J Clin Epidemiol Vol. 51, No. 12, pp. 1293–1298, 1998
Copyright © 1998 Elsevier Science Inc. All rights reserved.

0895-4356/98/$–see front matter
PII S0895-4356(98)00127-9

ELSEVIER

Evaluation of Diagnostic Imaging Tests: Diagnostic Probability Estimation

Olli S. Miettinen,[1,*] *Claudia I. Henschke,*[2] *and David F. Yankelevitz*[2]

[1]Department of Epidemiology and Biostatistics, McGill University, Montreal, Canada; and [2]Department of Radiology, The New York Hospital-Cornell Medical Center, New York, New York

ABSTRACT. In the evaluation of a diagnostic imaging test for the diagnosis of a particular illness in a particular category of patients, the test should be construed as leading to a test result in the sense of a set of descriptive readings from the image(s), not interpretation of these; and in the evaluation of the test, therefore, the first challenge is the translation of each test result (set of readings) into the corresponding probability that the illness is present. This interpretive translation should not be subjective, nor should it be based on an objective algorithm founded on clinical judgments. Instead, a suitable diagnostic probability function (of the elements in the test result) should be derived empirically by logistic regression analysis of suitable data. We illustrate this alternative outlook by reanalysis of the data from the Prospective Investigation of Pulmonary Embolism Diagnosis. J CLIN EPIDEMIOL 51;12:1293–1298, 1998. © 1998 Elsevier Science Inc.

KEY WORDS. Diagnosis, imaging, probability, test evaluation, logistic regression, Bayes' theorem

INTRODUCTION

Any evaluation of a diagnostic test has to do with a particular generic context of its potential application: concern to learn about the presence of a particular illness in a particular domain of presentation for testing. Thus, for ventilation-perfusion (V-Q) scanning of the lungs, the evaluation might focus on the diagnosis of pulmonary embolism (PE) in the patient giving rise to a suspicion for this illness by a specified set of domain-defining criteria.

For whichever context, evaluation must focus on a particular conceptual variant of the test. Thus, as for V-Q scanning in this context, the concept of the test without further specifications is so vague that one does not know even the broadest nature of its results: is it images per se, descriptive readings or data based on these (such as number of mismatched defects), or interpretation of the images or data with respect to presence of the illness (such as "low probability" of PE)? In other words, without such specification, it is unclear, even, where the test ends and the interpretation of its result begins. The choice among these three conceptualizations of an imaging test is, in and of itself, already a major basis for divergent outlooks on the evaluation of imaging tests.

Another important basis for divergence of outlooks relates to the theoretical framework for diagnosis and, hence, for diagnostic research. It was the radiologist Lusted who, in collaboration with Ledley, introduced the Bayes' theorem framework for this [1]. Yet, an alternative theoretical framework [2] deserves attention, one that in the context of diagnostic tests has particular merit with respect to imaging tests on the grounds that they produce descriptive readings or data on multiple aspects of the image(s).

In what follows, we outline very briefly the outlook that now prevails in the evaluation of diagnostic imaging tests, present critical questions about it, and then outline and justify the proposed alternative approach to setting diagnostic probabilities. We illustrate the prevailing outlook by the Prospective Investigation of Pulmonary Embolism Diagnosis (PIOPED) [3] and the alternative by reanalysis of the PIOPED data.

THE PREVAILING OUTLOOK

The PIOPED was an eminent, multicenter study about the presence of PE in the domain of adults in whom symptoms suggestive of PE were present within the most recent 24 hours and prompted a request for radiologic assessment. The radiologic test at issue was V-Q scanning in conjunction with chest roentgenography [3,4].

The definition of the V-Q test under evaluation involved three sequential elements:

*Address correspondence to: Olli S. Miettinen, M.D., Ph.D., Department of Epidemiology and Biostatistics, McGill University, 1020 Pine Avenue West, Montreal, PQ H3A 1A2, Canada.

Accepted for publication on 4 June 1998.

