Published in final edited form as:

Acad Radiol. 2010 September ; 17(9): 1079–1082. doi:10.1016/j.acra.2010.05.021.

Developing a New Reference Standard... Is Validation Necessary?

Rachel Gold¹, Melissa Reichman, MD², Edward Greenberg, MD², Jana Ivanidze³, Elliott Elias⁴, Apostolos J. Tsiouris, MD², Joseph P. Comunale, MD², Carl E. Johnson, MD², and Pina C. Sanelli, MD, MPH^{2,5}

¹ New York College of Osteopathic Medicine 1 Northern Boulevard P.O. Box 8000 Old Westbury, NY 11568

² Department of Radiology Weill Cornell Medical College New York Presbyterian Hospital 525 East 68th Street, Starr 8A New York, NY 10065

³ Ludwig Maximilians University of Munich Geschwister-Scholl-Platz 1 Munich 80539 Germany

⁴ Boston University School of Medicine 715 Albany Street Boston, MA 02118

⁵ Department of Public Health, Weill Cornell Medical College New York-Presbyterian Hospital 525 East 68th Street, Starr 630 New York, NY 10065

Abstract

A gold standard is often an imperfect diagnostic test, falling short of achieving 100% accuracy in clinical practice. Using an imperfect gold standard without fully comprehending its limitations and biases can lead to erroneous classification of patients with and without disease. This will ultimately affect treatment decisions and patient outcomes. Therefore, validation is essential prior to implementation of the reference standard into practice. Performing a comprehensive validation process is discussed along with its advantages and challenges. The different types of validation methods are reviewed. An example from our work in developing a new reference standard for vasospasm diagnosis in aneurysmal subarachnoid hemorrhage (A-SAH) patients is provided. Employing a new reference standard may result in a definitional shift of the disease and classification scheme of patients. Thereby, it is important to also assess the impact of a new reference standard on patient outcomes and its clinical effectiveness.

Introduction

The term “gold standard” is used to describe a diagnostic test that is regarded as definitive for a particular disease, and thereby becomes the ultimate measure for comparison. However, these gold standard methods are often imperfect and do not have 100% accuracy in practice. A “perfect” gold standard may only exist in theory and caution must be taken when classifying patients with a diagnosis based on a current gold standard. For example, colposcopy-directed biopsy of the cervix is the current gold standard for detection of cervical neoplasia. However, its sensitivity is only 60% which would seem that it is far from

Corresponding Author: Rachel Gold Department of Radiology New York Presbyterian Hospital 525 East 68th Street New York, NY 10065 Telephone: 212-746-2573 rlgold83@gmail.com.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

being a definitive test for this patient population [7]. In practice, diagnostic tests are sometimes assigned the status of gold standard, or as the “most sensitive test,” without verification [7]. Using these imperfect gold standards without understanding their limitations may lead to erroneous classification of patients affecting treatment decisions and patient outcomes. Therefore, validation of a new reference standard is an important step prior to its implementation in practice.

Validation is a term that has become synonymous with “accuracy.” To measure validity, or accuracy, is to assess how closely the results from a new diagnostic method approximate the current gold standard [14]. However, a diagnostic test is not validated based on accuracy statistics alone. It is important to assess that the reference standard does indeed do what it is intended to do in the target population [2]. The interpretation of these test results in the reference standard and the impact this has on patient outcomes are necessary factors to also consider in the validation process. Comprehensive evaluation of a reference standard includes assessing its clinical credibility, accuracy for diagnosis, generalizability to other target populations, and ideally assessing its clinical effectiveness [2].

Validation processes can be helpful in identifying flaws and bias in the reference standard that may result in misleading findings [13]. These include selection bias, poorly defined inclusion and exclusion criteria, missing data and unclear rationale for treatment decisions. Selection bias may exist in a reference standard that is only applicable to a subgroup of the target population. For example, digital subtraction angiography (DSA) is considered the gold standard for the diagnosis of vasospasm in aneurysmal subarachnoid hemorrhage (A-SAH) patients. Being an invasive technique with a complication rate of approximately 5% and a permanent stroke rate estimating 0.5–1% [3,8], DSA is not performed in all patients. Due to these associated risks, patients with high suspicion of vasospasm based on symptoms and/or other imaging tests are more likely to have a DSA performed compared to patients without symptoms, resulting in selection bias. The performance of the gold standard in this subgroup of high risk patients may not be the same for all patients in the population, thereby limiting its generalizability.

