

International Journal of Technology Assessment in Health Care

Official Journal of the International Society for Technology Assessment in Health Care

Editors

Egon Jonsson, Ph.D.

Karolinska Institute, Department of Medicine,
and SBU (The Swedish Council on Technology
Assessment in Health Care)
P.O. Box 5650
S-114 86 Stockholm, Sweden
Telephone: 46-8-412-3200
Fax: 46-8-411-3260
E-mail: management@sbu.se

Stanley J. Reiser, M.D., Ph.D.
Professor, Program on Humanities
and Technology in Health Care
The University of Texas - Houston
Health Science Center
P.O. Box 20708
Houston, Texas 77225, U.S.A.
Telephone: 713 500-5080
Fax: 713 500-5088
E-mail: sreiser@heart.med.uth.
tmc.edu

Aims and Scope: The *International Journal of Technology Assessment in Health Care* has as its specific scope of interest the generation, evaluation, diffusion, and use of health care technology. It addresses the diverse audience of health care providers within medicine, nursing, and the allied health professions; decision makers in government, industry, and health care organizations; and the scholarly disciplines such as ethics, economics, law, history, sociology, psychology, and engineering. The journal aims to fill the needs of those interested in the complexities of interaction between people and technology; technology as a force in social and organizational change; and technology as it is created, produced, applied, and paid for. It examines descriptively and analytically the effects of technology as perceived from the vantage point of different academic disciplines and policy-making organizations and examines methods necessary to conduct studies and evaluations of technology.

The focus of the journal is the international health care community. The experience of different countries in their encounter with health care technology, viewed comparatively, is invaluable to understanding its effects. Further, it is important to establish ties with the scholars, governments, and private institutions concerned with health care technology so that the experiences and learning of the international community in its parts can benefit the whole. The use of technology in health care has created some of the major dilemmas in society today. The journal serves as a forum for the wide range of professionals interested in the assessment of medical technology, its consequences for patients, and its impact on society.

Membership in the International Society for Technology Assessment in Health Care:

Members receive the journal at individual rates and can participate in the Society's activities, including the annual meeting. Individual dues for 2000, including the journal, are US \$95, payable to ISTAHC, 759 Victoria Square, RC4, Montreal (QC) H2Y 2J7, Canada.

**Publishing, Production, and Advertising
Offices:** Cambridge University Press, 40 West
20th Street, New York, NY 10011, U.S.A.

Subscription Offices: Cambridge University
Press, 110 Midland Avenue, Port Chester, NY
10573, U.S.A. (for U.S.A., Canada, and Mexico);
or Cambridge University Press, The Edinburgh
Building, Shaftesbury Road, Cambridge CB2
2RU, England (for U.K. and elsewhere).

2000 Subscription Information: *International
Journal of Technology Assessment in Health
Care* (ISSN 0266-4623) is published quarterly.
Annual institutional subscription rates for
Volume 16 (2000): US \$183.00 in the U.S.A.,
Canada, and Mexico; UK £115.00 elsewhere.
Individual rates: US \$107.00 in the U.S.A.,
Canada, and Mexico; UK £65.00 elsewhere.
Prices include postage; air mail or
registered mail is extra. Back volume
prices are available upon request.

Copyright © 2000 Cambridge University Press

All rights reserved. No part of this publication
may be reproduced, in any form or by any
means, electronic, photocopying, or otherwise,
without permission in writing from Cambridge
University Press.

Photocopying Information (for users in the
U.S.A.): The Item-Fee Code for this publication
(0266-4623/00 \$9.50) indicates that copying
for internal or personal use beyond that
permitted by Sec. 107 or 108 of the U.S.
Copyright Law is authorized for users duly
registered with the Copyright Clearance Center
(CCC) Transactional Reporting Service,
provided that the appropriate remittance of
\$9.50 per article is paid directly to: CCC,
27 Congress Street, Salem, MA 01970 U.S.A.
Specific written permission must be obtained
for all other copying; contact the nearest
Cambridge University Press office.

Periodicals postage paid at New York, NY,
and additional mailing offices.