1. *Production* of the images (imaging proper)—when to produce them (recency of symptoms) and how (equipment and its use)
2. *Reading* of the images—what to read (e.g., number of moderate segmental perfusion defects without . . .); who would read it (two "nuclear medicine readers" with their two levels of backup "adjudicators"); and under what conditions (presumably, unaware of preimaging data and angiographic findings)
3. *Interpretation* of the readings—their translation (by an algorithm adopted for the study) into overall result on a predefined unidimensinal ordinal scale of probability (categories of "high probability," etc.) for the illness (PE) being present.

The reading of the images results in more-or-less objective facts documented by the readers, such as the number of large, moderate, and small segmental mismatched perfusion defects. The interpretation of the readings, on the other hand, is a judgment based on these facts, addressing the probability of presence of PE. For example, in the PIOPED, the finding of two large segmental perfusion mismatches or the finding of 10 resulted in the same interpretation of "high probability" of PE. Under evaluation in the PIOPED was, more specifically, the "diagnostic usefulness" of the interpretation.

This type of test conceptualization—the test result not being the readings per se (item 2 in the elements just listed) but the interpretation of these (item 3)—was by no means peculiar to the PIOPED; it is used quite routinely in evaluations of diagnostic imaging. In terms of this conceptualization, the test is supposed to result in a diagnostic classification—positive result implying presence of the illness and negative result implying its absence—and interest tends to focus on the "accuracy" of this classification [5–11].

With the definitive diagnosis based on angiography, the key results of the study were the empirical values for test "sensitivity" (true-positive rate) and "specificity" (true-negative rate) corresponding to each of three different definitions of positive test result: "high probability," "high or intermediate probability," and "high, intermediate, or low probability" [3]. These results of the study are readily convertible to estimates of the likelihood ratios corresponding to each of the four possible test results ("high probability," etc.). When coupled with whatever prior probability exists for PE, these are taken to provide for calculation of the probability that PE is present, using Bayes' theorem [7].

As for comparison of alternative tests, equally ingrained is that idea of using a predefined ordinal scale of test results, such as the one in the PIOPED, and then proceeding to comparison in terms of the respective areas under the "receiver operating characteristic" (ROC) curve [8–11].

CRITICAL QUESTIONS

Taking some distance from this prevailing outlook and culture in the evaluation of diagnostic imaging tests, two important, interrelated questions arise. First, would it not be much more natural to take the development of categories of illness probability ("high probability," etc.)—insofar as they are of interest at all—to be the first-order objective of the study rather than an *a priori* constraint for it? In other words, why define the readings-based categories of illness probability in advance of the study, given that consensus on these categories in the absence of the actual data is difficult at best? Second, is it not better to ignore completely such categories, to adopt that alternative outlook even more radically, and thus to address the question of how, in the domain of the study, the prevalence of the illness is a joint function of the readings on the images, possibly together with documented diagnostic indicators other than those from the imaging? Were such a function to be addressed, the first-order result of the study would provide for deriving an evidence-based, actual diagnostic probability estimate from any given set of readings (descriptive) from the images (possibly together with the associated other information on the patient).

Our main concern in this article is to argue that, especially in the context of imaging diagnostics, the prevailing outlook should give way to the alternative one. Toward this end, we present reanalyses of the PIOPED data from the alternative vantage and address the theoretical issues surrounding these.

THE ALTERNATIVE OUTLOOK: ELEMENTS

The PIOPED "interpretation categories" were defined on the basis of the following input readings/data [3]:

- Number of large segmental (i.e., 75% or more of a segment) perfusion defects that were mismatched (i.e., without corresponding ventilation or roentgenographic abnormalities or substantially larger than these)
- Number of moderate segmental (i.e., 25%–75%) mismatched perfusion defects
- Number (0, 1–3, 4+) of small segmental (i.e., 25% or less) mismatched perfusion defects with normal roentgenogram
- Presence (yes, no) of any perfusion defect with a corresponding but substantially larger roentgenographic abnormality
- With respect to large or moderate segmental perfusion defects with matching ventilation defects (equal or larger in size), with or without matching roentgenographic defects (none or substantially smaller), maximum number of involved segments
  - in a single lung, and
  - in a single lung region
- Presence (yes, no) of nonsegmental perfusion defect(s) (effusion, cardiomegaly, etc.)
- Presence (yes, no) of exact correspondence between perfusion outlines and the shape of the lungs as seen on the roentgenogram.

