The need to develop guidelines for the evaluation of medical image processing procedures

Irène Buvat*, Virginie Chameroy*, Florent Aubry*, Mélanie Pélégrini*, Georges El Fakhri*, Céline Huguenin*, Habib Benali*, Andrew Todd-Pokropek*°, Robert Di Paola*
* U494 INSERM, 75634 Paris, France, ° UCL, London WC1E 6JA, UK

ABSTRACT

Evaluations of procedures in medical image processing are notoriously difficult and often unconvincing. From a detailed bibliographic study, we analyzed the way evaluation studies are conducted and extracted a number of entities common to any evaluation protocol. From this analysis, we propose here a generic evaluation model (GEM). The GEM includes the notion of hierarchical evaluation, identifies the components which have always to be defined when designing an evaluation protocol and shows the relationships that exist between these components. By suggesting rules applying to the different components of the GEM, we also show how this model can be used as a first step towards guidelines for evaluation.

Keywords: evaluation, medical image processing, efficacy, guidelines.

1. INTRODUCTION

Medical imaging generates a huge number of work and associated publications regarding new imaging devices and methods for image processing, image analysis or image interpretation. Several approaches are often available to deal with a given problem (e.g., image reconstruction in tomographic imaging), which leads to an increasing demand by researchers, companies or health authorities for evaluation and comparative assessment of the available imaging or processing methods. Yet, a survey of the literature shows that rigorous evaluation of the proposed procedures is often omitted. Moreover, when evaluation is performed, lack of details regarding the evaluation protocol frequently prevents from drawing sound conclusions regarding the performance and the practical implications of the method under investigation. One reason for this is the absence of clearly established and well codified rules for designing an evaluation protocol, such as rules that are commonly adopted when designing clinical trials. This absence of rules: 1) makes it difficult to assess the validity of a given evaluation study; 2) hinders the comparison of results obtained in a specific evaluation study with results previously reported in the literature; 3) prevents from performing quantitative systematic reviews, also called meta-analyses, aiming at combining data from different evaluation studies to establish the value of a method.

The need for some regulations pertaining to the evaluation of medical image processing methods has already been underlined and efforts have been pursued to create procedures for quality assessment (e.g., [1-4]). Yet, while there is a rapid increase in the development of medical image processing procedures, there is still no accepted norm for their assessment.

From a detailed bibliographic study, we have tried to analyze how evaluation studies are conducted and why they often yield unconvincing results. The purpose of this contribution is to share the results of this analysis and to suggest some preliminary directions (that we will express as rules) to better codify the way evaluation studies are performed. We do not here intend to solve the issues involved in the set-up of evaluation studies, but to assist any investigator in designing an evaluation protocol appropriate to his problem by identifying the questions that should be answered before starting an evaluation study. In addition, we will try to draw attention upon the potential caveats underlying evaluation studies to permit a more critical analysis of results published in the literature. More specifically, the purpose of the paper is to:

1) present a hierarchical structure for evaluation in medical image processing (section 2).

2) describe the framework we have developed for evaluation in medical image processing, that we will call a generic evaluation model, or GEM (sections 3 and 4). This framework synthesizes the components that have to be defined when designing an evaluation protocol and indicates the relationships and constraints between these components.

3) draw attention upon the potential biases that might affect the validity of an evaluation study and the applicability of the conclusions derived from it (section 4).

Whenever possible, the key points of our analysis will be expressed as rules. We think that a better understanding of these basic rules could be a first step towards evaluation guidelines that would facilitate the design of rigorous and reproducible evaluation studies, from which the value of medical image processing methods could be more soundly established.

2. A HIERARCHICAL FRAMEWORK FOR EVALUATION IN MEDICAL IMAGE PROCESSING

Evaluation is the process of judging the value of a method, here a medical image processing procedure. Evaluation in medical image processing is intrinsically difficult because of the large variety of quality assessment this can include. A key point to minimize confusion regarding evaluation is therefore to precisely define the nature of the evaluation study to be conducted. Evaluation is in fact a process including several stages, each one involving a specific investigation leading to a specific type of information regarding the method to be evaluated. In the literature dedicated to the evaluation of imaging technologies [5], six levels of evaluation studies have now gained wide acceptance. We show that such a hierarchical classification can actually also be used for evaluation of medical image processing.

• In the evaluation of imaging technologies, level 1 is called evaluation of technical efficacy. In medical image processing, this could be called validation or feasibility study and correspond to demonstrating the feasibility of the image processing method to be evaluated (what we will call "method" in the following) and the relevance of the results it produces. It is a prerequisite to any further evaluation study.

Example 1: showing that a method of attenuation correction in SPECT allows one to restore the homogeneity of the tracer distribution in the left ventricular wall of a normal subject is a level 1 study.

Example 2: showing that an algorithm for automatic detection of microcalcifications in X-ray mammography actually detects microcalcifications in mammograms is a level 1 study.

In medical image processing, this is the level which is most often considered. This stage can itself include two types of investigation: 1) demonstrating that the method produces "reasonable" results with respect to a priori knowledge regarding the expected results; 2) determining the optimal parameters for the method in a specific context (the notion of context will be detailed below), i.e., the parameters that will yield the "best" results with respect to some criteria or to the adequacy to some "ideal" results.