Our goal is to develop a reference standard that is applicable to the entire A-SAH population, including patients with and without symptoms, and importantly not to exclude patients that did not receive a particular imaging test. We hope that this novel approach in developing a reference standard, using both clinical and imaging criteria with consideration of treatment effects, will improve the accuracy in the classification of patients with and without vasospasm. This new reference standard will result in a definitional shift of the diagnosis of vasospasm, affecting the classification scheme and treatment decisions of A-SAH patients.

Development of a New Reference Standard for Vasospasm

Our development of a new reference standard for vasospasm in A-SAH patients is based on the definition of cerebral vasospasm which encompasses both clinical criteria of delayed onset of ischemic neurologic deficits and imaging criteria of narrowing of cerebral vessels documented by angiography or other imaging modalities. DSA is considered the gold standard for angiographic vasospasm and is a widely accepted method because of its superior spatial and temporal resolution compared to other imaging modalities. However, its applicability to only a subgroup of the population limits its usefulness and effectiveness as a reference standard technique. In order to design a reference standard that is applicable to the entire A-SAH population and to include a comprehensive definition of vasospasm, then multiple sequential tests are designed in a composite reference standard, incorporating both the clinical and imaging features of vasospasm. The composite reference standard is an

alternative method used when a “true” gold standard does not exist, or in most situations when the current gold standard has low detection of the disease in the population [1]. In theory, utilizing a combination of tests would result in a reference standard that has higher sensitivity and specificity than any individual test used alone. An advantage of this method includes that several sources of information are used for assessment of a complex disease, such as vasospasm, that has several definitions or criteria for diagnosis. Another advantage is that the multiple tests in the reference standard can be organized in a sequential fashion which avoids redundant and excess testing in the population [1].

Our new reference standard for the diagnosis of vasospasm in A-SAH patients is a novel approach consisting of a multi-stage hierarchical system incorporating patient outcome measures and consideration of treatment effects. It is unique for a reference standard to include the effects of treatment in the classification scheme. In general, a reference standard is not applicable to patients who have received treatment for the disease. However, this may be an important issue to address in patient populations that receive treatment as prophylactic measures.

The primary, secondary and tertiary levels developed in our reference standard are designed to include diagnosis of vasospasm using both imaging and/or clinical criteria. These levels are organized in a sequential fashion with weighted significance according to the strength of evidence for diagnosing vasospasm [10]. The primary level is considered the strongest level of evidence. Even though a patient may have multiple levels of evidence, only the highest level will be used to determine vasospasm according to this system. The following is a brief description of this reference standard. Reichman et al. has provided detailed explanation of the application of this reference standard in practice with its advantages and limitations [10]. It is important to emphasize that all A-SAH patients proceed through this hierarchical reference standard using the same criteria and methodology. The primary level in the reference standard uses DSA to determine the presence or absence of vasospasm. The severity of vasospasm are defined as: mild vasospasm as less than 50% degree of luminal narrowing; moderate vasospasm as 50%–75% degree of luminal narrowing; and severe vasospasm as greater than 75% degree of luminal narrowing. No vasospasm is defined as no evidence of luminal narrowing.

Patients that did not undergo a DSA during hospitalization proceed to the secondary level in the reference standard to evaluate for sequelae of vasospasm using both clinical and imaging criteria. The clinical criteria assesses permanent neurological deficits (distinct from the deficit at baseline produced by the A-SAH) evaluated on clinical exam. The imaging criteria evaluates for evidence of delayed infarction present on computed tomography (CT) and/or magnetic resonance imaging (MRI). Delayed infarction is defined as new infarction on CT or MRI after day 4 that has not been present on the initial CT within 3 days after onset [9,12]. A vasospasm diagnosis is assigned if the patient meets either (or both) of the clinical and imaging criteria. However, if the patient does not meet these criteria, and they have not been treated for vasospasm, then a no vasospasm diagnosis is assigned.

Patients who did not undergo a DSA exam and have no sequelae of vasospasm on clinical exam and/or imaging, but were treated for vasospasm, proceed to the tertiary level. The tertiary level assigns the diagnosis of vasospasm based on the patient’s response-to-treatment. “Triple H” (HHH) therapy stands for medically induced hypertension, hypervolemia and hemodilution. Medical HHH therapy has been shown to reverse the onset of ischemic neurologic deficits in patients who have A-SAH [6]. Those patients who showed improvement upon clinical exam and/or symptoms following the administration of HHH therapy are considered responders to appropriate therapy and are assigned a vasospasm

diagnosis at the tertiary level. In patients without a response-to-treatment and another etiology for the patient's symptoms is identified, then a no vasospasm diagnosis is assigned.