We defer to the PIOPED investigators' judgment that these are the descriptors to be abstracted from the images, that is, that the PE-diagnostic information in the images is, as a practical matter, imbedded in these descriptors of the images. Judgments of this type are, after all, unavoidable with respect to whatever category of inputs—history, physical examination, blood chemistry—in diagnostic research, and by no means is it our purpose here to advocate alternative radiologic judgments.

On the other hand, we do not share the PIOPED investigators' judgment as for how to translate these component descriptors of the images into an index, unidimensional, of the images' suspiciousness, of the degree to which they point to the presence or absence of PE. Instead, as was pointed out earlier here, the alternative outlook calls for learning about this integration from the study data, with this inquiry at the very core of the study concerns in this framework. So long as this option exists, clinging to pre-study classification schemes in research is contrary to the very idea of what research is about.

This problem is familiar from various contexts in medical and other sciences. The beginning is to translate the descriptors into a set of numerical "variates." In this, a binary descriptor typically becomes an "indicator" variate taking on value 1 for a designated one of the two categories, for the category that it "indicates," 0 otherwise; if the descriptor has three categories, an indicator is defined for two of these, etc.; and the tally of something might translate to two variates, a qualitative one indicating that the tally is at least 1, and a quantitative one representing the number per se or a transform of this; and, analogously for whatever other type of quantitative descriptor. With the variates (X_i) thus defined, the concern classically has been to find a suitable "weight"—coefficient (b_i)—for each, leading to a linear discriminant score $S = a + b_1X_1 + b_2X_2 \ldots$, one whose distribution is maximally different between the categories of interest, here presence and absence of PE [12]. In the last three decades, this classical outlook has become a subissue under the broader topic of logistic regression analysis [13,14] based on the defined variates, providing not only the discriminant scoring function but also its associated probabilities for the discriminated categories (e.g., PE present versus PE absent). It is a particular feature of logistic formulation for the regression model that the coefficients (b_i) in the linear "argument" ($a + b_1X_1 + b_2X_2 \ldots$) in it are invariant over the prevalence of the category of interest (PE) in the study experience; only the "intercept" (a) depends on this.

In this spirit, we fitted the logistic regression model specified in Appendix 1 to the PIOPED data on the 731 instances in which the V-Q test was satisfactorily carried out and the definitive diagnosis was established by angiography (supplemented by follow-up). Care was taken not to involve too many variates in the specification of that full model so as to avoid the problem of "over-fitting." In that full model, provision was made for different degrees of relevance for increasing number of mismatched perfusion defects, *a priori* presuming that a linear term would represent too strong a dependence. That model was then reduced in the usual stepwise fashion until all of the remaining coefficients differed "significantly" (two-sided $P < 0.10$) from zero. The result was, *a priori*, of the form

$$P = 1/(1 + e^{-S}), \qquad \textbf{(1)}$$

where P represents the estimated proportion with PE (the estimated probability of the presence of PE), e the base (2.72) of natural logarithms, and S the linear discriminant scoring function from the regression analysis. The latter took the form of:

$$S = -2.02 + 2.61(X_1)^{1/8} - (0.12 + 0.41X_3)X_2 + 0.65X_4 + 0.34X_5; \qquad \textbf{(2)}$$

X_1 = number of large or moderate perfusion defects, mistmatched

X_2 = maximum number of matched perfusion defects in any lung region

X_3 = indictor (see above) of X_1 greater than or equal to 1

X_4 = indicator of X_2 greater than or equal to 1

X_5 = indicator of presence of any perfusion defect with a corresponding but substantially larger roentgenographic abnormality

The values for the probability of PE that Equation 1 yields with this scoring function range from 12% to 90%, with 25th, 50th, and 75th centiles of the distribution equal to 15%, 17%, and 67%, respectively, based on the PIOPED data. For selected ranges of the estimated probability, Table 1 shows the actual frequencies (prevalence) of PE, together with the corresponding numbers of patients. It is seen that this diagnostic function placed 29% of the patients in the greater than 60% probability range (a high probability of PE relative to the overall prevalence of 34%), and 60% of them in the less than 20% range (a low probability of PE in the same sense), thus leaving only 11% in the intermediate, 20 to 60% range.