• In the evaluation of imaging technologies, level 2 is called evaluation of the diagnostic accuracy efficacy of the method. In medical image processing, this could be called evaluation of the method accuracy and it corresponds to measuring precisely the performance of the method for the problem it is intended to solve. This can include characterizing the performance of the method and comparing the performance of n methods (n=2).

This stage of evaluation is addressed more or less thoroughly in medical image processing. Note that, as for evaluation of level 1, it does not necessarily involve real subjects, but can be conducted using simulations or phantom acquisitions.

Example 1: comparing several attenuation correction methods in SPECT using Monte Carlo simulations or phantom data is a level 2 evaluation study.

Example 2: studying the sensitivity and the specificity of an algorithm for automatic detection of microcalcifications in X-ray mammography using images of normal breast parenchyma in which simulated microcalcifications have been added is a level 2 evaluation study.

• Level 3 is called evaluation of the diagnostic thinking efficacy in evaluation of imaging technologies. In medical image processing, this level can be called identically and corresponds to measuring whether the method is judged helpful by the clinician when making its diagnosis, either because it helps him formulate a more appropriate diagnosis, or because it helps him give a diagnosis with more confidence, or because it saves him time when he makes a diagnosis. Note that it is possible

that a method performs well at level 2 but not at level 3. Indeed, a method can be accurate (and therefore pass the level 2), but can have no impact upon the interpretation of the images.

Example 1: a level 2 study can demonstrate that an attenuation correction method in SPECT improves the uniformity of tracer distribution in normal subjects but a level 3 study can show that it does not help the interpretation of the images of normal subjects by the clinicians because the clinicians are highly experienced in reading "through" attenuation artifacts.

Example 2: determining whether a radiologist finds it helpful to observe the results of an algorithm for automatic detection of microcalcifications in X-ray mammography when interpreting mammograms is a level 3 study.

• In evaluation of imaging technologies, level 4 is the evaluation of the therapeutic efficacy of the imaging test. In medical image processing, evaluating the therapeutic efficacy of a method is showing that the method provides information which contributes to the appropriateness of the patient management. Therefore, this involves comparing diagnoses made without and with the help of the method to be evaluated. Compared to level 3, this evaluation stage does not concern the impact of the method as subjectively assessed by the observer but measures the value of the method by considering *independent* information making it possible to objectively judge the relevance of the diagnosis.

Example 1: demonstrating that an attenuation correction method in SPECT improves the sensitivity and the specificity of the detection of coronary artery disease, as detected by coronarography, is a level 4 study.

Example 2: determining whether an algorithm for automatic detection of microcalcifications in X-ray mammography detects microcalcifications that are not seen by the radiologist but that are actually present in the breast is a level 4 study.

• Level 5 is evaluation of the patient outcome efficacy. This corresponds to determining how the method affects patient outcome. Studies of this level aim at determining whether results from studies of level 4 (regarding diagnosis and therapy) will have repercussions on the outcome of the patient.

Example 1: a level 5 study regarding attenuation correction in SPECT would be to determine whether the use of attenuation correction actually avoids unnecessary surgery in patients that would have been otherwise wrongly diagnosed hence operated) or whether the therapeutic management is in fact dominated by results from complementary examinations so that attenuation correction has no impact upon the patient outcome.

Example 2: showing that using an algorithm for automatic detection of microcalcifications in X-ray mammography results in earlier detection of microcalcifications which translates into an increase of the percentage of "cured" patients is a level 5 study.

• Level 6 is the evaluation of the societal efficacy. In medical image processing, a level 6 study would therefore aim at examining the benefit of the method for the society as a whole. It obviously includes cost-effectiveness considerations.

Example 1: a level 6 study for attenuation correction in SPECT would be to determine whether attenuation correction in SPECT leads to a decrease of useless examinations in patients suspected of coronary artery disease.

Example 2: regarding an automatic algorithm for the detection of microcalcifications in mammograms, a level 6 study could concern the role of the algorithm in the throughput of woman screening.

Evaluations of levels 5 and 6 are very rarely considered in medical image processing because the rapid changes in processing methods and technology contrast with the long time such evaluations would take to complete. Because of these rapid changes, there is also a high risk the results of such study become rapidly obsolete.

This hierarchical analysis yields us to a first rule for designing an evaluation protocol in medical image processing.

RULE 1: Before starting an evaluation study, care should be taken to consider only one level of evaluation at a time. This would avoid confusion regarding what the study is supposed to demonstrate and would also avoid ambiguous conclusions or unfounded extrapolation of the results.

3. A GENERIC EVALUATION MODEL (GEM)

Designing a thorough evaluation protocol requires the specification of a number of entities such as, for instance, the data that will be processed by the method to be assessed. Reviewing the different evaluation approaches used in the literature, including, among others, visual assessment, methods extracting some quantitative measures comparing before and after, and task-based methods (e.g., 2 alternative forced choice experiments), we extracted a number of entities common to any evaluation protocol [6]. We found that these components and the relationships between them could be described by what we will call the generic evaluation model (GEM) shown in Figure 1. The purpose of the GEM is to identify the components which have to be defined when designing an evaluation protocol and to explicitly show the relationships that exist between these components. We now describe the GEM and introduce its different components.