Validation Methods

A comprehensive validation process includes both internal and external validation strategies. Internal validation refers to methods performed on a single data set to determine the accuracy of a reference standard in the classification of patients with or without disease in the target population. Conflicts may arise when a new reference standard challenges the current gold standard. Both clinical reasoning and statistical analysis are used to determine replacement of the current gold standard [15]. There are several essential issues to consider regarding the reproducibility, amount and quality of information provided, and the accuracy of the new reference standard [15]. Ultimately, a reference standard is a matter of choice by the investigator and needs to fit best for the diagnostic outcome and the patient population for which its use is intended.

External validation evaluates the generalizability of the reference standard by demonstrating its reproducibility in other target populations. In other words, how precise is the reference standard and what is the test-retest reliability? Precision is measuring the degree to which a result obtained by the reference standard is repeated on a second occasion in the same study population [14]. Although a test may be very accurate, as demonstrated by internal validation methods, it can have poor precision. For example, vaguely defined criteria in the reference standard can lead to variability in the classification of patients, and thus poor reproducibility. When validating a reference standard, it is important to not only consider accuracy, but to also ensure reproducibility and generalizability to other target populations.

Our internal validation process includes two phases to comprehensively address the statistical accuracy of the new reference standard and also its ability to replace the current gold standard. Phase I is designed to compare the secondary/tertiary levels of the reference standard with the current gold standard of using DSA alone. Patients who had a DSA performed will then be applied to the secondary/tertiary levels in order to compare the diagnostic outcomes with DSA. Phase II is designed to evaluate the accuracy and feasibility of applying the new reference standard to the target population by comparison with the chart diagnosis for vasospasm. Phase II represents the application of the reference standard in practice with A-SAH patients.

Discussion

Diagnostic methods continue to advance and improve with development of new technology that will continuously challenge our current gold standards and classification schemes. It is important to recognize that a definitional shift may occur when using a new reference standard. As a result, additional cases of disease may be detected creating uncertainty about whether these additional cases truly represent new cases that are detected by a superior method or it represents false-positive results that should not be treated for disease [5]. Therefore, it is essential to adhere to a set of principles that will guide replacement of a current gold standard through a vigorous validation process that not only assesses accuracy but also the consequences of the switch, both nosologically and clinically [5].

A review of the literature reveals that accuracy measures are not always essential in the validation process of a new reference standard. The accuracy paradigm of assessing sensitivity and specificity of diagnostic tests is abandoned in regards to validation of a reference standard [11]. Rather, clinical events (the number of events in those tested positive and negative for the disease) are the primary focus and calculating measures of event rates,

relative risks and correlation statistics are performed [11]. Currently, there is no consensus as to an acceptable performance level of a new reference standard prior to its acceptance into practice. There is a need for the scientific and clinical community to define a threshold of performance considered sufficient to allow the implementation of a new reference standard with confidence in its classification of patients and understanding of its limitations [11].

Initially in the development of a new reference standard, it is important to achieve a high profile in the standard accuracy characteristics; including sensitivity, specificity, positive and negative predictive values for that particular disease. Once an accurate reference standard is established, further efforts are then focused on assessing its impact on treatment decisions and patient outcomes. For example, Fraile describes two different approaches for validation of sentinel node biopsy (SNB) in cancer staging [4]. The first approach is to compare SNB with the current gold standard, axillary lymph node dissection (ALND), and calculate the standard accuracy parameters. A high sensitivity rate of 96%, negative predictive value of 97.3%, and false-negative value of 4% were reported in this study for SNB. The next step was then focused on the prognostic and therapeutic impact of SNB on the patient population [4]. This study had established that SNB more accurately stages cancer in a considerable number of patients, and it is superior to the current gold standard for lymph-node assessment in breast carcinoma [4]. Therefore, the replacement of the current gold standard is accepted based on evidence that the new reference standard has the potential to improve diagnosing/staging in a very significant way. SNB ultimately leads to a change in adjunct therapy with an expected positive impact on survival [4]. This suggests that both accuracy characteristics as well as patient outcomes need to be considered when validating a new reference standard.

From a different perspective, Altman describes two types of validation processes [2]. The first is a statistical validation, described as one which passes all appropriate statistical checks, including goodness-of-fit on the original data set and unbiased prediction on a new data set. The second is a clinical validation, representing one which performs satisfactorily on a new data set according to its predetermined context-dependent criteria. It is important to emphasize that a clinically validated model may be statistically invalid. For example, its predictions are biased, or it fails a goodness-of-fit test. The contrary may also be true in which a statistically validated model may be clinically invalid. For example, the intrinsic prognostic information is too weak. Generally it may be more challenging to achieve statistical validation because of the considerable difficulty in overcoming the introduced bias of over optimism at the model building stage. However, clinical validation may be more useful in uncommon disease entities because of small sample size and limited transportability of the reference standard.