Postmaster: Send address changes in the U.S.A.
and Canada to: *International Journal of
Technology Assessment in Health Care*, Journals
Department, Cambridge University Press,
110 Midland Avenue, Port Chester, NY 10573,
U.S.A.

International Journal of Technology Assessment in Health Care, 16:3 (2000), 834-841.
Copyright © 2000 Cambridge University Press. Printed in the U.S.A.

SENSITIVITY ANALYSIS OF THE DIAGNOSTIC VALUE OF ENDOSCOPIES IN CROSS-SECTIONAL STUDIES IN THE ABSENCE OF A GOLD STANDARD

Sebastian Schneeweiss

Harvard Medical School and Harvard University School of Public Health

Abstract

Objectives: The evaluation of the diagnostic value of endoscopic procedures usually lacks a gold standard when performed in cross-sectional studies. The objective is to demonstrate an easily applicable method to assess the possible range of sensitivity, specificity, and predictive values of endoscopic procedures in the absence of a gold standard method.

Methods: Data from a study of 328 endoscopies comparing two different methods to diagnose superficial bladder cancer were used as a numerical example. Both endoscopic procedures were performed in the same patients in one session. Under the assumption of a systematic misclassification process, a model to correct sensitivity estimates is developed.

Results: The lowest possible sensitivity estimate for a new fluorescence endoscopy technique (FE) was 78%, the maximum 97.5%. Depending on realistic assumptions made upon the misclassification, a reasonable estimate for sensitivity was 93.4% (95% confidence interval [CI]: 90%–97.3%) for the FE technique. The sensitivity of the traditional white-light endoscopy method ranged from 47.2% to 53%, with a reasonable estimate of 46.7% (95% CI: 39.4%–54.3%).

Conclusions: This method to determine the theoretically possible range of sensitivity estimates in endoscopic procedures is helpful in cross-sectional studies with a missing gold standard method. It is easily applicable for a variety of endoscopic procedures, including upper and lower gastro-intestinal tract, urogenital tract, or diagnostic laparoscopic surgery.

Keywords: Endoscopy, Diagnostic test, Sensitivity analysis, Missing gold standard, Methods

The development of endoscopic procedures with high sensitivity is critical for early and effective medical treatment of many diseases. The knowledge and comparison of the diagnostic value of existing and new endoscopic procedures are important prerequisites for high-quality health care. The sensitivity of a diagnostic test is the proportion of detected

This paper was made possible by support from grant no. Schn527/3-1 and Schn527/4 by the Deutsche Forschungsgemeinschaft. Further support came from the Pharmacoepidemiology Research and Training Grant, Harvard University School of Public Health, and MEDAC GmbH. The author is grateful to Dr. M. Kriegmair for contributing the example data.

lesions that could have been detected had a perfect gold standard method been applied. However, for many diagnostic procedures in medical practice no gold standard method exists. A theoretical gold standard test for colon cancer would mean to remove the total colon and get a detailed pathologic and/or histochemical examination of the complete tissue. Such a procedure is not justifiable because of ethical reasons. Therefore, only biopsy tissue from endoscopies undergoes a pathologic work-up. In longitudinal studies, subjects can remain under diagnostic scrutiny and can be retrospectively classified as having had colon cancer at the point the original test was done, if the cancer becomes clinically apparent after a lag time period (3,161;5). The lead time must be defined often by using untestable biological assumptions and complicated models of the specific tumor growth.

Many recent studies evaluating urological or gastrointestinal endoscopic methods are cross-sectional studies, since they are less complex, less expensive, and less time-consuming than longitudinal studies (6;10;11;12;15;16;17). However, in cross-sectional studies it is even more difficult to identify a gold standard procedure. Investigators usually use the results of the most sensitive existing procedure as the standard of comparison, which might or might not be a good proxy of a gold standard, leading to biased results.

The objective is to provide an easily applicable approach to define an interval of sensitivity, specificity, and predictive value estimates in cross-sectional studies comparing endoscopic procedures that include the value had a true gold standard been available. The method will be illustrated with a comparison of a new aminolevulinic acid-induced fluorescence endoscopy (FE) versus a traditional white-light endoscopy (WLE) to detect superficial bladder cancer (8).

