Journal of Biomedical Optics 5(2), 119–130 (April 2000)

Principles and pitfalls of diagnostic test development: implications for spectroscopic tissue diagnosis

Maryann Fitzmaurice
University Hospitals of Cleveland
Pathology Department
Cleveland, Ohio 44106
and Case Western Reserve University,
Cleveland, Ohio

Abstract. Diagnostic spectroscopy has the potential to supplant the time-honored ''gold standard'' of light microscopy and herald an era of *in vivo* tissue diagnosis. However, the lessons in disease diagnosis learned by pathologists over the years should not be forgotten. This discussion will focus on the basis principles and pitfalls of diagnostic test development, and how they apply to optical spectroscopy tissue diagnosis.
[S1083-3668(00)00102-7]

Keywords: biomedical optics; spectroscopy; disease diagnosis; test development.

Paper JBO-42006 received Sep. 1, 1999; revised manuscript received Dec. 17, 1999; accepted for publication Dec. 17, 1999.

1 Introduction

The compound microscope was invented in the 17th century. But, it was not until the mid-19th century that refinements in the light microscope and the introduction of the mechanical microtome to cut thin tissue sections allowed examination of human tissues with sufficient resolution to describe their normal histology and disease pathology. Even then, microscopic examination of diseased tissue was largely an academic exercise and was thought by many in the medical community to be not only clinically unnecessary but even counterproductive. In fact, in 1853, Velpeau, a prominent breast surgeon of his day, said ''the intervention of the microscope is not at all necessary to decide whether such and such tumor, which has been removed, is or is not of a cancerous nature.'' [1] And, only two years later, Virchow, regarded by many as the father of academic pathology, said ''it must be understood that in addition to applied (diagnostic) microscopy, there is scientific microscopy. What in the end will be of importance in the development of medicine is whether the microscope proves to be an agent *merely* of diagnosis or truly of reform.'' [2]

Interestingly, it was surgeons and not pathologists who first became convinced that diagnostic microscopy was clinically useful. Carl Ruge and Johann Weit, both gynecologic surgeons at the University of Berlin, were the first to espouse the use of microscopy for preoperative tissue diagnosis in their studies of carcinoma of the uterine cervix, a major focus of spectroscopic tissue diagnosis today. The first hysterectomy for cervical cancer was performed in Breslow in 1878, providing the opportunity for cure. But, given the morbidity and mortality of the then primitive surgical procedure, it was particularly important to avoid preoperative misdiagnosis and unnecessary surgery. In 1880, Ruge and Weit reported that of the first 23 hysterectomies ever performed for presumptive cervical cancer, only 13 had a correct preoperative clinical diagnosis when compared to the microscopic postoperative diagnosis. They suggested that preoperative microscopic examination of uterine scrapings (what we would today call an endocervical or endometrial curettage) should be used to confirm the clinical suspicion of malignancy prior to hysterectomy.[3]

So, ironically, the first surgical pathologists were surgeons and not pathologists. And, the first surgical pathology laboratories arose primarily in departments of surgery and not pathology. In fact, Arthur Purdy Stout, the founder and namesake of what is today the American Society of Surgical Pathologists, began his career in the Department of Surgery in what would later become Columbia–Presbyterian Medical Center in New York.[4]

Eventually, pathologists began to assume a major role in surgical pathology. And, in 1898, one of the first ever written reports of the microscopic pathology of a surgically excised tissue was issued by William Travis Howard in the newly created Department of Pathology at Cleveland Lakeside Hospital, later to become University Hospitals of Cleveland (UHOC).[5]

But, true integration of surgical pathology into everyday clinical practice would await technological advancements. One of the firsts in the field of tissue diagnosis came shortly after World War I with the development of the cryostat, a specialized microtome that could be used to prepare frozen tissue sections for intraoperative diagnosis. With this development, the importance of microscopic tissue diagnosis became more widely accepted.

However, it was not until after World War II that surgical pathology laboratories came under the control of trained pathologists in most American hospitals. And, even today, some clinicians (especially dermatologists) still insist on doing their own pathology. Humphreys, a surgeon at Columbia University, probably said it best when he said ''surgical pathology was born out of necessity and out of wedlock'' and was never acknowledged by its father (pathology).[6,7]

Since that time, there have been a number of advancements in tissue diagnosis. Hematoxylin and eosin have been used as the standard histochemical stain for microscopic tissue diagnosis since the mid-1800s.[8] It is essentially a contrast agent, which combines a basophilic natural dye (hematoxylin) with an acidophilic counterstain (eosin) to give contrast to tissue that is essentially transparent microscopically if un-

Address all correspondence to Dr. Maryann Fitzmaurice.
E-mail: mxf39@po.cwru.edu

1083-3668/2000/$15.00 © 2000 SPIE

stained. However, in the past 20 years, a large inventory of specialized, enzyme- and immunohistochemical stains have been developed to identify specific chemical moieties within the tissue and, thereby, improve diagnosis. The most recent advances in tissue diagnosis have come largely in the form of ancillary studies performed *in vitro* to support a microscopic tissue diagnosis, and include electron microscopy, morphometry (computer-assisted quantitative image analysis), ploidy analysis, and molecular biology techniques.

Similar advances have occurred in methods of tissue fixation. Formalin (or formaldehyde) has been the fixative of choice since the late 1800's.[8] However, formalin alters the chemistry of the tissue by cross linking its proteins and is, therefore, incompatible with many of the special stains and other ancillary studies that have come into common use. So, fixation methods have evolved to keep pace, and a number of non-formalin-based fixatives are also currently used for microscopic tissue diagnosis.

And now, only 50 years or so since the practice of surgical pathology became routine, techniques such as optical spectroscopy are being developed as an alternative to microscopic tissue diagnosis. These techniques offer the potential for real-time *in vivo* tissue diagnosis, a possibility that could revolutionize the clinical diagnosis of disease and ultimately the practice of pathology. Hopefully, pathologists will more readily acknowledge this offspring and be more willing to foster optical spectroscopy diagnosis, than they were to adopt microscopic disease diagnosis, as the way of the future.