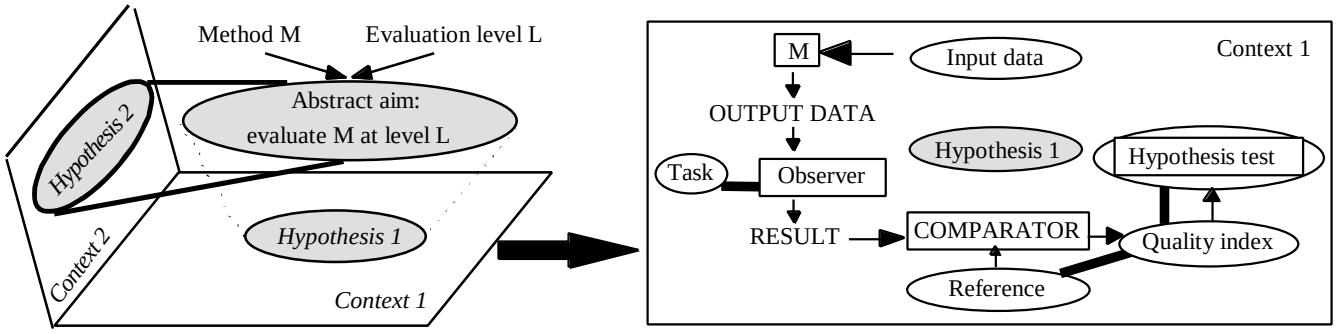

Figure 1a

Figure 1b

Figure 1: The generic evaluation model (GEM). a) a method M combined with the considered level of evaluation L yields an abstract aim. A given abstract aim can translate into different concrete hypotheses depending on the context it is projected onto. b) arrangement of the component of the GEM for a given context (see details in the text).

Considering the **method** to be assessed and a **level of evaluation** (see section 2), one should be able to formulate the general purpose of the evaluation study, that we will call the **abstract aim** (Figure 1a).

Example 1:when evaluating a method of attenuation correction in SPECT at level 4, an abstract aim could be determining whether the method significantly improves the specificity of MIBI Tc-99m cardiac SPECT in detecting CAD.

Example 2: when considering an automatic algorithm for the detection of microcalcifications in X-ray mammogram at level 2, an abstract aim could be characterizing the performance of the algorithm for the detection of microcalcifications.

A fundamental point is that the **abstract aim** can be approached only in regard of a specific environment, that we will call a **context** (the plane in Figure 1). The context is the environment in which the method is to be evaluated and include all the components that have to be determined for defining the evaluation protocol. As will be shown below, an important feature of the context is that it also determines the applicability of the conclusions that will be drawn from the study.

The projection of the study **abstract aim** in the **context** gives a **concrete hypothesis** the test of which will complete an evaluation process. A given abstract aim can lead to different concrete hypotheses depending on the context (Figure 1a) it is projected onto. The hypothesis belongs to the context (Figure 1b) and the specification of all the context components will be related in some way to this hypothesis which plays a central role. Because the hypothesis is the projection of the abstract aim onto a context, testing the hypothesis obviously gives information regarding the abstract aim. However, achieving the abstract aim might require several contexts to be considered, and therefore several hypotheses to be tested. What we call here an evaluation study is the set of operations needed to test the concrete hypothesis. Therefore, achieving the abstract aim (i.e., testing a method at a given level), might need several evaluation studies.

*For instance, evaluating a tomographic SPECT reconstruction algorithm at level 4 is an **abstract aim**. This aim can only be approached if it is projected onto a concrete **context**, e.g., for Tc-99m MIBI SPECT cardiac studies acquired on a 3 headed gamma camera equipped with high resolution fan beam collimators, for the detection of coronary artery disease (CAD) in*

*patients with high likelihood of CAD. A **hypothesis** could be that the reconstruction algorithm allows to accurately detect perfusion defects in the left ventricle in this context. Now, if one considers another context, namely the detection of tumors in whole body SPECT oncology, a hypothesis could be that the use of the reconstruction algorithm permits the detection of tumors that are = 6 mm in diameter provided their contrast with respect to background activity is higher than 7 for instance. As the clinician might want to use the same reconstruction algorithm for several applications, achieving the abstract aim might be required for the clinician to adopt the reconstruction algorithm, i.e. several evaluation studies corresponding to different contexts hence different hypotheses to be tested might have to be conducted.*

Another example of abstract aim would be characterizing the performance of an algorithm for automatic detection of microcalcifications in X-ray mammography at level 2. In a context of simulated data, a concrete hypothesis could be that simulated microcalcifications are detected by the algorithm with a sensitivity and a specificity greater than 90%.

The hypothesis determines the **input data** (this link is illustrated by using a similar oval box for both entities in Fig 1b), i.e. the data the method is to be applied on. *In the latter example, they could be real normal mammograms in which microcalcifications are simulated using an algorithm.*

The **method** operates on the input data (all operator are represented in square boxes in Fig 1b) and the input data and the method entirely determine the **output data**, which is a component of the model that does not have any degree of freedom on its own (which is symbolized by uppercases).

The hypothesis will also motivate a **task**, which is the mean by which the output data will be converted into a **result** that will be appropriate to judge the value of the method. *For instance, the task could be comparing the binary mask corresponding to the detected microcalcifications with that corresponding to the simulated microcalcifications.*

The task will be performed by an **observer** using the output data resulting from the method to be assessed. *For instance, an algorithm can be designed to perform the comparison between the two binary masks or a human observer can do the job.*

The output data, the observer and the task entirely determine the **result** of the observation (which we will simply call result in the following), which does not have any degree of freedom on its own.