METHODS

Subjects

All consecutive patients with a total of 704 endoscopies admitted to the Department of Urology at the University of Munich Hospital Center in Großhadern between January 1995 and August 1996 were eligible; 337 endoscopies were excluded either due to patient refusal or improper documentation. An additional 39 endoscopies were excluded according to protocol criteria. A total of 208 patients with 328 endoscopies participated in the study, 72 were inspected for primary bladder cancer, and 136 were under surveillance because of prior transitional cell carcinoma of the urinary bladder.

More than one endoscopy session may be performed within a single patient. The unit of observation is an endoscopy. Although 38.5% of subjects had two or more endoscopies, the possibility of a correlation between these observations was disregarded to illustrate the method. Violations of this assumption would lead to falsely narrow confidence intervals, but would not affect sensitivity estimates.

Design

Methodologic details of the endoscopic procedures are described elsewhere (7). In this cross-sectional evaluation, a conventional WLE immediately preceded a 5-aminolevulinic acid-induced FE in the same session. FE succeeded WLE because it was assumed to be more sensitive from the experience of prior case series (7;9). Since FE and WLE were performed in the same session, all patients that gave informed consent and thus enrolled in the study underwent both types of endoscopy.

Statistical Methods

For the following analysis we apply the standard terminology used in the evaluation of diagnostic tests (Figure 1). Sensitivity is a proportion of the true positive patients over the

		True outcome		
Test outcome	Positive	Negative		
Positive	True positives	False positives	Test positives	
Negative	False negatives	True negatives	Test negatives	
	$n \times p^*$	$n \times (1-p)$		n

Figure 1. Standard notation for diagnostic tests. *p is the true prevalence of cancerous lesions in the endoscoped urinary bladders, and n is the total number of endoscopies (328).

sum of true positives plus false negatives. This proportion follows a binomial distribution. The 95% confidence intervals are calculated using the normal approximation (1,121). The sum of true positives and false negatives equals the true prevalence of bladder cancer (p) times the total number of study subjects (n).

A gold standard was not available as a reference for both endoscopic procedures. Only biopsy tissue underwent a pathologic work-up. The pathologic work-up is known to have a very high sensitivity and specificity, and is considered a gold standard for tissue specimens. However, since it is impractical and unethical to remove or biopsy the whole urinary bladder, there is an unknown likelihood of missing cancerous lesions because they were not biopsied. This likelihood can vary and may depend on the test's outcome, in case the test outcome guides the number and locations of biopsies (13). More assumptions and more complicated methods are necessary to account for such a bias (2).

Effects of Having No Gold Standard. In the reported study, a gold standard test is only available for regions or lesions that are actually biopsied. This can lead to misclassification among the four cells in Figure 1 and cause over- or underestimation within each cell.

Lesions counted in the true-positive cell did show positive in the test and were confirmed through a biopsy. It is very unlikely that any observation in the cell should apply to another cell, because a pathological work-up is generally assumed to be a gold standard method. The observed number is the most conservative estimate, since some cancerous areas might have been missed by the biopsy and thus appeared in the false-positive cell (see arrow A in Figure 2). The magnitude of this misclassification is assumed to be small but is unknown.

		True outcome	
Test outcome	Positive	Negative	
Positive	True positives A	False positives	
Negative	False negatives B	True negatives	
	$n \times p$	$n \times (1-p)$	

Figure 2. Possible misclassifications because of an insufficient gold standard measurement method. A: Some true positives will be falsely classified as false positives. B: Some false negatives will be falsely classified as true negatives.

Table 1. Final Results of 328 5-Aminolevulinic Acid-induced FE and WLE

Type	Number of endoscopies (%)	Number of biopsied areas per endoscopy, range (mean)	Result of FE
a	77 (23%)	1-10 (2.6)	All neoplastic lesions detected were fluorescence-positive and also visible under white light.
b	82 (25%)	1-11 (4.1)	Neoplastic lesions were detected additionally only because of fluorescence signal but not under white light.
c	97 (30%)	1-7 (2.2)	All fluorescing lesions that were biopsied were benign.
d	20 (6%)	1-10 (1.5)	The complete urothelium of the bladder was fluorescence-negative. All biopsies taken had been benign.
e	48 (15%)	0 (0)	The complete urothelium of the bladder was fluorescence-negative. No biopsies were taken.
f	4 (1%)	1-5 (1.9)	The complete urothelium of the bladder was fluorescence-negative. Additionally, fluorescence-negative small papillary tumors were found.
	328 (100%)	0-11 (2.2)	Total

Source: Kriegmair et al.