2 Diagnostic Principles and Pitfalls

Many of the early studies of optical spectroscopy as a diagnostic technique were small-scale proof-of-principle studies, intended primarily to show that optical spectroscopy could be performed *in vivo* and information obtained that could be the basis of a clinically useful diagnostic test. These studies have clearly shown the diagnostic potential of several types of spectroscopy, including reflectance, light-scattering, fluorescence, and even Raman spectroscopy, for tissue diagnosis in a variety of clinical settings, including atherosclerotic cardiovascular disease,[9–14] premalignant lesions in the bronchopulmonary tree,[15–17] upper aerodigestive tract,[18] gastrointestinal tract,[19–22] and female genital tract,[23–25] breast cancer[26,27] and other solid tumors, and even degenerative neurologic diseases such as Alzheimer's.[28] This work has recently been reviewed in detail.[29,30]

Now, the principle having been proven to a considerable extent, optical spectroscopy is maturing as a diagnostic modality. And, specific diagnostic spectroscopic tests are being proposed for more extensive testing in larger-scale clinical trials. There are four basic questions that need to be answered early in the process of developing a diagnostic test for large-scale clinical use, whether it be an *in vitro* clinical laboratory test for a blood analyte or a spectroscopic test for *in vivo* tissue diagnosis: Who should you study? What test parameters should you choose? Where should you set your diagnostic thresholds? And, how do you handle outliers and line sitters?

Table 1 Murphy's seven meanings for normal; adapted from Ref. 31.

1. Most probable (±2 SD)
2. Most representative of its class (average, median, modal)
3. Commonly encountered in its class (usual laboratory reference range)
4. Carrying no penalty (harmless)
5. Commonly aspired to (conventional)
6. Most suited to survival (optimal)
7. Most perfect of its class

2.1 Who Should You Study?

2.1.1 *How do you define normal?*

It seems fairly obvious that you need to study both patients with the disease in question and people free of that disease, the so-called normal controls. But, how should you define normal? Murphy[31] has identified at least seven different ways to define normal in the clinical setting (Table 1). Normal can be defined in a descriptive way as the most representative of its class, e.g., the average, mean, or mode. It can also be defined as the most probable result, often given as a range of ±2 standard deviations (SD) from the mean. Or, it may be defined as the most commonly encountered in its class, which corresponds to the usual laboratory reference range, determined by studying a large group of ''normal'' volunteers with no known disease. But, normal can also be defined in a more functional way, as that carrying no penalty (harmless), that commonly aspired to (conventional), that most suited to survival (optimal), or even the most perfect of its class (ideal).

Consider, for example, blood cholesterol, a significant risk factor for atherosclerotic cardiovascular disease (Table 2). The mean blood cholesterol varies for different populations depending upon age, sex, geographic, ethnic, and racial origin, and cultural factors such as diet and exercise. The mean blood cholesterol is lower for boys less than 12 years of age (yoa) (115 mg/dl) than for men 40–60 years of age (215 mg/dl) in the United States. It is also lower for men 40–60 years of age in Japan (195 mg/dl) than for men of comparable age in the United States. The current blood cholesterol recommendation of the American Heart Association (AHA) is <200 mg/dl. Yet, a typical reference range for blood cholesterol in a hospital clinical chemistry laboratory is 100–270 mg/dl. Which of these is normal?

Table 2 Blood total cholesterol.

Mean for boys <12 yoa in USA	115 mg/dl
Mean for men 40–60 yoa in USA	215 mg/dl
Mean for men 40–60 yoa in Japan	195 mg/dl
Current recommendation of AHA	<200 mg/dl
Laboratory reference range (UHOC)	100–270 mg/dl

Table 3 Diagnostic classification of coronary artery and aortic atherosclerosis.

1. Normal
2. Intimal fibroplasia
3. Atherosclerotic plaque
4. Atheromatous plaque
5. Fibrotic–sclerotic plaque
6. Calcified atherosclerotic plaque
7. Calcified atheromatous plaque
8. Calcified fibrotic–sclerotic plaque

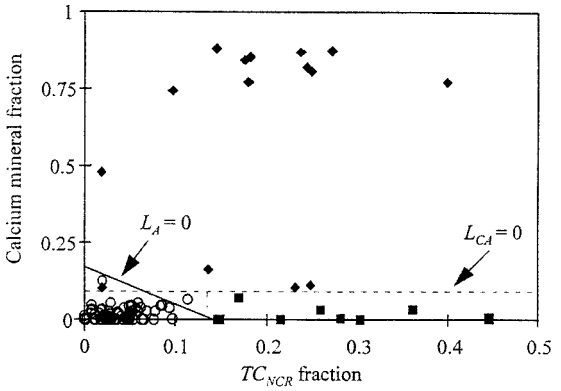

Fig. 1 NIR Raman spectroscopy diagnosis of coronary artery atherosclerosis, using relative weight fractions of total cholesterol (TC_{NCR}) and calcium minerals (● nonatherosclerotic, ■ noncalcified plaque, and ◆ calcified plaque).

Similar issues arise when trying to define normal for the purposes of tissue diagnosis. For example, a number of studies have been done exploring the potential of fluorescence and Raman spectroscopy for the *in vivo* diagnosis of atherosclerosis. In their studies of coronary artery and aortic atherosclerosis, a number of investigators,[9–14] have used a histologic diagnostic classification scheme based on that proposed in the Systematized Nomenclature of Medicine (SNOMED) (Ref. 32) that includes eight categories representing the progression from normal arteries to end-stage calcified plaques (Table 3).

But, atherosclerosis is a ubiquitous disease that affects the entire general population, beginning in infancy and progressing throughout adult life.[33] Therefore, truly normal arteries are seen only in infants and young children. Most arteries in even young adults show intimal fibroplasia, a thickening of the luminal intimal layer of the artery wall, which is one of the earliest manifestations of atherosclerosis but may also be seen in other types of arterial disease such as hypertension. So, as a practical matter, if the patient population available for study is comprised solely of adults, as it was in these studies, few if any normal arteries will be available for study, and it may not be possible to use a truly normal control group. In fact, in these studies, the control group was defined as nonatherosclerotic rather than normal, and included both histologically normal arteries and arteries with intimal fibroplasia.

This type of control group definition is appropriate for proof-of-principle studies of the diagnostic potential of fluorescence and Raman spectroscopy. Figure 1 shows data from the study of Romer et al.[13] in which the relative weight fractions of two biochemical components of atherosclerotic plaque, total (free and esterified) cholesterol, and calcium mineral salts, determined by Raman spectroscopy, were used as the basis of an algorithm for the diagnosis of atherosclerosis. Using this algorithm, nonatherosclerotic arteries (normal+intimal fibroplasia) could be distinguished from calcified and noncalcified plaque.

