Open Journal of Medical Imaging, 2014, 4, 205-209
Published Online December 2014 in SciRes. http://www.scirp.org/journal/ojmi
http://dx.doi.org/10.4236/ojmi.2014.44028

Scientific Research

Medical Image Acquisition and Processing: Clinical Validation

Michael L. Goris

Stanford University School of Medicine, Stanford, USA
Email: mlgoris@stanford.edu

Received 27 October 2014; revised 26 November 2014; accepted 22 December 2014

Abstract

The validation of medical imaging (processing and acquisition) can be achieved in multiple ways, somewhat influenced by the context. There are three traps to avoid: First reliance on ground truth requires the knowledge of it before the end of the trial, second comparison to gold standards cannot show improvement and finally one needs to deal with confirmation bias. In this paper we discuss those topics and alternative validation schemes.

Keywords

Medical Imaging, Clinical Validation

1. Introduction

When new imaging technologies begin to be used in clinical settings, little is known about their potential to improve care. They are usually "sold" on technological or impressionistic criteria. Clinical validation is rare, because the consensus is that it would delay the application. In this paper we discuss the different methods for validation, not all of them which would cause implementation delay.

Ideally, for the sake of improving the appropriateness of medical imaging, one would hope for more rapid progression to what can be called *scientific clinical research* or *technology assessment*.

The first level of evaluation can be called *diagnostic efficacy*[1]. The appropriate research question to ask in this phase, while the technology is just beginning to diffuse into clinical practice or when a substantive advance in the technology occurs, is "How well does the new technology detect specific disease conditions?" The measures of efficacy are the operating characteristics (sensitivity, specificity). Effectiveness includes the positive and negative predictive value for a given prevalence of the disease in the study population and receiver-operating cha-

[1]Effectiveness or efficacy is the extent to which planned outcomes, goals, or objectives are achieved as a result of an activity. Effectiveness assumes a field result, efficacy ideal circumstance. Efficiency is the ratio of the output to the inputs of any system. A test with perfect effectiveness and efficacy could be inefficient because of cost or side effects.

racteristic (ROC) analysis [1]. The test has diagnostic efficacy if it classifies the patient in the correct category[2]. The correct category in earth sciences is called the ground truth or in medicine, the defining diagnosis. One type of defining diagnostic techniques is based on the analysis (under the microscope) of tissues obtained from a lesion (e.g. histology, from a biopsy or in autopsy) or microorganism detection in the case of infection. The diagnosis is not changed if the patient dies earlier or later than expected, or responds differently to therapy.

The defining diagnosis is really a type of taxonomy[3] and defines the ground truth[4]. The assumption is that the result of the taxonomy is related to outcome or the result of a specific therapy. While diagnostic efficacy is principally of interest to radiologists, referring clinicians may be more interested in how the information derived from an imaging test affects how they care for patients, represented by the concepts of *therapeutic thinking efficacy*. For therapeutic thinking, the corresponding question relates to the effect of imaging on considerations of treatment.

The validation of defining diagnostic technique would be solipsistic. We will concentrate on non-defining diagnostic techniques and as much as we can on techniques based on computer processing of medical images. But why would we want a new diagnostic procedure? The global answer is that the existing one or the combination of existing ones is too costly or lacks in efficacy, effectiveness or efficiency. Costly should be understood as combining expenses in material and personnel, pain, danger, the lack of accuracy (strictly speaking unfavorable operating characteristics for the examined population: e.g. false positive rates), lack of predictive value

Not all diagnostic techniques aim to properly classify the patients, but rather to predict either the outcome, or the best therapy to obtain the desired outcome (measurements of plasma cholesterol do not provide a diagnosis but a prognosis, staging is not diagnostic, but predictive). Ultimately, the diagnostic technique should be evaluated in its role in the management of the patient, or the outcome[5].

2. Validation Approaches

2.1. Outcome Analysis

Outcome analysis is actually based on large population studies. Originally it was presented as the method to evaluate cost-effectiveness (e.g. did people live longer if there were more MRI scanners in the region). More pointedly, it has been used to look at the efficacy of screening studies. The analysis of those data is complicated by the concatenation of increased detection (an increase in incidence) and the fact that early detection may not always predict progression. For breast cancer and mammography it did take a long time to show the beneficial outcome [2].