For comparison, Table 2 shows the results of our application of the *a priori* classification used in the PIOPED. Strikingly, while 18% of the patients fell in the "high probability" category and 85% of these had PE, the majority, 62%, of the patients fell in the "intermediate" probability category, in which the prevalence of PE was 25%; and only 20% of the patients fell in the two lowest-probability categories, representing, approximately, the less than 20% probability range. The pattern is distinctly less attractive than that in Table 1.

The probability estimates provided by Equation 2 (in conjunction with Equation 1) are predicated on the overall prevalence of PE in the dataset leading to the scoring function in Equation 2, this prevalence—the mean $\overline{P}$ of the

TABLE 1. Prevalence of PE by range of estimated probabilty of PE from a logistic regression model (Equations 1 and 2) together with the corresponding input numbers of patients, in the PIOPED experience

Estimated probability	PE prevalence, %	Patients PE	Patients No PE	Patients Total
80%+	89	67 (27%)	8 (2%)	75 (10%)
60–80%	66	91 (36%)	47 (10%)	138 (19%)
40–60%	62	13 (5%)	8 (2%)	21 (3%)
20–40%	21[a]	12 (5%)	45 (9%)	57 (8%)
15–20%	18	42 (17%)	191 (40%)	233 (32%)
10–15%	13[b]	26 (10%)	181 (38%)	207 (28%)
Any	34	251 (100%)	480 (100%)	731 (100%)

[a]Forty-six of the 57 patients had estimates in the 20 to 25% range.

[b]One hundred twenty-six of the 207 patients had the smallest possible probability of PE, obtainable from the function, 12%.

Abbreviations: PIOPED = Prospective Investigation of Pulmonary Embolism Diagnosis; PE = pulmonary embolism.

posttest estimates based on the data—being 251/731 = 0.34. If the "practice-specific" overall prevalence is $\bar{P}'$ ($\neq \bar{P}$), then the scoring function in Equation 2 requires adjustment by the addition of log $[\bar{P}'/(1 - \bar{P}')] - \log[\bar{P}/(1 - \bar{P})] = \log[\bar{P}'/(1 - \bar{P}')] + 0.65$.

As an example of diagnostic probability estimation, then, the local prevalence of PE in the PIOPED diagnostic domain might be 25% (as best is known), and the V-Q scan might show five mismatched segmental perfusion defects representing at least 25% of the segment, as many as three matched defects in a single lung region and no perfusion defect under a roengenographic abnormality. Referring to Equation 2, the intercept (−2.02) needs to be adjusted by adding log $[0.25/(1 - 0.25)] + 0.65 = -0.45$; $X_1 = 5$, $X_2 = 3$, $X_3 = 1$, $X_4 = 1$, and $X_5 = 0$. In these terms, $S = -0.22$, so that antilog $(-S) = 1.24$ and, thus, $P = 1/(1 + 1.24) = 0.45$ as the estimate of the probability of PE.

THE ALTERNATIVE OUTLOOK: EXTENSIONS

In accordance with the spirit of the PIOPED, addressed earlier here was the situation in which the radiologist expresses diagnostic probability on the basis of the radiologic data alone. Yet, ultimately the diagnostic probability that guides the decision about intervention is based on added inputs from the patient's history and physical examination as well as tests other than imaging. Some aspects of history are relevant to differential risks for the illness at issue and its differential-diagnostic alternatives, whereas the other aspects of history, together with the physical examination and nonimaging tests, akin to imaging, address differential manifestations between the illness and its alternatives.