However, intimal fibroplasia worsens with age, and may result in a clinically significant stenosis of the involved artery, in the absence of other features of atherosclerosis seen in more advanced plaques. So, a relatively arbitrary limit must be set as to how much intimal fibroplasia is acceptable in the control group. And, eventually, as these spectroscopic techniques are more extensively tested in clinical practice, it will be necessary to not only distinguish histologically normal arteries from those with intimal fibroplasia, but to quantitate the intimal thickness or degree of luminal stenosis in arteries with intimal fibroplasia. Current studies are focusing on this and other diagnostic questions, and it may be that a combination of Raman spectroscopy with reflectance or some other type of spectroscopy may be able to provide this type of information.

Another problem in defining normal controls is frequently encountered in developing tests for tissue diagnosis, when comparing studies performed by different investigators. In many cases, normal will be defined differently by each investigator, making comparison difficult if not impossible. For example, Follen Mitchell et al.[23] recently performed a meta-analysis and found laser-induced fluorescence (LIF) spectroscopy comparable or superior to more conventional diagnostic techniques for the diagnosis of squamous intraepithelial lesions of the cervix (dysplasia), including colposcopy, Pap smear, cervicography, HPV testing, and speculoscopy, using receiver operating characteristic curves (Figure 2). In their review of the literature, they found negative or''normal'' results defined in at least five different ways by different investigators, including negative colposcopic biopsy, negative cone

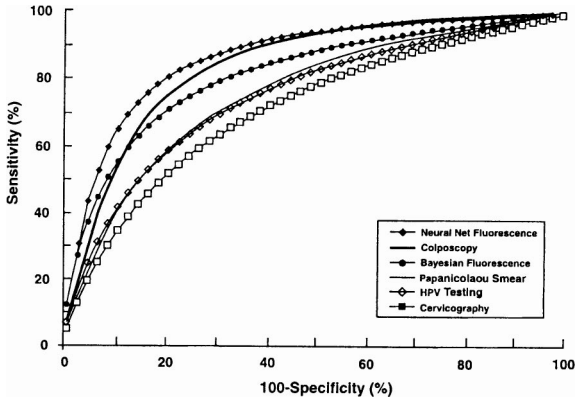

Fig. 2 Meta-analysis comparison of laser-induced fluorescence spectroscopy with conventional techniques for the diagnosis of cervical dysplasia, using receiver operator curves.

biopsy, negative Pap smear cytology, negative cervicographic findings, and negative colposcopic findings.

2.1.2 *How do you define diseased?*

The previous examples show that great care must be taken not only to define the normal control group for a specific study, but also to establish the definitions of normal used in other studies to which the study data will be compared. Great care must also be taken to define the patient population to be studied with the disease in question. The current generally accepted so-called ''gold standard'' for tissue diagnosis is microscopic examination of biopsy or surgical resection specimens. Ironically, the clinical areas of greatest need for new diagnostic modalities are often the areas where the gold standard of histologic diagnosis fails. This leaves the investigator with no reliable standard for comparison.

Why does the gold standard fail? Microscopic diagnosis, like spectroscopic diagnosis, is criteria driven or multiparametric. The diagnosis of most diseased or neoplastic tissues depends upon fulfillment of a number of microscopically defined criteria (or parameters). For a particular neoplasm, for example, these might include nuclear and cytoplasmic features of the neoplastic cells, the type of cell–cell junctions, the architectural arrangement of the cells, the relationship of the cells to surrounding normal tissue structures, etc. Fulfillment of each of these criteria may, in turn, require fulfillment of additional criteria. For example, for a particular neoplasm, the nuclear criteria may include size and shape of nuclei, size and number of nuceoli, chromatin pattern, etc.

This type of tissue diagnosis is by its very nature qualitative and subjective, even when the criteria are well established. Unfortunately, in some cases there is lack of consensus among pathologists as to the appropriate criteria.[34] In other cases, the criteria are poorly defined or difficult to recognize. And, finally, as Rambo said, ''pathologists are physicians *and* human beings,'' [35] and therefore, subject to human fallability. As a result, microscopic tissue diagnosis is subject to significant interobserver and intraobserver variability. This is a particularly difficult problem in the diagnosis and grading of dysplasia, a premalignant lesion seen in patients at high risk for development of carcinoma in a variety of clinical settings, including Barrett's esophagus,[36] inflammatory bowel disease,[37] adenomatous colon polyps,[38] cervical squamous intraepithelial lesions,[39] oropharyngeal cancer,[40] and superficial bladder cancer,[41] and the subject of intense spectroscopic investigation.

What can you do when the gold standard fails? A number of different approaches have been used to deal with the problem of interobserver variability in establishing the true tissue diagnosis as a base line for comparison with spectroscopic data. One is the ''superman'' approach of consulting an expert in the field, and using his or her expert diagnosis. In some ways this is the most commonly used, as a single observer has established the tissue diagnosis in the vast majority of published reports. Perhaps a more objective method is to employ the diagnosis of more than one pathologist, but then one has to have a strategy to deal with their differences of opinion.

One such strategy is to use the consensus of the entire group of pathologists as the diagnosis. A consensus diagnosis can be arrived at in several ways. One way is to have all of the pathologists perform the microscopic examination at the same time, usually at a multiheaded microscope, and agree on a single diagnosis. A second is to have each of the pathologists perform the microscopic examination independently, and to define the consensus diagnosis as that diagnosis rendered by the majority of pathologists. Using this approach, patients or specimens for which there is no consensus of a majority of pathologists may be either eliminated from analysis or assigned to a category of diagnosis unknown. A third way is to have each of the pathologists perform the microscopic examination independently, and then have a review by all of the pathologists together to reach a consensus diagnosis for those cases where there was a difference of opinion. This approach was employed by Ramanujam et al.[25] in developing a LIF spectroscopy technique for the diagnosis of cervical dysplasia.

Another strategy to deal with diagnostic differences of opinion is to use a scoring system to assign a numerical value to each possible diagnosis, and to use the arithmetic mean of the scores for each pathologist to establish the diagnosis. This allows data from all of the patients or specimens to be analyzed. Wallace et al. employed the latter two strategies in developing a light-scattering spectroscopy (LSS) technique for the diagnosis of dysplasia in Barrett's esophagus.[19] In their *in vivo* endoscopic biopsy study, they compared a diagnosis of dysplasia determined by LSS-based quantitative analysis of epithelial cell nuclear enlargement and crowding, two criteria used by pathologists in the microscopic diagnosis of dysplasia, with both the average diagnosis, using an adaptation of the scoring system of Riddel et al.[37] (Table 4), and the majority consensus diagnosis of four pathologists. Their data showed that, as expected, there was significant interobserver variation among the four pathologists, with Kappa statistics ranging from 0.31 to 0.37 (62%–66% agreement) for one-to-one comparisons of each pathologist with his or her colleagues. The LSS-based diagnosis faired better, with Kappa statistics of 0.57 and 0.63 (80% and 90% agreement) when compared to the average and consensus diagnoses, respectively (Table 5).