Imaging usually represents only one or a few steps in a chain of diagnostic and therapeutic interventions, so how can we ascribe an outcome to any one of these? The performance of an imaging test may be excellent, but patients might still have adverse outcomes because the treatment was inappropriate or no adequate treatment exists. As a result of all of these factors, outcomes evaluations of imaging technologies are rare. In the case of screening for early detection, the outcome is affected only to the extent that the treatment is (relatively) effective in early stages and not in more advanced disease. Outcome studies measure effectiveness rather than efficacy. Outcome analysis as validation would take too long and prevent the introduction of new techniques.

2.2. Predictive Power

A taxonomic exact diagnosis may not be predictive. Consider that in some diseases the median survival time is n years: fifty percent die earlier, 50% later. The prognosis is not necessarily well defined by the diagnosis. Staging refines the prognosis, or the expected response to a particular therapy. Other techniques can be used to predict earlier which therapy will fail or succeed so that alternatives can be used [3] [4]. Early response to therapy may predict the long term results. There is a time lag between development and the definition of predictive power, but this approach is less burdensome than outcome analysis.

[2]Diagnosis (Greek: *διάγνωση*, from *δια* dia- "apart-split", and *γνώση* gnosi "to learn, knowledge") is the identification of the nature of anything, either by process of elimination or other analytical methods. Therefore defining diagnostic techniques are not really diagnostic but classifiers.

[3]Taxonomy is the practice and science of classification. The word finds its roots in the Greek *τάξις*, *taxis* (meaning "order", "arrangement") and *νόμος*, *nomos* ("law" or "science"). Taxonomy uses taxonomic units, known as taxa (singular taxon)

[4]Ground truth is a term originally used in remote sensing; it refers to information collected on location. Ground truth allows image data to be related to real features and materials on the ground. In medicine it refers to the verified diagnosis.

[5]From the mid-17th century via late Latin < Greek *prognōsis* "knowledge beforehand" < *gignōskein* "know".

2.3. Predicting the Taxonomy

This is the most common type of validation for diagnostic techniques. The most relevant aspects of this approach are 1) that a ground truth is assumed to exist and be known and 2) that at some point there has to be a defining test or the next best thing (e.g. a gold standard)[6].

The major problem is verification bias in the first case always and in the second case mostly. An example of Verification Bias is the evaluation of Myocardial Perfusion Scintigraphy (MPS). The gold standard for the presence of (significant) coronary artery diseases (CAD) was originally the coronary arteriogram (CA).The MPS study would select patients more likely to need the CA. However, trust came too early. A value was ascribed to MPS (prematurely) and soon the probability of a CA being performed following a negative MPS decreased while the probability of a CA being performed following a positive MPS increased. The result was verification bias: with an over-estimation of the sensitivity and an under-estimation of the specificity. The proper validation would have been to perform MPS only on patients who had a positive or negative CA and do it blindly.

There are ways to overcome verification bias: one of them is to look at populations with a known prevalence [5] or stratified populations, another to correct the bias on the assumption of a neutral pre-selection [6]. The former is based on the fact that if groups are known to have a prevalence of CAD, without defining which individuals actually have it, it is axiomatic that the prevalence of positive test should correspond with the prevalence of the disease in the group.

If existing populations with known prevalence exist, this is a fairly direct approach, but expensive to implement.

2.4. Discriminating Power

There are two concatenated conditions for a test to be discriminating: the metric has to have intrinsic discriminating value and the measurement has to be precise enough, so that variability in testing does not reach the magnitude of the difference between affected and unaffected.

2.4.1. Patient Study A

25 patients with “early” CF and 10 control cases. Patients are defined by genetics or sweat test. Controls are non-affected siblings in the same age range and same sex distribution. Did the pulmonary function tests discriminate between both groups [7] **Table 1** & **Table 2**?

