/PDF_files/2000reports/Surgical_robots.PDF [Accessed September 2007].

Nelson AA (1980). "Research design: Measurement, reliability and validity." *Am J Hosp Pharm*, 37, 851–857.

AQ: This reference is not cited anywhere in the text. Kindly insert its citation or delete it from the list of references.

OHTAC Recommendation (2004a). "Computer assisted hip and knee arthroplasty: Navigation and robotic systems."http://www.health.gov.on.ca/english/providers/program/mas/tech/reviews/pdf/rev_arthro_020104.pdf [Accessed September 2007].

OHTAC Recommendation (2004b). "Computer assisted surgery using telemanipulators."http://www.health.gov.on.ca/english/providers/program/mas/tech/reviews/pdf/rev_teleman_020104.pdf [Accessed September 2007].

Paleologos TS, Wadley JP, Kitchen ND, Thomas DGT (2000). "Clinical utility and cost-effectiveness of interactive image-guided craniotomy: Clinical comparison between conventional and image-guided meningioma surgery." *Neurosurgery*, 47(1), 40–48.

Pichler C von, Radermacher K, Rau G (1996). "The state of 3D technology and evaluation." *Min Invax Ther Allied Technol*, 5, 419–426.

Pocock SJ (2004). *Clinical Trials: A Practical Approach*, Wiley, New York, ISBN 0-471-90155-5.

Solomon MJ, Stephen MS, Gallinger S, White GH (1994). "Does intraoperative hepatic ultrasonography change surgical decision making during liver resection?". *The Am J Surg*, 168, 307–310.

Strauss G, Koulechov K, Röttger S, Bahner J, Trantakis C, Hofer M, Korb W, Burgert O, Meixensberger J, Manzey D, Dietz A, Lüth T (2006). "Evaluation of a navigation system for ENT with surgical efficiency criteria." *Laryngoscope*, 116(4), 564–572.

Van Veelen MA, Meijer DW, Goossens RHM, Snijders CJ (2001). "New ergonomic design criteria for handles of laparoscopic dissection forceps." *J Laparoendosc Adv Surg Tech*, 11(1), 17–26.