Not only the hypothesis induces input data and task, but it also induces one or several **quality indices**, i.e. some sorts of a measure that will allow one to test the hypothesis. *In our example, the quality indices could be the sensitivity and specificity of the detection of microcalcifications by the algorithm.* The definition of the quality index is highly dependent on the **reference**, which is the a priori knowledge that can be used to assess the performance on the method under study. *For instance, to use sensitivity or specificity as quality indices, whether there is a microcalcification has to be known hence the use of simulated microcalcifications.*

The **reference** is also closely connected with the hypothesis, hence with the input data. Indeed, the hypothesis is meaningful only with respect to the reference and the very formulation of the hypothesis always implicitly implies a reference: *to determine whether the algorithm detects microcalcifications with the expected sensitivity and specificity, one has to know where the microcalcifications are, hence the use of simulated microcalcifications.*

The quality index is calculated by a **comparator**, using the result and the reference as input data. However, once the quality index and the reference have been fully defined, the comparator does not have any degree of freedom of its own (hence uppercase). *In our example, it will just be the algorithm that would determine the sensitivity and the specificity of the method for detecting microcalcifications.*

Finally, the **hypothesis test** is the final step used to complete the evaluation study. It uses the quality index as input and determines whether the hypothesis should be rejected or whether there is no sufficient evidence to reject the hypothesis. *In our example, the test could compare the measured sensitivity and specificity with the expected 90% values.*

From this example, it can clearly be seen that all components of the GEM are very closely connected and that they can only be defined by taking the whole context into account. Most components have their own degrees of freedom, but are nevertheless constrained: the hypothesis constrains the input data, task, reference, quality indices and hypothesis test, i.e. these five entities can not be defined while ignoring the hypothesis. The task is not an input of the observer but the observer is constrained by the task (thick link), since he/it must be able to perform the task. The comparator is constrained by the

result and the reference (which is not of a surprise since these two components enter the comparator). The comparator is also constrained by the quality indices since the definition of the quality index precedes that of the comparator, which is only an operator providing a quality index from a result and a reference. The hypothesis test is constrained by the quality index (thick link), since it must be appropriate for giving the quality index as an output. Three of the model components are entirely determined by others (uppercases), which means that they do not have any degree of freedom of their own, and that provided the other components are properly defined, these should simply be derived from the others. These are the output data which are determined by the input data and the method, the observer's result, which is fully determined by the observer and the task, and the comparator, which is fully determined by the result, reference and quality index. In the following, we will give more details regarding the different components and explicit some of the relationships that exist between these components. We will show how this analysis can yield guidelines for evaluation.

RULE 2: As a first try, define all the components of the context as precisely as possible.

4. FROM THE GEM TO EVALUATION GUIDELINES

We will now describe the different model components in more details and especially insist upon the underlying caveats.

4.1. THE METHOD

The most obvious component common to any evaluation study is the method to be assessed. The very definition of the method to be assessed is an issue in itself because of the parametric nature of most image processing methods. Indeed, an image processing method often involves some parameters (for instance, an iteration number in iterative tomographic reconstruction, a cut-off frequency in filtering). A crucial question to be addressed when defining the method is how these parameters will be chosen. Two situations should be distinguished:

1) the method will be used with a set a “standard” parameters, “standard” meaning parameters which are conventionally applied and therefore easy to reproduce, but which are not necessarily “optimal” for the considered context. These parameters should be systematically explicitly defined: sentences such as “conventional X method was used” in the description of the evaluation protocol should be prohibited.

2) the method will be used with a set of “optimal” parameters, i.e., will give the best possible results for that particular context. The term optimal is meaningful only with respect to a criterion, which should be specified in the description of the protocol. The definition of these optimal parameters obviously requires a prior evaluation of level 1 (see section 2).

Choosing in which conditions the method will operate (standard or optimal) is of paramount importance because it fully determines the applicability of the conclusions drawn from the evaluation study: this is a first way the context affects the applicability of the study. For instance, if the method is evaluated when operating using optimal parameters, results will provide an optimistic limit of the performance of the method and it can be deduced that in the same context, the method will not give better results. On the other hand, evaluating a method operating in standard conditions does not give information regarding the best results the method could achieve.

RULE 3: Mention explicitly the values of the parameters underlying the method and how they have been determined. State whether using these parameters, the method is used in standard or optimal conditions. In the latter case, state with respect to which criterion the method can be considered as being used in optimal conditions.

4.2. THE INPUT DATA

The method obviously always operates on some input data, which should be relevant to the hypothesis to be tested. Depending on the abstract aim and context, a large variety of input data have been considered. In spite of the diversity of potentially relevant input data, there are some issues common to all evaluation studies that should be bared in mind. Defining the input data involves defining: 1) the type of the data; 2) the sample of input data; 3) the sample size.

The input data can be: simulated (*in silico* studies), real but corresponding to physical phantoms (roughly similar to *in vitro* studies), or clinical (for *in vivo* studies). The choice between these 3 types of data depends on the level of the evaluation study (section 2). Evaluation of level 1 or 2 can be performed using in silico or in vitro studies, while evaluation of any higher level should be *in vivo*. For some studies of level 1 or 2, there might be test objects conventionally adopted to characterize the performance of a method. If and only if they are relevant, these test objects should be considered in priority to facilitate comparison of the results with previously published data. The relevance of the test objects should be carefully examined however, because widespread use of an object does not guarantee relevance. For instance, using the "Lena" image or the "cameraman" image to assess image compression methods in medical image processing is not sufficient because the characteristics (texture, type of details) of these images are too different from those of medical images. If a compression method performs well on Lena, it does not guarantee that it will be good enough for compressing X-ray chest radiography.