Table 1. Pulmonary function tests in 25 children with cystic fibrosis, compared to unaffected siblings: RV = respiratory volume; TLC = total lung capacity; IC% = Inspiratory Capacity; SVC = Slow Vital Capacity; FVC = Forced Vital Capacity; FEV1 = Forced Expiratory Volume in the first (3, 5) second following max inhalation; FEF = forced expiratory flow.

	RV/TLC	IC%	FVC%	SVC%	FEV1%	FEV1/FVC	FEF25-75%	FEFMax%
Mean CF	28.2	102.3	115.4	110.7	104.0	80.4	83.6	99.6
STDV CF	10.2	18.9	18.5	19.8	17.1	7.6	29.5	25.6
Mean Nl	21.6	97.0	111.6	112.3	106.6	84.1	103.9	100.5
STDV NL	4.0	6.3	15.1	13.3	13.7	6.4	26.5	20.9
T-test	**0.011**	**0.226**	**0.568**	**0.814**	**0.668**	**0.189**	**0.068**	**0.926**

Table 2. The two groups are better discriminated by looking at quantitative air-trapping.

		A1	A2	A3
25 CF	Mean	16.16	9.83	4.50
	SD	14.71	10.30	5.11
10 NL	Mean	5.22	2.27	0.82
	SD	3.64	1.72	0.62
	T-test	**0.0013**	**0.0012**	**0.0013**

[6]The defining diagnosis defines the ground truth; the next best thing is a gold standard.

2.4.2. Patient Study B

To evaluate quantitative air trapping measurements in children with mild cystic fibrosis (CF) lung disease during a one year double-blind placebo-controlled rhDNase intervention trial and compare results from quantitative air trapping with those from spirometry or visually scored HRCT scans of the chest [8] (**Table 3**).

In a certain sense discriminating power is the el dorado of image processing, because if the exactitude of the measurement can often be determined (e.g. the shape and size of an hearing aid determined from an image of the external auricular canal), this is not true in all cases: there is no life verification of early air trapping in children with Cystic Fibrosis and there is no life diagnosis of Alzheimer disease in elderly. In this case discriminating power overcomes the lack of ground truth and shows efficacy.

2.5. Internalized Validation

The prototype of internalization is automation of region of interest (ROI) definition. The creator or user of the routine assumes that the user knows if a ROI is correctly placed and limited. The question is whether the automated program yields a result acceptable to the observer, and how frequently.

Another is image processing that extracts an image feature, and hence not only facilitates interpretation, but makes it more reproducible. Again, the validation is an agreement with the observer.

The validation is internalized, not because of clinical criteria, but because it standardizes interpretation to the satisfaction of the user. It is a weak validation, but immediate and cheap.

2.6. Equivalence

Equivalence is based on the comparison with an established diagnostic procedure. The established procedure is sometimes referred to as "gold standard", even if it is not a perfect procedure. More precisely this approach is referred to as a "**no worse than**" design. The "not worse than" denomination refers to the fact that at best the evaluated diagnostic procedure perfectly matches the "gold standard", but cannot be shown to be better: all discrepancies are demonstrating a worse performance.

It takes different forms. In MPS the gold standard was the CA; however the metric was not the same: the arteriographic measurement of stenosis does not necessarily determine the relative decrease of flow in the dependent myocardium. In addition, a normal MPS predicts a lowering of risk for myocardial ischemic events independently of the CA findings [9].

In imaging a common study design is to compare the automatic analysis to the judgment of a panel of (experienced) experts (see also internal validation). Again, the performance cannot be shown to improve in the new modality since the human observers define the truth.

The equivalence design is not altogether valueless since the new procedure may globally decrease the cost (expenses in material and personnel, pain and danger). What can be demonstrated is an improvement in reproducibility of the interpretation if the method is automated or quantitative.

However, the use of gold standards, while easily performed, may be dangerous if the metric differ in a physiological important manner.

3. Conclusion

Cost effectiveness or efficiency evaluation would be the next step: Assuming that we reach a satisfactory clinical validation, we would also like to know how much a successful technology will cost. This allows us to see, regardless of the health benefit, whether society can afford to implement it on a broad scale. Since cost is rela-

Table 3. Discriminating the effect of treatment.