In an attempt to validate a reference standard both statistically and clinically, limitations exist partly due to the idealization of the gold standard concept. A perfect gold standard may only exist in theory. If the current gold standard is flawed or has low detection of the disease, then using it as the comparison for the new reference standard will result in false conclusions. The new reference standard will be limited by the accuracy of the current gold standard and will not be able to achieve higher sensitivity and specificity. Currently, there are no obvious means of deciding whether the proposed newly developed reference standard is in fact really better than the current gold standard [5]. In this situation, it remains uncertain if the additional cases identified as disease by the new reference standard represent true-positive or false-positive cases [5]. In certain disease entities, long-term follow-up of these patients may be helpful in resolving these conflicts. Another limitation in the validation process is small sample size limits evaluation of the new reference standard. This sample-size effect is of concern in validating a new reference standard for uncommon diseases in the population.

In the era of evidence-based medicine, the development of a new reference standard is expected to achieve both clinical and statistical validation. A new reference standard is accepted into practice if the proposed criteria are accurate in classifying patients with and without disease and demonstrate reproducible results that are generalizable to other target populations. However, once the new reference standard is implemented in practice, it is emphasized that it not only alters our diagnostic processes; but may also result in a definitional shift in whom we classify as having a disease. Thereby, an assessment of these clinical consequences on treatment decisions and patient outcomes is recommended in the validation process.

Acknowledgments

This publication was made possible by Grant Number 5K23NS058387-02 from the National Institute of Neurological Disorders and Stroke (NINDS), a component of the National Institutes of Health (NIH). Its contents are solely the responsibility of the authors and do not necessarily represent the official view of NINDS or NIH.

References

1. Alonzo TA, Pepe MS. Assessing the Accuracy of a New Diagnostic Test When a Gold Standard Does Not Exist. UW Biostatistics Working Paper Series 1998;3–32.
2. Altman DG, Royston P. What do we mean by validating a prognostic model? *Statistics in Medicine* 2000;19:453–473. [PubMed: 10694730]
3. Earnst F, Forbes G, Sandok BA, et al. Complications of diagnostic cerebral angiography: prospective assessment of risk. *Am J Roentgenol* 1984;142:247–253. [PubMed: 6198889]
4. Fraile M, Rull M, Julian FJ, et al. Sentinel node biopsy as a practical alternative to axillary lymph node dissection in breast cancer patients: an approach to its validity. *Annals of Oncology* 2000;11:701–705. [PubMed: 10942059]
5. Glasziou P, Irwig L, Deeks JJ. When should a new test become the current reference standard? *Annals of Internal medicine* 2008;149(11):816–821. [PubMed: 19047029]
6. Macdonald, LR.; Weir, B. Cerebral vasospasm. Academic Press; 2001. Medical aspects of vasospasm; p. 353-458.
7. Pfeiffer RM, Castle PE. With or without a gold standard. *Epidemiology* 2005;15(5):595–597. [PubMed: 16135933]
8. Pryor JC, Setton A, Nelson PK, et al. Complications of diagnostic cerebral angiography and tips on avoidance. *Neuroimaging Clin N Am* 1996;6(3):751–758. [PubMed: 8873102]
9. Rabinstein AA, Weigand S, Atkinson JLD, Wijdicks EFM. Patterns of cerebral infarction in aneurysmal subarachnoid hemorrhage. *Stroke* 2005;36:992–997. [PubMed: 15831836]
10. Reichman M, Greenberg E, Gold R, Sanelli P. Developing patient-centered outcome measures for evaluating vasospasm in aneurysmal subarachnoid hemorrhage. *Acad Radiol* 2009;16:541–545. [PubMed: 19345894]
11. Rutjes AWS, Reitsma JB, Coomarasamy A, et al. Evaluation of diagnostic tests when there is no gold standard. A review of methods. *Health Technology Assessment* 2007;11(50)
12. Shimoda M, Takeuchi M, Tominaga J, Oda S, Kumasaka A, Tsugane R. Asymptomatic versus symptomatic infarcts from vasospasm in patients with subarachnoid hemorrhage: Serial magnetic resonance imaging. *Neurosurgery* 2001;49:1341–1350. [PubMed: 11846933]
13. Simon R, Altman DG. Statistical aspects of prognostic factor studies in oncology. *British Journal of Cancer* 1994;6:979–985. [PubMed: 8198989]
14. Streiner DL, Norman GR. “Percision” and “accuracy”: two terms that are neither. *Journal of Clinical Epidemiology* 2006;59:327–330. [PubMed: 16549250]
15. ten Bosch JJ, Angmar-Mansson B. Characterization and validation of diagnostic methods. *Monographs in Oral Science* 2000;17:174–189. [PubMed: 10949840]