RULE 4: Define the type of data you will be using: *in silico*, *in vitro* or *in vivo* data.

Given the type of the input data, the composition of the sample should be precisely defined. Indeed, results should only be extrapolated to samples of similar composition: this is a second way the context affects the applicability of the study. A sample is a set of individuals, which can be subjects, test-objects or images. These individuals should be chosen so that they provide a good representation of the population one would like to extrapolate the conclusions to.

For studies of level 1, it is possible to include only one or very few individuals in the sample, provided no unfounded extrapolation of the results is made. For instance, if one shows that a method failed in a specific configuration observed for a given individual, this can have important consequences, even if a single individual was considered. Similarly, if one shows that a method gives promising results on one or few individuals only, this can be enough to demonstrate the feasibility of the method (e.g., [7]). However, illustrating the relevance of a method using one or very few images of one type is usually not sufficient to conclude at the relevance of the method for other images of similar types. For instance, it is common that authors presenting a new method for image compression illustrate the method using one or two medical images (e.g., [8]). Given the large variety of medical images, these examples should always be interpreted with great cautious and should never be considered as a proof that the method will be appropriate for medical images.

As soon as there are several individuals in the sample, one might be interested in considering an homogeneous sample or an inhomogeneous sample. Actually, the sample is always inhomogeneous in some respects, so what might be clearly stated are the features with respect to which the sample is homogeneous and the features with respect to which it is not. For instance, if one wants to evaluate an automatic method for detecting hot sources in a warm background in SPECT using simulated data, the input data should include only images with warm background and without or with hot sources (and in that sense, the sample will be homogeneous), but sources corresponding to different sources/background contrasts could be included to test the method in a variety of "hot sources" configurations (and in that sense the sample would not be homogeneous). One essential point is to try to control as much as possible all sources of inhomogeneities in the sample and to mention all characteristics that might be of importance to reproduce a sample of similar composition.

RULE 5: Determine the sample composition, especially to what respect the sample is homogeneous and to what respect it is not.

RULE 6: Make sure that enough information is given for other investigators to create a sample of similar composition.

A sample is never ideal and the very process of defining the inclusion and exclusion criteria can introduce biases in the analysis. All potential sources of bias should be carefully identified. Classical biases are centripetal or popularity bias, when the sample only includes rare, difficult or challenging cases, filtering biases (e.g., when the samples only include cases for which a gold standard is known) and spectrum bias, when the sample only includes a limited range of types (e.g., lesion types) or severity of the feature of interest. These biases do not necessarily invalidate the study, provided they are clearly identified, i.e., they do not affect the internal validity of the study. However, they affect the applicability of the results to another sample, i.e., they will affect the external validity of the study.

RULE 7: State whether there is a centripetal or popularity bias, a filtering bias or a spectrum bias in the choice of the sample.

The last point regarding the input data is the size of the sample to be considered. Apart from level 1 studies for which a single individual can be considered, the appropriate sample size depends on the magnitude of the effect to be demonstrated or on the precision with which a quality index must be determined. It is of paramount importance to determine the appropriate size of the sample *before* starting the evaluation study. Many evaluation studies do not produce informative results just because too few individuals are included. When dealing with simulations or phantom data, this means that time has been wasted to study cases from which no clear conclusion could be obtained. When the evaluation study involves patients, including too few patients could even cause ethical issues, since when no conclusion can be drawn from the data, it could have been more ethical to include none of them in the study.

Several results in the literature can be used to determine the appropriate sample size. For instance, when determining the sensitivity Se of a method, the associated standard error is given by $\sigma_{Se}=[Se(1-Se)/(P-1)]^{1/2}$, where P is the number of cases which are truly positive in the study [9]. As a result, if the expected sensitivity is roughly known, the number of positive cases to be included for estimating the sensitivity with a standard error σ_{Se} can be deduced by $P=Se(1-Se)/\sigma^2_{Se}+1$. If the prevalence in the sample is Pr, a sample of P/Pr individuals should be considered. Similarly, the number of truly negative cases N to be considered when studying the specificity Sp can be derived by $N=Sp(1-Sp)/\sigma^2_{Sp}+1$, where σ_{Sp} is the standard deviation associated with the specificity. In ROC analyses also, tables have been derived to determine the appropriate size of the sample to be considered, given an estimate of the magnitude of the effect to be demonstrated (e.g., [10, 11]).

RULE 8: Use as much information as possible to estimate the size of the sample that will be appropriate for the study.

4.3. THE TASK

The input data processed by the method give the output data, the analysis of which will lead to information regarding the performance of the method. Deriving information regarding the performance of the method from the output data can be performed only by considering the hypothesis. The manner the output data is converted into a result relevant to the hypothesis is by the specification of a task, which is always defined either implicitly or explicitly in any evaluation study. The task is the mean by which the output will be converted into a result from which a judgment regarding the value of the method will be possible.

Example: if the abstract aim of the study is to determine whether an algorithm satisfactorily detects microcalcifications in X-ray mammograms, the hypothesis can be "the algorithm detects microcalcifications with a sensitivity and a specificity which are greater than 90%". The method will output binary masks corresponding to regions where microcalcifications have been detected. The task will be analyzing these masks to count and localize the microcalcifications.