Metric	Pulmozyme™	Placebo	P
N	**11**	**14**	
A1	−2.10 ± 44.80	34.40 ± 62.10	0.102
A2	−9.30 ± 42.90	43.40 ± 73.20	0.035
A3	−13.10 ± 40.50	48.20 ± 81.20	−0.02

tive, researchers generally relate it to how much benefit is obtained for how much money. They have developed the ratio of cost per years of life saved and refer to this ratio as a technology's *cost-effectiveness* [10]. The next step in the evaluation must include the result in the targeted population or efficiency, which is also a function of the disease prevalence in that population.

References

[1] Hanley, J.A. and McNeil, B.J. (1982) The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology*, **143**, 29-36. http://dx.doi.org/10.1148/radiology.143.1.7063747

[2] Berry, D.A., Cronin, K.A., Plevritis, S.K., Fryback, D.G., Clark, L.C., Zelen, M., Mandelblatt, J.S., Yakovlev, A.Y., Habbema, J.D.F. and Feuer, E.J. (2005) Contributions of Screening and Adjuvant Treatment to Reduction in Breast Cancer Mortality in the US from 1975 to 2000. *New England Journal of Medicine*, **353**, 1784-1792. http://dx.doi.org/10.1056/NEJMoa050518

[3] Zhu, H.J. and Halkar, P.K. (2007) An Evaluation of the Predictive Value of During Treatment 18F-Fluorodeoxyglucose PET/CT Scans in Pediatric Lymphomas. *RSNA Scientific Assembly and Annual Meeting Program*, 954.

[4] Zhu, H.J., Halkar, R., Alavi, A. and Goris, M.L. (2013) An Evaluation of the Predictive Value of Mid-Treatment 18F-FDG PET/CT Scans in Pediatric Lymphomas and Undefined Criteria of Abnormality in Quantitative Analysis. *Hellenic Journal of Nuclear Medicine*, **16**, 169-74.

[5] Goris, M.L., Bretille, J., Askienazy, S., Purcell, G.P. and Savelli, V. (1989) The Validation of Diagnostic Procedures on Stratified Populations: Application on the Quantification of Thallium Myocardial Perfusion Scintigraphy. *American Journal of Physiological Imaging*, **4**, 11-15.

[6] Diamond, G.A., Rozanski, A., Forrester, J.S., Morris, D., Pollock, B.H., Staniloff, H.M., Berman, D.S. and Swan, H.J.C. (1986) A Model for Assessing the Sensitivity and Specificity of Tests Subject to Selection Bias: Application to Exercise Radionuclide Ventriculography for Diagnosis of Coronary Artery Disease. *Journal of Chronic Diseases*, **39**, 343-355. http://dx.doi.org/10.1016/0021-9681(86)90119-0

[7] Goris, M.L., Zhu, H.J., Blankenberg, F., Chan, F. and Robinson, T.E. (2003) An Automated Approach to Quantitative Air Trapping Measurements in Mild Cystic Fibrosis. *Chest*, **123**, 1655-1663. http://dx.doi.org/10.1378/chest.123.5.1655

[8] Robinson, T.E., Goris, M.L., Zhu, H.J., Chen, X., Bhise, P., Sheikh, F. and Moss, R.B. (2005) Dornase Alfa Reduces Air Trapping in Children with Mild Cystic Fibrosis Lung Disease: A Quantitative Analysis. *Chest*, **128**, 2327-2335. http://dx.doi.org/10.1378/chest.128.4.2327

[9] Hachamovitch, R., Berman, D.S., Shaw, L.J., Kiat, H., Cohen, I., Cabico, J.A., Friedman, J. and Diamond, G.A. (1998) Incremental Prognostic Value of Myocardial Perfusion Single Photon Emission Computed Tomography for the Prediction of Cardiac Death: Differential Stratification for Risk of Cardiac Death and Myocardial Infarction. *Circulation*, **97**, 535-543. http://dx.doi.org/10.1161/01.CIR.97.6.535

[10] Beinfeld, M.T., Wittenberg, E. and Gazelle, S.G. (2005) Cost-Effectiveness of Whole-Body CT Screening. *Radiology*, **234**, 415-422. http://dx.doi.org/10.1148/radiol.2342032061