For the purposes of the ultimate diagnosis, the diagnostic probability function that addresses the risk information as well as the nonimaging manifestational information jointly with the imaging information (readings) is an obvious extension of what is presented here—in principle, that is. Realistically, though, the radiologist is concerned to understand how to integrate the imaging readings into an imaging score S_i, and to use this for entirely imaging-based diagnosis, whereas the clinician in direct charge of the patient might first focus on the rest, on the integration of the risk and preimaging manifestational information into a clinical score S_c; and the latter, even, might best be ad-

TABLE 2. Prevalence of PE by PIOPED's *a priori* categories[a] of interpretation (probability of PE), together with the corresponding input numbers of patients, in the PIOPED experience

A *priori* probability	PE prevalence, %	Patients PE	Patients No PE	Patients Total
"High"	85	112 (45%)	19 (4%)	131 (18%)
"Intermediate"	25	115 (46%)	340 (71%)	455 (62%)
"Low"	17	19 (8%)	92 (19%)	111 (15%)
"(Near-)normal"	15	5 (2%)	29 (6%)	34 (5%)
Any	34	251 (100%)	480 (100%)	731 (100%)

[a]Our categorization based on the PIOPED definitions and input data.

Abbreviations: PE = pulmonary embolism; PIOPED = Prospective Investigation of Pulmonary Embolism Diagnosis.

dressed in terms of its components, a risk score S_r and a pre-imaging manifestational score S_m.

Given such a segmentation of the research issues, leading to the component scores S_r, S_m, and S_i, the ultimately relevant overall scoring function can be developed by fitting a regression model (logistic) with the independent variates (Xs) defined in terms of these scores. The fitting would inherently account for whatever degree of redundancy there is among these three input elements in terms of information about the presence or absence of the illness at issue.

DISCUSSION

Our orientational proposition is that diagnostic interpretation of the readings from a (set of) diagnostic image(s) should not be construed as part of the test itself. Instead, the test should be construed as ending with the readings (descriptive) constituting the test result.

Given this conceptualization of an imaging test in diagnosis, we strongly propose that *a priori* definition of a scale (unidimensional) a result interpretation should be replaced by logistic regression analysis of the data on image-descriptors, leading empirically to unidimensional scoring of the multidimensional result and ready translation of this score to diagnostic probability estimate, possibly with practice-specific adjustment of the "intercept" in the scoring function. The logistic regression framework would not only substitute for the *a priori* definition of the scale of result interpretation but also for the Bayes' theorem framework [1,5] in setting the posttest probability, as the fitted logistic regression function in and of itself implies the posttest probability.

Insofar as one wants to think in terms of Bayes' theorem, it is to be noted that the posttest probability in these terms can be put to the form of Equation 1, with $S = \log [\overline{P}'/(1 - \overline{P}')] + \log (LR)$, where $\overline{P}'$ is the local pretest probability in the sense of the local overall prevalence (or the arithmetic mean of the posttest probabilities) and *LR* is the likelihood ratio—the probability of the result among cases divided by that among noncases of the illness at issue. This means that in the Bayes' theorem framework of setting diagnostic probabilities, the requisite *LR* inputs are readily derived by visiting the logistic regression framework: $LR = S - \log [\overline{P}/(1 - \overline{P})]$, where $\overline{P}$ is the overall prevalence in the study data leading to the scoring function S.

Given that the use of the Bayes' theorem framework requires the use of the logistic regression approach to address to requisite likelihood ratios, there really is no point in retaining the Bayes' theorem framework for setting diagnostic probabilities. The logistic regression framework is sufficient in itself, and the Bayes' theorem framework is unworkable without input from the logistic regression framework so long as the test result is multidimensional, as is characteristic of diagnostic imaging tests in general.

We thank H. Dirk Sostman, M.D., for providing us with access to the Prospective Investigation of Pulmonary Imbolism Diagnosis database and for helpful discussions on the manuscript.