As indicated above, some lesions could be missed because not enough biopsies were performed. These additional observations, classified as false positives, would in fact belong in the true-positive cell (arrow A in Figure 2).

For the false-negative and true-negative cells, the mechanism is analogous. If an insufficient number of biopsies were performed, test-negative lesions become classified as negative, but in truth they were positive. The magnitude of this misclassification (see arrow B in Figure 2) is also assumed to be small, but larger than the misclassification in the test positives (arrow A), because fewer biopsies per endoscopy were taken when the endoscopy is fluorescence-negative and suggest an unaffected bladder (Table 1). This misclassification pattern is the same for both techniques.

Correction for Misclassification. From the misclassification model explained above, a method to correct the counts in a 2 by 2 table, and subsequently the sensitivity estimate, can be developed according to the matrix correction method (4). Let p(A) be the probability of false positives being falsely classified as such, and p(B) be the probability of true negatives being falsely classified as such. Then true positives = "true positives" + p(A) × "false positives"; false positives = "false positives" - p(A) × "false positives"; true negatives = "true negatives" - p(B) × "true negatives"; and false negatives = "false negatives" + p(B) × "true negatives." Terms in quotation marks are "classified as."

We can now derive a more general formula for estimating test sensitivity: Sensitivity = ("true pos." + p(A) × "false pos.") / ("true pos." + p(A) × "false pos." + "false neg." + p(B) × "true neg."). The magnitude of the misclassification, which can be differential with respect to the test result, is unknown. In the case of our example, experienced urologists claim that, in the case of bladder cancer, both p(A) and p(B) are arguably small, and p(A) is smaller than p(B).

Specificity, positive predictive values and negative predictive values can be estimated only for FE. The reason lies in a typical characteristic of this parsimonious study design: the WLE exam was always followed by the FE exam. All decisions of whether, where, and how many biopsies to take in macroscopically unaffected areas were exclusively made using FE

(compare type c to f in Table 1), which was *a priori* assumed to be more sensitive. Biopsies of areas that were not suggestive of cancerous lesions under WLE were not performed, and data are absent. This would have required a two-step approach with a complete work-up under WLE before starting FE, probably not in one session. There would be an additional source of bias, since any random biopsies taken under WLE might lead physicians away from those spots under the subsequent FE, if performed in one session. In our framework of misclassification, this means that there is another set of misclassification probabilities, $p(A')$ and $p(B')$, which describe what would happen if we would assume WLE as our observed gold standard. Since those data are not observed in our example, no specificity and predictive value estimates can be given for WLE.

There is no need to correct standard errors, because this approach works with structural assumptions of the misclassification with not-quantifiable errors attached. If more empirical information about the misclassification probabilities would be available from internal or external validation studies and used for the correction, the additional variation from those results must be considered.¹

RESULTS

Table 1 summarizes the results of 328 endoscopic procedures in the 208 study subjects. The mean number of biopsies varies from 0 to 4.1, depending on the macroscopic presentation of the bladder. Figure 3 represents data from Table 1 in a 2 by 2 table comparing test outcome with "true outcome." Test outcome is determined for FE or WLE; the classification of true outcome depends on the results of FE-guided biopsies. Results classified as type e (no biopsies performed) are also classified as "true negative," such as patients of type d.

According to our model, the number of 97 false positives might be too high, and some of the observations in this cell should be shifted to the 77 + 82 observations in the cell of the true positives. Conversely, some of the 48 + 20 true negatives should be added to the four false negatives. Further note that the additional 82 observations in the true-positive cell are only detected due to the FE procedure and would fall in the false-negative cell for the WLE procedure (type b in Table 1).