Table 4 Scoring system for the histologic diagnosis of dysplasia in Barrett's esophagus on endoscopic biopsy.

Scoring System
1 = NDB (nondysplastic Barretts)
2 = IND (indefinite for dysplasia)
3 = LGD (low-grade dysplasia)
4 = HGD (high-grade dysplasia)
5 = Invasive adenocarcinoma
Mean Score
NDB = 1 – 1.74
IND = 1.75 – 2.49
LGD = 2.5 – 3.24
HGD = > 3.25

Table 5 Interobserver variability in the microscopic and LSS diagnosis of dysplasia in Barrett's esophagus.

	Kappa	% Agreement
Pathologist 1 vs. colleagues	0.31	66%
Pathologist 2 vs. colleagues	0.22	62%
Pathologist 3 vs. colleagues	0.34	65%
Pathologist 4 vs. colleagues	0.37	65%
Spectroscopy vs. pathology, average diagnoses	0.57	80%
Spectroscopy vs. pathology consensus diagnoses	0.63	90%

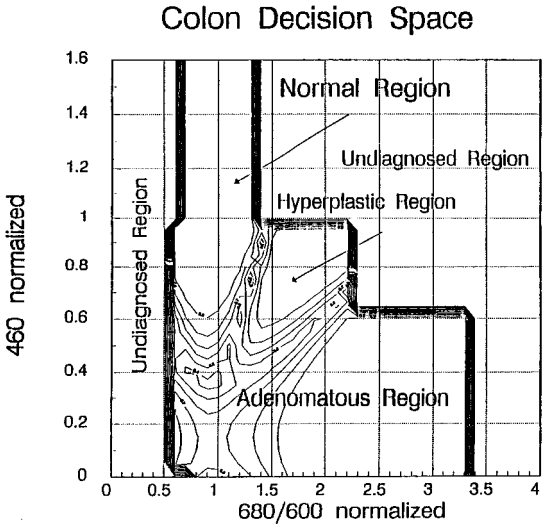

Fig. 3 Probabilistic diagnostic algorithm for the diagnosis of diminutive colon polyps using LIF spectroscopy.

Other approaches for coping with the flawed gold standard have also been tried. Morphometric analysis of biopsies or surgical specimens has been suggested as a more quantitative and objective approach to the assessment of microscopic criteria for tissue diagnosis,[42] but has not been widely adopted in clinical practice because it is time consuming and laborious. However, it has been used by some investigators as a basis for comparison with optical spectroscopy. In developing their LSS technique for the diagnosis of dysplasia, Perelman et al.[20] initially compared quantitative measures of nuclear size determined by morphometry with those determined by LSS in normal colon epithelial cell and T84 colon tumor cell culture monolayers. Ikeda et al. have also compared the diagnosis of premalignant changes in bronchial epithelium by fluorescence endoscopy with the nuclear features of endobronchial brushings assessed by morphometry.[15] Yet another approach is to compare the spectroscopic diagnosis to both the microscopic tissue diagnosis and another independent measure of disease or tumor involvement, such as the presence of a disease- or tumor-associated genetic abnormality. Studies of this type are in progress.

In the final analysis, whether an average, consensus, or expert microscopic diagnosis or other independent measure of disease or tumor involvement is used as the basis for comparison, the only thing that matters is whether or not the spectroscopic technique works in clinical practice. That is, whether or not it predicts the biologic end points of disease progression or response to therapy. So, ultimately, at some stage in the development of a spectroscopic (or any other) diagnostic test, studies with long-term patient follow up must be conducted. Since the field of optical spectroscopic tissue diagnosis is so young, few, if any, of this type of longitudinal study have been done as yet.

2.1.3 *Who else do you study?*

Most preliminary studies performed during the process of developing a new diagnostic test include, as we have discussed, a normal control group and a patient group with the disease in question. Many newly developed diagnostic tests, such as the spectroscopic techniques just discussed, show a clear distinction between groups in this type of small-scale twofold comparison of preselected populations. However, before testing these new techniques in larger-scale studies of unselected patient populations, it is important to study not only normal controls and the patient population of interest, but also patients with other diseases that may be clinically confused with the disease in question.

For example, Richards-Kortum et al. reported a LIF spectroscopic technique for the diagnosis of diminutive adenomatous colon polyps at endoscopy. In their initial *in vitro* study,[21] their best diagnostic algorithm had a sensitivity of 100% for adenomatous polyps (versus normal colonic mucosa). (The subject of sensitivity and specificity will be discussed in more detail later.) However, in the subsequent *in vivo* study of Cothren et al.,[22] a similar algorithm had a sensitivity of only 92% (Figure 3). One might assume that this apparent loss of diagnostic sensitivity was due to nontranslatability of *in vitro* data to *in vivo* studies. However, in this case, the apparent loss of diagnostic sensitivity was due, at least in part, to the fact that the patient population encountered in the second study included not only patients with adenomatous polyps, but patients with hyperplastic polyps, a commonly occurring diminutive colon polyp with intermediate spectroscopic features not included in the initial study. In fact, in the *in vivo* study, the sensitivity reported is for the diagnosis of adenoma versus nonadenoma (normal +hyperplastic polyp). In this case, hyperplastic polyps are just the tip of the iceberg, since there are many other less common types of colon polyps (juvenile polyps, retention polyps, hamartomatous polyps, inflammatory polyps, etc.) which may be encountered in patients undergoing endoscopic surveillance for adenomatous polyps. A similar situation exists for virtually every disease for which a spectroscopic diagnostic test is under development.

This is not to suggest that initial studies should include all possible confounding lesions or diseases. But rather, that it should be anticipated that the sensitivity and specificity of a diagnostic technique would fall when it is tested in large-scale studies in unselected or less selected patient populations. And that, in the later stages of test development, these confounding lesions or diseases will need to be identified and studied in systematic fashion.