The task might be: 1) a detection task (corresponding to a yes/no answer); 2) a localization task (a localization task can include a detection task, but not necessarily, since the task can be to locate an object that is known to be in the image); 3) a classification task (in which a ranking should be associated with features); 4) a characterization task (in which classes among a finite set of possible classes should be associated with features); 5) a quantification task involving the measurement of some features of interest.

RULE 9: Define precisely the task.

4.4. THE OBSERVER

Whatever the evaluation study, the task has to be performed by an observer. It can be a human observer or an algorithm. Some considerations are specific to human observers while others apply to both human observers and algorithms.

If the observer is a person, his degree of experience with respect to the task and to the context should be clearly stated. There are at least two choices in defining the profile of the observer: experts or "average" observers can be considered (worst cases can also be relevant in some instances). An expert observer is expected to be an observer with a large experience in

interpreting the data he will be provided. However, it should be precisely defined in which respect the expert can be considered as such.

Example: for interpreting mammograms, an expert could be a senior radiologist or a radiologist (junior or senior) used to interpret many mammograms in his practice. An "average" observer could be a general physician, non radiologist.

RULE 10: Define the observer. If human, determine whether experts or "average observers" will be considered. If "experts", clearly state in what respect the observers can be considered as experts.

The conditions in which the observer will perform the task should also be clearly defined, especially the information made available to him when performing the task. A human observer can be blinded to part or all additional information pertaining to the case he interprets or not. Blind interpretation can be relevant in study of levels 1 to 4, but not of higher levels. The information being made available to an algorithmic observer should also be clearly stated. For human observers, the conditions in which they perform the task can also include the conditions with which they interpret the output data (e.g., type of screen, color scale), and whether the observers are free to adjust some of these conditions at their convenience or whether they have to operate in strictly fixed conditions.

RULE 11: Determine the information that will be made available to the observer, i.e., the conditions in which he/it will operate.

When human observers are considered for studies involving patients, a potential source of bias is the prospective or retrospective nature of the study. Indeed, the observer decision can be affected by the potential consequences of his judgment. If the observer knows (because he has been informed or the context of the study is such that it cannot be otherwise) that his judgment will have no clinical impact, he might take more risks than otherwise. This makes the extrapolation of the results of evaluation studies to "real situation" particularly difficult and partially explains why evaluation study of level =4 are rare in medical image processing. This is also a part of the context that will strongly affect the applicability of the results.

RULE 12: For a human observer, state whether the study is a prospective or a retrospective study. Say whether the observer is aware that his judgment will or will not have an impact upon the patient management.

For any observer, the decision of the observer for doubtful cases should be defined. Indeed, if some input data are excluded during the analysis because of a failure of the observer, this can yield an overestimation of the performance of the method. The number and identity of data that are withdrawn from the analysis during the evaluation study should therefore be mentioned. This also gives information regarding the robustness of the evaluated method.

RULE 13: Clearly define all possible outcomes especially what will happen for "outlying" cases.

Finally, the observer reproducibility should be studied. For human observers, information regarding the intra and inter observer reproducibility should be given, since this will strongly affect the applicability of the evaluation results.

RULE 14: Give results regarding the intra and inter observer variability.

4.5. THE REFERENCE

The observer provides a result, which is entirely determined by the input data, the method, and the observer. To use this result for assessing the method, it has to be compared to some "expected" or ideal result: the reference. Two cases should be considered: either the ideal solution of the task is known (the "gold standard") and this gold standard can be used as the reference to judge the relevance of the observer's result, or the ideal solution is not known and the reference is then some a priori information which can be used in some way to assess the quality of the observer's result.

Defining the reference is of paramount importance since the conclusion of the evaluation study will be derived directly from the comparison between the observer's result and the reference. A golden rule is that the reference must neither depend on the output of the method nor on the observer's result. Although this rule might appear obvious, it is often violated.

Example: when assessing a method for the detecting activation regions in functional Magnetic Resonance Imaging, the reference is often the result itself, from which it is rather subjectively claimed that the regions detected as being the site of activation are realistic and that activation could indeed have occurred in these regions. A less disputable approach would consist in designating the regions where activation is expected BEFORE applying the method and analyzing the results provided by the method.

Depending on the context, the reference can be obtained from the characteristics of the simulated data (when performing *in silico* studies), from the characteristics of the phantom (when performing *in vitro* study), from a consensus of experts using other sources of information than the observer's result, from an histopathology or autopsy result.

RULE 15: Define a reference which is independent of the output produced by the method and of the result provided by the observer.

When there is a lapse of time between the instant the reference is determined and the instant the observer gives a result, care should be taken to account for this delay, if it can alter the relevance of the reference.

RULE 16: Check that time lapse does not jeopardize the relevance of the reference.

To derive relevant conclusions from the evaluation study, it is important to associate a level of confidence with the reference, that will indicate whether the reference is certain (for instance when performing *in silico* studies), or whether the reference is not a “true” gold standard but only an imperfect standard (sometimes also called bronze standard, or fuzzy gold standard). Indeed, the methods to be used to compare the result with the reference (see section 4.6 below) should ideally depend on the uncertainty associated with the reference and there are some methods to do so (e.g., [12-15]).

RULE 17: Associate a level of confidence with the reference.