To determine the maximum range of sensitivity estimates, misclassification probabilities will be set to 0 and 1 each. A probability of 0 means that there is no misclassification; 1 assumes that all subjects in whom endoscopy did not suggest cancer actually had cancer. Table 2 presents sensitivity, specificity, and positive and negative predictive values for those scenarios.

"True outcome" according to FE-guided biopsy

Test outcome	Positive	Negative
Positive	77+82=159	97
Negative	4	48+20=68
	328 × p*	328 × (1-p)

Figure 3. Observed results for 5-aminolevulinic acid FE. *The true prevalence is unknown; 328 endoscopies were performed.

Table 2. Estimates of Sensitivity, Specificity, Positive and Negative Predictive Value, and 95% CI Depending on Different Structural Assumption of the Misclassification

Parameter	Estimate	95% CI
<i>p(A) = 0; p(B) = 0 (most optimistic assumption)</i>		
FE: Sensitivity	0.98	0.95-1.0
WLE: Sensitivity	0.47	0.40-0.55
FE: Specificity	0.41	0.34-0.49
FE: Positive predictive value	0.62	0.56-0.68
FE: Negative predictive value	0.94	0.89-1.0
<i>p(A) = 1; p(B) = 1 (most pessimistic assumption)</i>		
FE: Sensitivity	0.78	0.74-0.83
WLE: Sensitivity	0.53	0.48-0.58
FE: Specificity	— ^a	—
FE: Positive predictive value	1	1-1
FE: Negative predictive value	0	0-0
<i>p(A) = 0.05; p(B) = 0.11</i>		
FE: Sensitivity	0.93	0.90-0.97
WLE: Sensitivity	0.47	0.39-0.54
FE: Specificity	0.40	0.32-0.47
FE: Positive predictive value	0.64	0.58-0.70
FE: Negative predictive value	0.84	0.76-0.93
<i>p(A) = 0.10; p(B) = 0.22</i>		
FE: Sensitivity	0.98	0.95-1.0
WLE: Sensitivity	0.47	0.40-0.55
FE: Specificity	0.41	0.34-0.49
FE: Positive predictive value	0.62	0.56-0.68
FE: Negative predictive value	0.94	0.89-1.0

^a Not meaningful because only zero cells for "true" negative outcomes.

It is reasonable to assume that the magnitude of misclassification is inversely related to the average number of biopsies taken. Of the 97 observations classified as false positives, on average 2.2 biopsies per endoscopy were taken (compare type c in Table 1). Of the 48 classified as true negatives, on average 1.5 biopsies were performed, whereas from the remaining 20 in the same cell none was taken. Grouping both types together, on average one biopsy was performed. Therefore, the misclassification in the test negatives, $p(B)$, will be assumed to be 2.2-fold larger than in the test positives, $p(A)$. According to expert opinion, it is sufficiently conservative to assume that 5% of the observation could be falsely classified as false positives ($=p(A)$). Then the estimate of $p(B)$ would be 0.11. Table 2 also presents a less optimistic view with $p(A) = 0.1$ and $p(B) = 0.22$.

DISCUSSION

In clinical research endoscopic procedures are frequently compared in cross-sectional studies, performing two procedures sequentially within one session. We demonstrated a simply applicable and intuitive method to correct sensitivity estimates of diagnostic tests for misclassification under a variable set of structural assumptions when a gold standard method is absent. The method is presented for a frequently applied cross-sectional design that compares two endoscopic procedures sequentially in one session.

In our example, data show the superior sensitivity of FE to detect cancerous lesions compared with WLE. Even in a worst case scenario—assuming that in all endoscopies that did not suggest cancer, patients do have cancerous lesions—the sensitivity of FE appears superior by 25% points. The method can be analogously used to interpret the value of diagnostic endoscopies in the upper and lower gastrointestinal tract, or of diagnostic laparoscopy.

It should be noted that, with increasing misclassification from the most optimistic scenario to the most pessimistic, the difference between sensitivity estimates between FE and WLE is diminishing. This is analogous to effect attenuation due to random misclassification in fourfold tables (14, 127-131). Because of the particularly parsimonious one-session sequential design, specificity and positive and negative predictive values can be estimated only for the second, usually more sensitive endoscopic method, which guides biopsies in macroscopically unaffected areas.