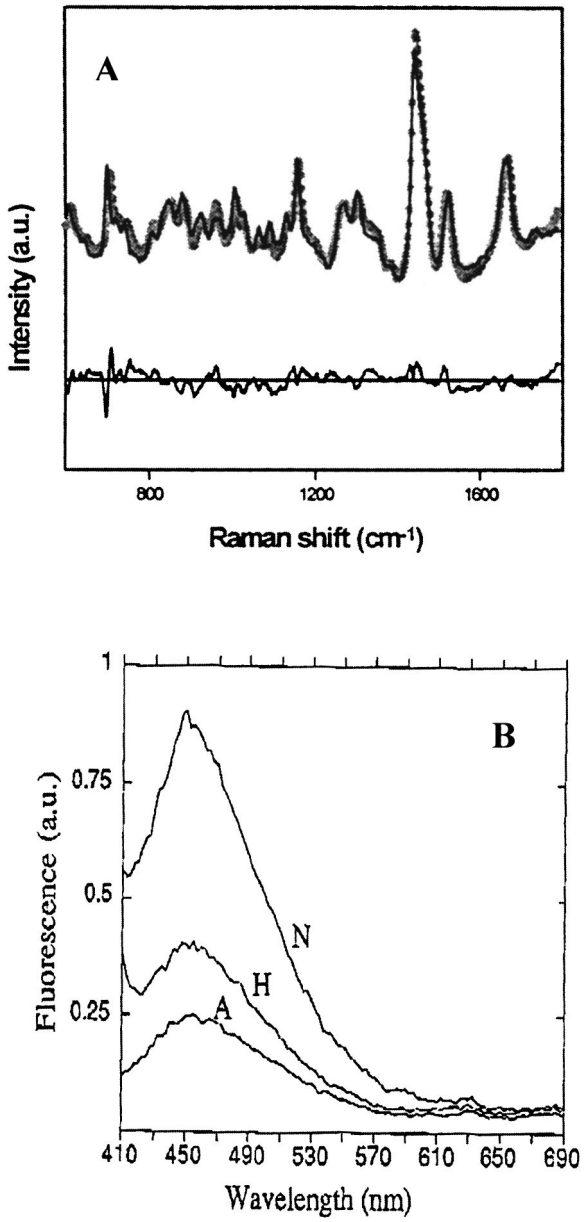

Fig. 4 Variability in the wealth of diagnostic parameters in different types of spectra: feast or famine.

2.2 What Diagnostic Parameters Should You Choose?

When spectroscopic techniques are considered, it seems that there is either a feast or famine in terms of the number of possible test parameters. LIF spectra, for example, are rather broad, smooth, and featureless on visual inspection, yielding a paucity of obvious diagnostic parameters. Raman spectra, on the other hand, can present an alarming complexity of sharp spectral features, any one of which might be a useful diagnostic parameter (Figure 4). Determining the optimal number of parameters can be as daunting a task as selecting the optimal type of parameter when faced with such a wealth of information.

Table 6 Probability of obtaining an abnormal diagnostic test result; adapted from Ref. 43.

No. of independent tests	Percentage of normals with abnormal result
1	5
2	10
4	19
6	26
10	40
20	64
50	92
90	99

A number of strategies can be used to identify and select diagnostic spectral parameters, including visual inspection of peak intensity or peak ratios, principal component, logistic or other statistical analyses, or a more empirical method using spectral features of known chemical or morphologic constituents of the diseased tissue. Whichever method is used, it is important to keep in mind that the probability of obtaining an abnormal result increases with the number of independent tests performed (or parameters measured), from 10% for two independent tests to 40% for ten independent tests (Table 6).[43] This is a well-known phenomenon in clinical medicine and has led to the discontinuance of the once common practice of routinely ordering large panels of (sometimes 20 or more) clinical chemistry tests on all patients admitted to the hospital. This practice predictably resulted in a high frequency of spurious abnormal tests, which in turn, resulted in costly and unnecessary follow-up laboratory testing.[44]

Fortunately, as with histologic diagnostic criteria, spectroscopic diagnostic parameters are not always independent. Nevertheless, even with dependent parameters, the likelihood of an abnormal result still increases, albeit less sharply, with increasing number of diagnostic parameters.

It may be that no specific criterion exits for determining the optimal number of independent tests to perform (or parameters to include) in a diagnostic algorithm. Perhaps the best approach is to test the performance of the algorithm for each possible combination of parameters under consideration. The commonly used measures of test performance are discussed in detail later.

Another interesting point regarding spectroscopic diagnostic parameters is the observation of Shafer[45] that those parameters that contribute most to the fit of spectroscopic data to a model may not be the parameters with the most diagnostic utility. Using principal component analysis to identify near-infrared (NIR) Raman spectral parameters for the diagnosis of benign and malignant breast lesions[26] (Figure 5), Shafer et al. found that principal component 8, which contributed only

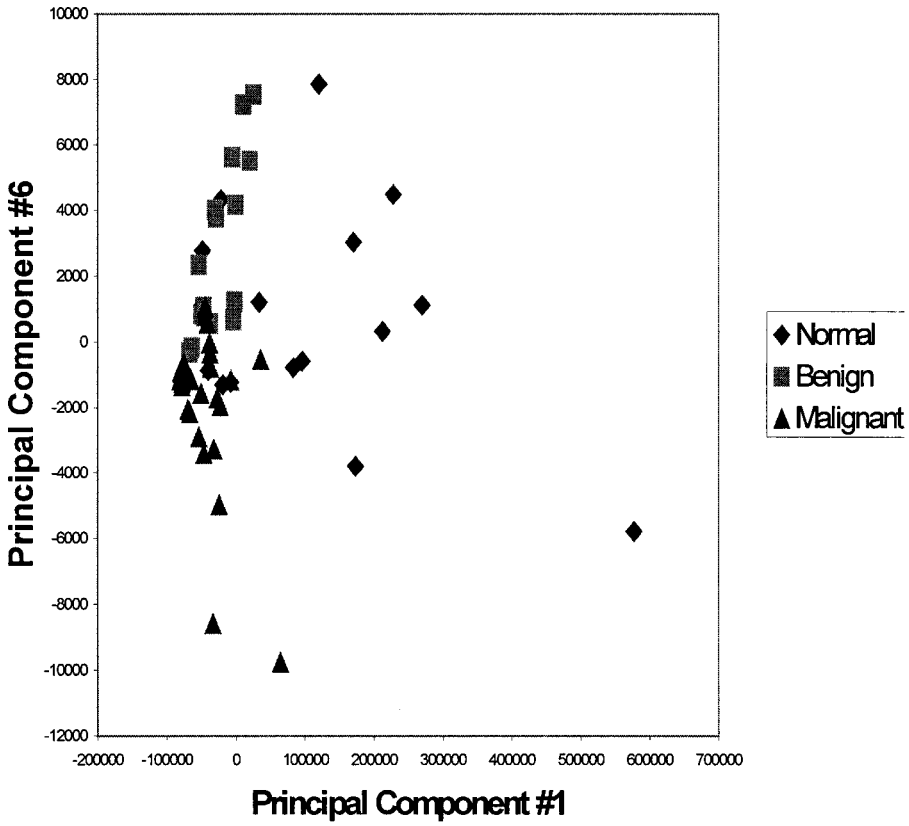

Fig. 5 Diagnosis of benign and malignant breast lesions using NIR Raman spectroscopy (♦ normal; ■ benign; and ▲ malignant).