4.6. QUALITY INDICES

The result given by the observer should be compared with the reference to state on the quality of the result hence on the value of the method being tested. A quality index allows one to characterize the adequacy of the result to the reference. The appropriate index(es) depends on the nature of the results and on the reference, and also on the hypothesis. Some indices (like those whose name include the term “error”) can only be defined when a precise reference is known, while other (such as signal-to-noise ratio) can be meaningful even when only imprecise a priori knowledge about the result (fuzzy reference) is available. A quality index is not necessarily a number but can be more complex, for instance a function or a vector. A great deal of work has been performed to design quality indices to evaluate medical images (e.g., [16]), including quality indices that would be correlated with human judgment (e.g., [17]).

Examples of indices requiring a precise reference are quantitative error, sensitivity, specificity, accuracy, area under an ROC curve [18], percentage of accurately classified features, regression curve between result and reference.

Examples of indices that can be used even if only a fuzzy reference is known are contrast or signal-to-noise ratio.

There are some standard quality indices commonly used to characterize the performance of a method (e.g., sensitivity or specificity of a method, positive predictive value and negative predictive values). Whenever possible and when they are relevant, they should be used to facilitate the comparison with values published in the literature. However, even standard indices should always be used and interpreted with cautious. All indices have their own limits and several indices are often required to achieve a sound conclusion. Examples of considerations that should be kept in mind are as follows.

When sensitivity and specificity are used, it should be reminded that although these two indices are theoretically not dependent on the prevalence, their estimated values depend on the composition of the considered sample. As sensitivity and specificity do not give useful information regarding the practical value of a method, they should always be given with the positive predictive value and the negative predictive value.

Another example concerns linear regression analyses, which are often presented to compare one set of values (y) with another (x), and the correlation coefficient is frequently use to determine the quality with which one method (giving y values) estimates the result provided by another (the x values). A correlation coefficient should always been interpreted with cautious since its value can be strongly affected by the couples of values corresponding to the highest and lowest points of the x axis. The standard error of the estimation should always be given. It is usually helpful to present results of a Bland-Altman analysis in addition to results of a correlation analysis [19]. Also, possible non linear relationship between the two variables should always be consider, in which case linear regression analysis is not appropriate to demonstrate the relationship between the two variables.

RULE 18: Carefully choose a quality index appropriate to the hypothesis to be tested and use a combination of quality indices if necessary. Beware of the limits of the considered quality indices.

4.7. HYPOTHESIS TEST

To complete the evaluation study, the hypothesis must actually be tested. A hypothesis test must therefore be defined. It is obviously constrained by the initial hypothesis. It is also constrained by the nature of the quality index since it must be appropriate to deal with the properties of the quality index (e.g., some known statistical properties of this index).

Examples of hypothesis tests are analysis of variance, comparison of two estimated values or of two sets of estimated values (e.g., mean, standard deviations, areas under ROC curves [18]), comparison of an estimated value with a theoretical value or comparison of a set of estimated values with a set of theoretical values.

There are a whole panel of statistical tests that can potentially be used to test a hypothesis, each one presenting advantages and drawbacks. However, some general considerations can be given.

Two options must sometimes be considered: 1) choosing a powerful hypothesis test which will strongly constrain the choice of the quality index so that it present the properties required by the test; 2) choosing a less powerful test (for instance nonparametric) to deal with a very relevant quality index that does not have the properties required by a more powerful test.

RULE 19: Choose a hypothesis test compatible with the properties of the quality index.

When one must choose between parametric and non parametric tests, parametric tests should be used whenever the underlying assumptions are verified, because of their higher power compared to that of non parametric tests. When the statistical distributions of the quality indices are not known, non parametric tests [20] or bootstrap approaches [21] should be considered.

RULE 20: Choose a parametric test if the assumptions underlying such a test are true. Otherwise, prefer a non parametric test or a bootstrap approach.

Whenever results can be paired (e.g., for each image, one result without and with the method to be assessed), paired tests should be used rather than unpaired tests as they are more powerful.

RULE 21: Use paired tests if possible.

Whenever multiple comparisons have to be performed, care should be taken to account for the multiple comparison effect. A appropriate correction for multiple comparison (e.g., [22]) should be performed when necessary.

RULE 22: Account for multiple comparisons if necessary.

5. DISCUSSION AND CONCLUSION

In medical image processing, most evaluation studies rely on some implicit rules in the design of the evaluation protocol, but unlike in clinical trials, there are no formal rules that are commonly accepted and that have to be followed to ensure the appropriateness of the investigations. This lack of "norms" definitely slows down progress in the field for the three main reasons given in the introduction. One explanation for this lack of norm could be the large variety of evaluation studies that can be conducted in medical image processing depending, among other things, on the nature of the data to be processed and on the aim of the evaluation study. This variety makes it difficult to give "recipes" that would ensure the appropriateness of an evaluation study and guarantee the quality of the results it provides.

To try better control the elements that impact the validity of an evaluation study and to facilitate the design of protocols that answer questions that we encounter when developing image processing methods, we propose to model the evaluation process in medical image processing. The idea was to extract a model that is common to any evaluation study, i.e., that can be used to describe any evaluation protocol. The resulting generic evaluation model therefore features components common to any evaluation protocols, such as input data, method to be evaluated or observer. It is important to underline that although it might appear obvious that these components enter the evaluation model, some of them are often only implicitly defined when describing an evaluation protocol. We think that these implicit definitions partially explain why evaluation studies described in the medical image processing literature are not always convincing and cannot be easily reproduced by other investigators. We are convinced that an explicit specification of the model components when designing and reporting an evaluation protocol would facilitate the research in the field.