The presented results assume that no other sources of bias might have led to systematic distortions of the sensitivity estimates. Many more problems can bias results of studies evaluating diagnostic tests that rely partially on the examiner's subjectivity (e.g., no blinding of examiner to prior clinical information or no blinding with respect to the method used) (13). However, it is not the objective of this paper to discuss the overall validity of the particular results.

In conclusion, the paper illustrates an easily applicable method for correction of a specific misclassification problem frequently found in the diagnostic evaluation of endoscopic procedures, which will support a more realistic interpretation of their diagnostic value for researchers and practitioners.

NOTE

¹A self-explanatory spreadsheet (MS Excel) for the calculations of this method can be obtained by sending an e-mail to the author: schneeweiss@post.harvard.edu.

REFERENCES

1. Armitage P, Berry G. *Statistical methods in medical research*, 3rd ed. Oxford: Blackwell, 1994.
2. Diamond GA. Of Bayes: Effect of the verification bias on posterior probabilities calculated using Bayes' theorem. *Medical Decision Making*. 1992;12:22-31.
3. Fletscher RH, Fletscher SW, Wagner EH. *Clinical epidemiology*, 2nd ed. Baltimore: Williams & Wilkins; 1988.
4. Greenland S, Kleinbaum DG. Correcting for misclassification in two-way tables and matched-pair studies. *Int J Epidemiol*. 1983;12:93-97.
5. Hosokawa O, Tsuda S, Kindani E, et al. Diagnosis of gastric cancer up to three years after negative upper gastrointestinal endoscopy. *Endoscopy*. 1998;30:721-723.
6. Kimchi NA, Mindru V, Broide E, Scapa E. The contribution of endoscopy and biopsy to the diagnosis of periampullary tumors. *Endoscopy*. 1998;30:538-543.
7. Kriegmair M, Baumgartner R, Kneuchel R, et al. Fluorescence photodetection of neoplastic urothelial lesions following intravesical instillation of 5-aminolevulinic acid. *Urology*. 1994;44:836-841.
8. Kriegmair M, Stepp H, Stepp H, et al. Transurethral resection and surveillance of bladder cancer supported by 5-aminolevulinic acid-induced fluorescence endoscopy. *Eur Urol*. 1999;36:386-392.
9. Kriegmair M, Waidelich R, Lumper W, et al. Integral photodynamic treatment of refractory superficial bladder cancer. *J Urol*. 1995;154:1339-1341.
10. Matsushita M, Hajiro K, Okazaki K, Takakuwa H, Tominaga M. Efficacy of total colonoscopy with a transparent cap in comparison with colonoscopy without the cap. *Endoscopy*. 1998;30:444-447.
11. Murphy WM, Rivera-Ramirez I, Medina CA, Wright NJ, Wajzman Z. The bladder tumor antigen (BTA) test compared to voided urine cytology in the detection of bladder neoplasms. *J Urol*. 1997;158:2102-2106.
12. Planz B, Striepecke E, Jakse G, Bocking A. Use of Lewis X antigen and deoxyribonucleic acid image cytometry to increase sensitivity of urinary cytology in transitional cell carcinoma of the bladder. *J Urol*. 1998;159:384-387.
13. Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. *JAMA*. 1995;274:645-651.

14. Rothman KJ, Greenland S. *Modern epidemiology*, 2nd ed. Philadelphia: Lippincott-Raven, 1998.
15. Stepp H, Sroka R, Baumgartner R. Fluorescence endoscopy of gastrointestinal diseases: Basic principles, techniques, and clinical experience. *Endoscopy*. 1998;30:379-386.
16. Tribl B, Turetschek K, Mostbeck G, et al. Conflicting results of ileoscopy and small bowel double contrast barium examination in patients with Crohn's disease. *Endoscopy*. 1998;30:339-344.
17. Wiener HG, Mian C, Haitel A, et al. Can urine bound diagnostic tests replace cystoscopy in the management of bladder cancer? *J Urol*. 1998;159:1876-1880.