References

1. Ledley RS, Lusted LB. Reasoning foundations of medical diagnosis. **Science** 1959; 130: 9–21.
2. Dawid AP. Properties of diagnostic data distributions. **Biometrics** 1976; 32: 647–658.
3. The PIOPED Investigators. Value of the ventilation/perfusion scan in acute pulmonary embolism: Results of the Prospective Investigation of Pulmonary Embolism Diagnosis (PIOPED). **JAMA** 1990; 263: 2753–2759.
4. Gottschalk A, Juni JE, Sostman HD, Coleman RE, Thrall J, McKusick KA, *et al.* Ventilation-perfusion scintigraphy in the PIOPED study. I. Data collection and tabulation. **J Nucl Med** 1993; 34: 1109–1118.
5. Sackett DL. A primer on the precision and accuracy of the clinical examination. **JAMA** 1992; 267: 2638–2644.
6. Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? **JAMA** 1994; 271: 389–391.
7. Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? **JAMA** 1994; 271: 703–707.
8. Tempany CM, Zhou X, Zerhouni EA, Rifkin MD, Quint LE, Piccoli CW, *et al.* Staging of prostate cancer: Results of Radiology Diagnostic Oncology Group project comparison of three MR imaging techniques. **Radiology** 1994; 192: 47–54.
9. Zerhouni EA, Rutter C, Hamilton SR, Balfe DM, Megibow AJ, Francis IR, *et al.* CT and MR imaging in the staging of colorectal carcinoma: Report of the Radiology Diagnostic Oncology Group II. **Radiology** 1996; 200: 443–451.
10. Panicek DM, Gatsonis C, Rosenthal DI, Seeger LL, Huvos AG, Moore SG, *et al.* CT and MR imaging in the local staging of primary malignant musculoskeletal neoplasms: Report of the Radiology Diagnostic Oncology Group. **Radiology** 1997; 202: 237–246.
11. Mushlin AI, Detsky AS, Phelps CE, O'Connor PW, Kido DK, Kucharczyk W, *et al.* The accuracy of magnetic resonance imaging in patients with suspected multiple sclerosis. The Rochester-Toronto Magnetic Resonance Imaging Study Group. **JAMA** 1993; 269: 3146–3151.
12. Tourassi GD, Floyd CE, Coleman RE. Improved noninvasive diagnosis of acute pulmonary embolism with optimally selected clinical and chest radiographic findings. **Acad Radiol** 1996; 3: 1012–1018.
13. Truett J, Cornfield J, Kannel W. A multivariate analysis of the risk of coronary heart disease in Framingham. **J Chron Dis** 1967; 20: 511–524.
14. Walker SH, Duncan DB. Estimation of the probability of an event as a function of several independent variables. **Biometrika** 1967; 54: 167–179.

APPENDIX 1

In addition to the Xs in Expression 2, including $X_6 = X_2X_3$, we defined the following independent variates:

$X_7 = X_1X_4$

$X_8 = X_3X_5$

$X_9 = X_4X_5$

$X_{10} = X_1^{1/2}$

X_{11} = $X_1^{1/4}$

X_{12} = number of large mismatched perfusion defects divided by the sum of the numbers of large and intermediate ones

X_{13} = number of small mismatched perfusion defects with normal roentgenogram coded 0, 1, and 2 for 0, 1-3 and 4+, respectively, and multiplied by X_3

X_{14} = indicator of nonsegmental perfusion defects(s) multiplied by X_3

X_{15} = maximum number of matched perfusion defects in a single lung multiplied by X_3

No data are available as to nonsegmental perfusion defects with exact correspondence between perfusion outlines and the shape of the lungs as seen on the roentgenogram.

Stepwise reduction of the full model having led to the retention of the variates in Expression 2 (with X_3 deleted but $X_6 = X_2X_3$ retained), each of the deleted variates was reentered alone as an additional one, but none of them made a significant (one-sided $P < 0.05$) contribution.

With the final model, the standard errors associated with the intercept and the coefficients of $X_1^{1/8}$, X_2, X_4, X_5, and X_6 were 0.21, 0.21, 0.13, 0.30, 0.20, and 0.15.