Defining the model components to design an evaluation protocol is not as straightforward as it might seem at a first glance. Indeed, the links and constraints that exist between the different components are quite intricate, and the consistency of an evaluation protocol is not that easy to achieve. In this paper, we have illustrated some of the relationships between the components that should be taken into account when designing an evaluation protocol. One reason why the specification of the model components should be performed with so much care is that it entirely determines the internal and external applicability of the results. We pointed out some of the limits implied by the choice of specific components and these limits should be bared in mind when designing an evaluation protocol.

The model we propose could be a first step towards evaluation guidelines in medical image processing. By highlighting some basic rules, we pointed out directions for which more specific guidelines should be established. Much work has to be performed to achieve sort of a norm for evaluation in medical image processing. Such a norm would indeed require a more comprehensive analysis of the different classes of cases that can be encountered and more suggestions regarding solutions to define the appropriate components for these different classes. For instance, also the GEM applies for any level of evaluation, each level presents some specificities that should be taken into account. Similarly, a list of tasks could be defined and the specific features of each one should be indicated. Hypothesis tests appropriate for different classes of hypothesis could also be suggested, and so on.

Although the GEM has been described and proposed for medical image processing procedures, we feel that the GEM is generic enough to describe evaluation protocols used not only for medical image processing assessment, but also for the evaluation of imaging devices and protocols for instance. Characteristics of the model that are specific to medical image processing appear only at a certain level of interpretation of the model (e.g., when defining some constraints between the components), but many model components and constraints should be identical or similar regardless the nature of the entity being tested (e.g., an image processing procedure, an acquisition device, an imaging protocol).

6. REFERENCES

1. K. M. Hanson. Method to evaluate image-recovery algorithms based on task performance. *SPIE*, 914:336-343,1988.
2. K. E. Britton and E. B. Sokole. COST B2: why and wherefore. *Eur. J. Nucl. Med.*, 19:563-568,1992.

3. S. Furuie, G. Herman, T. Narayan, P. Kinahan, J. Karp, R. Lewitt and S. Matej. A methodology for testing for statistically significant differences between fully 3D PET reconstruction algorithms. *Phys. Med. Biol.*, 39:341-354,1994.
4. S. Matej, G. Herman, T. Narayan, S. Furuie, R. Lewitt and P. Kinahan. Evaluation of task-oriented performance of several fully 3D PET reconstruction algorithms. *Phys. Med. Biol.*, 39:355-367,1994.
5. D. G. Fryback and J. R. Thornbury. The efficacy of diagnostic imaging. *Med. Decis. Making*, 11:88-94,1991.
6. I. Buvat, F. Frouin, G. El Fakhri, H. Benali, V. Chameroy, F. Aubry and R. Di Paola. Rôle de la simulation dans la méthodologie d'évaluation en imagerie médicale. *Méd. Nucl.*, 20:472,1996.
7. G. N. Hounsfield. Computerized transverse axial scanning (tomography). Part I: description of a system. *Br. J. Radiol.*, 46:1016-1022,1973.
8. A. Uhl. Generalized wavelet decompositions in image compression: arbitrary subbands and parallel algorithms. *Opt. Eng.*, 36:1480-1487,1997.
9. D. Green and J. Swets. *Signal detection theory and psychophysics*. Krieger, Huntington, NY, 1974.
10. J. A. Hanley and B. J. McNeil. The meaning and use of the area under the receiver operating characteristic (ROC) curve. *Radiology*, 143:29-36,1982.
11. J. A. Hanley and B. J. McNeil. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148:839-843,1983.
12. C. Berche, F. Aubry, C. Langlais, J. Vitaux, C. Parmentier and R. Di Paola. Diagnostic value of transverse axial tomoscintigraphy for the detection of hepatic metastases: results on 53 examinations and comparison with other diagnostic techniques. *Eur. J. Nucl. Med.*, 6:435-452,1981.
13. R. M. Henkelman, I. Kay and M. J. Bronskill. Receiver Operator Characteristic (ROC) analysis without truth. *Med. Decis. Making*, 10:24-29,1990.
14. P. N. Valentein. Evaluating diagnostic tests with imperfect standards. *Am. J. Clin. Pathol.*, 93:252-258,1990.
15. C. E. Phelps and A. Hutson. Estimating diagnostic test accuracy using a “fuzzy gold standard”. *Med. Decis. Making*, 15:44-57,1995.
16. R. F. Wagner and D. G. Brown. Unified SNR analysis of medical imaging systems. *Phys. Med. Biol.*, 30:489-518,1985.
17. W. E. Smith and H. H. Barrett. Hotelling trace criterion and its correlation with human observer performance. *J. Opt. Soc. Amer. A.*, 3:717-723,1986.
18. C.E. Metz. Basic principles of ROC analysis. *Semin. Nucl. Med.*, 8:283-298,1978.
19. J. M. Bland and D. G. Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, 307-310,1986.
20. S. Siegel and N. J. Castellan. *Nonparametric statistics for the behavioral sciences*. McGraw-Hill, New York, 1988.
21. B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. Chapman & Hall, New York, 1993.
22. E. R. DeLong, D. M. DeLong and D. L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44:837-845,1988.