0.06% to the total variance, contained more diagnostic information (*p*-value 0.1114) than principal component 5 (*p*-value 0.4079), which contributed 0.21% to the total variance (Table 7).

Table 7 Principal components (PC) in the NIR Raman spectra of benign and malignant breast lesions.

PC	% Total variance	*p* value*
1	96.77	
2	1.24	0.0001
3	0.84	0.0077
4	0.58	0.0000
5	0.21	0.4079
6	0.09	
7	0.08	0.9438
8	0.06	0.1114*
9	0.04	0.3478

* When added to the basic model of PC1 + PC6.

2.3 Where Should You Set Your Decision Thresholds?

In order to establish the diagnostic utility of any test, one has to evaluate some objective measure of its diagnostic performance. The most common measures used in clinical medicine are statistical and include sensitivity, specificity, predictive value, and test efficiency (Table 8). Ideally, one would like to develop a diagnostic test with 100% sensitivity, specificity, and predictive value. But, in the real world this is for all

Table 8 Statistical measures of diagnostic test performance. TP=true positive (No. of diseased patients correctly diagnosed); FP=false positive (No. of healthy patients misdiagnosed as diseased); TN=true negative (No. of healthy patients correctly diagnosed); and FN=false negative (No. of diseased patients misdiagnosed as healthy). Sensitivity=TP/(TP+FN)×100; specificity=TN/(TN+FP)×100; positive predictive value=TP/(TP+FP)×100; Negative predictive value=TN/(TN+FN)×100; and test efficiency=(TP+TN)/(TP+FP+TN+FN)×100.

	No. with positive test results	No. with negative test results	Total
No. with disease	TP	FN	TP+FN
No. without disease	FP	TN	FP+TN
Totals	TP+FP	TN+FN	TP+FP+TN+FN

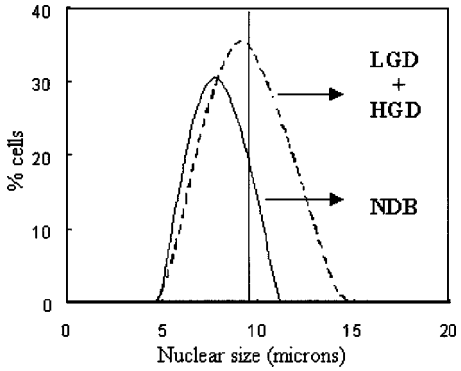

Fig. 6 Overlapping populations of cell nuclei in nondysplastic and dysplastic Barrett's epithelium studied by LSS (LGD=low-grade dysplasia; HGD=high-grade dysplasia; and NDB=nondysplastic Barretts).

practical purposes impossible. The reason is that, for most test parameters, you have overlapping populations, as observed by Backman[46] in the nuclear size distributions determined by LSS for dysplastic and nondysplastic Barrett's epithelium (Figure 6).

Fortunately, test parameters with substantial overlap can provide useful diagnostic information. Note that in the case of Figure 6, as will be discussed in greater detail later, you could conceivably achieve high sensitivity and specificity, even with the overlap shown, if you select the appropriate diagnostic parameter, for example, the percentage of nuclei larger than 10 μm.

Unfortunately, with overlapping populations, there is usually a tradeoff between sensitivity (or positive predictive value) and specificity (or negative predictive value). When the diagnostic threshold is changed, one goes up and the other goes down. So, the real question is where to set the diagnostic threshold. Should you optimize sensitivity, specificity, or predictive value? The answer is that it depends upon the clinical situation.

Consider, for example, a spectroscopic test for dysplasia to be used during endoscopic surveillance of patients with Barrett's esophagus. There are several ways that this type of test might be used clinically. It might be used to direct endoscopic biopsy to areas of increased likelihood of dysplasia, to be confirmed *in vitro* by conventional microscopy. Or, it might be used to make a real-time *in vivo* diagnosis of dysplasia in order to identify patients requiring more intensive endoscopic surveillance. Or, it might be used to make a real-time *in vivo* diagnosis of dysplasia in order to direct laser ablation therapy during the same endoscopic procedure. This spectroscopic diagnostic test may need to be optimized differently for use in each of these clinical situations. In fact, the definitions of positive and negative results themselves may need to differ in these different clinical situations.

2.3.1 *When do you want high sensitivity?*

The two statistical measures of diagnostic performance reported most often in the medical literature are sensitivity and specificity. Yet, in the majority of clinical situations, it is sensitivity, positive predictive value, and test efficiency that best reflect a diagnostic test's clinical utility.

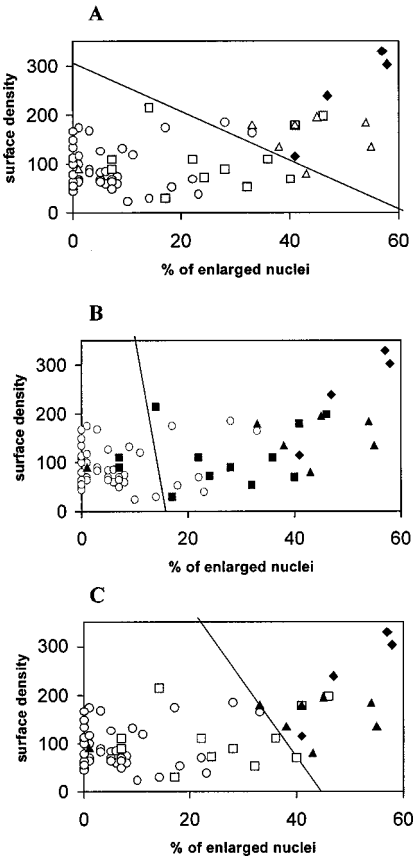

Fig. 7 Endoscopic diagnosis of dysplasia in Barrett's esophagus *in vivo,* using nuclear enlargement and crowding determined by light-scattering spectroscopy (○ nondysplastic Barrett's; ■ indefinite for dysplasia; ▲ low-grade dysplasia; and ◆ high-grade dysplasia). Note: hollow symbols indicate test results defined as negative and solid symbols test results defined as positive in each scenario.

Sensitivity is defined as the ratio of the number of true positive (TP) tests to the number of patients tested with the disease in question or TP/(TP+FN) (FN=false negative) (see also Table 8). When do you want high sensitivity? When the disease to be diagnosed is serious, should not be missed, and is treatable, and false positive results do not have serious adverse consequences for the patient. In this case, you want to identify every single patient with the disease for treatment or further clinical evaluation, even at the cost of misdiagnosing some healthy people as diseased.

Such would be the case for the LSS test for dysplasia in Barrett's esophagus of Wallace et al.,[19] were it to be used to direct endoscopic biopsies to be confirmed later by conventional microscopy, in order to, for example, identify patients with high-grade dysplasia requiring esophagectomy. In this case, the risk of a false positive spectroscopic diagnosis is small, since the diagnosis would be confirmed microscopically. So, the goal would be to identify every possible patient with high-grade dysplasia for biopsy, even if it meant biopsying some patients without high-grade dysplasia. In this case, only spectroscopic diagnoses of high-grade dysplasia would be defined as positive. Using this definition of positive and the decision threshold shown in Figure 7(A), the LSS test of Wallace et al. has a sensitivity, specificity, positive predictive

value, and test efficiency of 100%, 86%, 29%, and 89%. One might assume that a test with a positive predictive value of 29% has little diagnostic utility. But, in this case, it is the sensitivity of 100% that best reflects the test's clinical utility.

2.3.2 *When do you want high positive predictive value?*

Positive predictive value is defined as the ratio of the number of true positive tests to the total number of positive tests or TP/(TP+FP) (FP=false positive) (see also Table 8). When do you want high positive predictive value? When the disease to be diagnosed is serious, should not be missed, and is treatable, and false positive results *may* have serious adverse consequences for the patient. In this case, you want to be certain that every patient with a positive test has the disease in question, even at the risk of missing some diseased patients.

Such would be the case for the LSS test for dysplasia in Barrett's esophagus of Wallace et al., were it to be used make a real-time *in vivo* diagnosis of dysplasia in order to enroll the patient in a more intensive endoscopic surveillance program. In this case, a false positive diagnosis of dysplasia in a patient without dysplasia could subject the patient to additional unnecessary endoscopic surveillance. However, all patients at risk of dysplasia, even those with indefinite findings, should be enrolled in annual surveillance. In this case, spectroscopic diagnoses of high-grade dysplasia, low-grade dysplasia, and indefinite for dysplasia would be defined as positive. Using this definition of positive and the decision threshold shown in Figure 7(B), the LSS test of Wallace et al. has a sensitivity, specificity, positive predictive value, and test efficiency of 88%, 94%, 78%, and 75%. In this case, the sensitivity and specificity are lower, but the positive predictive value, the value that best reflects the test's utility in this clinical situation, is substantially higher.

2.3.3 *When do you want high test efficiency?*

Test efficiency is the least well known of the statistical measures of test performance. But, it is most often the best measure of the clinical utility of a diagnostic test. In fact, experience has shown that, given the choice of several different diagnostic tests for a specific disease, with no prior knowledge of the relative performance of the tests, clinicians will usually end up using the test with the highest test efficiency.

Test efficiency is defined as the ratio of the total number of correct test results to the total number of tests performed (TP+TN)/(TP+FP+TN+FN) (TN=true negative) (see also Table 8). When do you want high test efficiency? When the disease to be diagnosed is serious, should not be missed, and is treatable, and false positive and false negative results are equally serious or potentially injurious to the patient. In this case, you want to be certain that the test result is accurate whether it is positive or negative. This is most often the case in clinical practice.

And, such would be the case for the LSS test for dysplasia in Barrett's esophagus of Wallace et al., were it to be used to make a real-time *in vivo* diagnosis in order to direct laser ablation therapy of foci of dysplasia during the same endoscopic procedure. In this case, the risk of endoscopic laser ablation of a patient without dysplasia is roughly comparable to the risk of not treating a patient with dysplasia. Spectroscopic diagnoses of low- and high-grade dysplasia would be defined as positive and lead to laser ablation, but a diagnosis of indefinite for dysplasia would not. Using this definition of positive and the decision threshold shown in Figure 7(C), the LSS test of Wallace et al. has a sensitivity, specificity, and positive predictive value 92%, 98%, and 85%, respectively, and a test efficiency of 96%, the highest of the three scenarios.

2.3.4 *When do you want high specificity?*

As mentioned previously, specificity is one of the measures of diagnostic test performance most often reported, whereas, in fact, there are relatively few clinical situations in which specificity or negative predictive value are of utmost importance. Specificity is defined as the ratio of the number of true negative tests to the number of healthy individuals (individuals free of the disease in question) or TN/(TN+FP) (see also Table 8). When do you want high specificity? When the disease is serious but *not* treatable, knowledge that the disease is absent has psychological or public health value, and false positive results may have serious adverse consequences for the patient. In this case, you want to identify every single nonaffected or healthy individual, even at the cost of misidentifying some diseased patients as healthy. An example would be a test to diagnose an untreatable degenerative neurologic disorder such as Alzheimer's disease, where knowledge that the patient does not have the disease is reassuring but misdiagnosis of a patient with another treatable form of dementia may deny him or her appropriate medical treatment. Spectroscopic techniques for the diagnosis of Alzheimer's disease are currently under development.[28]

2.3.5 *When do you want high negative predictive value?*

Negative predictive value is defined as the ratio of the number of true negative tests to the total number of negative tests or TN/(TN+FN) (see also Table 8). When do you want high negative predictive value? When the disease is serious but *not* treatable, knowledge that the disease is absent has psychological or public health value, and false negative results will *not* have serious adverse consequences for the patient. This is the least common clinical reality. An example would be a test to identify individuals at risk of an untreatable inheritable disease such as Huntington's chorea by virtue of having an affected parent, for the purposes of genetic counseling. In this case, individuals who test positive could not be treated themselves, but might be counseled not to have a family in order to prevent passing on the disease to their children.

In the end, the decision as to which statistical measure of performance to optimize, for a specific diagnostic test, for a specific clinical use, must be made together with the appropriate clinicians with an understanding of the relative risks and benefits to the patient.

2.3.6 *Berkson's fallacy*

Statistical measures of test performance, such as sensitivity and positive predictive value, are not only influenced by where the diagnostic threshold is set and how positive and negative results are defined. They are also influenced by disease prevalence, the frequency of the disease in the study

Table 9 Disease prevalence and positive predictive value; adapted from Ref. 43.

Disease Prevalence (%)	Positive Predictive Value* (%)
0.1	9.0
1.0	50.0
2.0	66.9
5.0	83.9
50.0	99.0

[a] 99% sensitivity; 99% specificity.

population. As shown in Table 9, if sensitivity and specificity are held constant, the positive predictive value increases dramatically with disease prevalence, from 9% at a disease prevalence of 0.1% to 50% at a disease prevalence of 1.0% in the test population (at a sensitivity and specificity of 99%).[43] This is predicted by Bayes' theorem,[47] which can be expressed as the following equation:

$$\text{positive predictive value} = [(\text{prevalence})(\text{sensitivity})] / [(\text{prevalence})(\text{sensitivity}) + (1-\text{prevalence})(1-\text{specificity})].$$

One of the earliest pitfalls of diagnostic test development recognized is Berkson's fallacy, which deals with the effect of disease prevalence on diagnostic test performance. It states that ''the interplay of differential admission rates from an underlying population to the study population, can lead to the observation of a spurious association in the study group.'' [48] Simply stated, this means that unintentional bias in selecting patients for your study groups may lead to the conclusion that your test is a better (or worse) diagnostic tool than it really is.

One of the most common biases unknowingly introduced during the course of diagnostic test development is disease prevalence. Patients or specimens are often selected for small-scale proof-of-principle studies in a nonrandom fashion, so as to insure that a reasonable number of the patients or specimens studied are diseased. In fact, many investigators strive for roughly equal numbers of normal and diseased patients or specimens in these types of studies. This corresponds to a disease prevalence of 50%, which is extremely high compared to the prevalence of most diseases in the general population. The positive predictive value of a diagnostic test studied in this type of artificially high prevalence study population will undoubtedly fall when the test is studied in a larger-scale clinical trial where the disease prevalence in the study population is likely to be much lower. This is not to say that small-scale proof-of-principal studies of the type described are badly constructed. But rather, that the prevalence of the disease in the study population should be taken into account, when evaluating the predictive value of a diagnostic test in both initial small-scale and subsequent larger-scale studies.

Table 10 Pathologist's hedges.

''Consistent with ...''
''Most consistent with ...''
''More consistent with ... than ...''
''Favor ...''
''Suggestive, but not diagnostic of ...''
''Indefinite for ...''
''No definite evidence of ...''
''Cannot rule out ...''

Further, in order to optimize the performance of a diagnostic test, it may be necessary to adjust the decision threshold, based on the disease prevalence in the patient population studied clinically. Therefore, the same diagnostic test may need different decision thresholds when used as a diagnostic test in high-risk patients (where the disease prevalence in the test population is high) than when used as a screening test in the general population (where the disease prevalence is low).

In addition, the effects of institutional bias on disease prevalence must be taken into account. Disease prevalence often varies from institution to institution for the same patient population. This variation may be due to any of a number of factors, largely beyond the control of the investigator, that determine which patients from the general population receive their medical care in a particular institution. These include the type of medical institution (primary care versus specialty care, community versus university, for profit versus not for profit, etc.) and the community in which the institution is located (inner city, urban, suburban, rural, etc.). Therefore, the predictive value of a diagnostic test may vary from institution to institution in multi-institutional trials.

2.4 How Should You Handle Outliers and Line Sitters?

Once the decision threshold for a diagnostic test is established, and the test goes into clinical testing, there will inevitably be results that are outliers or line sitters. There are, once again, several strategies to deal with outliers and line sitters. One strategy is to render a definitive diagnosis in every case, no matter where the result lies on the decision surface. This is undoubtedly the worst choice, for as Voltaire once said, ''doubt is an unpleasant state of mind, but certainty is ridiculous.''

Pathologists have evolved another strategy over the years for dealing with the microscopic equivalent of outliers and line sitters, and that is the hedge, a descriptive phrase that modifies their subjective microscopic diagnoses and reflects their relative degree of certainty, such as ''consistent with ...,'' ''favor ...,'' or ''indefinite for ...'' (Table 10). Even the most objective and quantitative spectroscopic diagnostic test needs an equivalent to the hedge.

One statistical approach to the diagnostic hedge is to establish a region of diagnostic uncertainty on the decision surface, defined by a specific confidence limit above and below

the diagnostic threshold, determined for the specific test population. This result would lead to three diagnostic categories: positive, negative, and indeterminate. This method may deal relatively effectively with line sitters, but does not recognize outliers. It is also not easily translatable to different patient populations. And, how do you select the appropriate confidence limit?

Another approach is the probabilistic approach used by Cothren et al.[22] in developing their LIF technique for the diagnosis of diminutive colon polyps at endoscopy. They determined a series of probability contours for their diagnostic parameters that they then used to divide the decision surface into diagnostic regions for each normal and disease category based on its prior probability, a measure of disease prevalence, in the study population (Figure 3). Their probabilistic diagnostic algorithm provides not only a diagnosis for every data point, but also an objective measure of the degree of certainty of the diagnosis in that specific patient population. Unlike the confidence limit approach, the probabilistic approach deals effectively with both line sitters and outliers, since it defines decision surfaces where no diagnosis can be rendered as the probability that the diagnosis is correct is unacceptably low. It is also transferable to any patient population where the posterior probability of the disease is known or can be ascertained. Finally, it allows each clinician to determine for each individual patient in each clinical setting what the diagnostic confidence limit should be.

3 Conclusion

Optical spectroscopy is about to enter a new era of rigorous clinical testing and evaluation. As spectroscopic techniques for tissue diagnosis are tested in large-scale clinical trials rather than small-scale proof-of-principle studies, it is especially important that the basic principles and potential pitfalls of diagnostic test development be well understood. Although the hope and expectation may be that optical spectroscopy may one day supplant conventional microscopy, it is important not to forget the hard-earned lessons learned by pathologists as they blazed the trail of tissue diagnosis.

Acknowledgments

The author would like to acknowledge Michael Feld and other colleagues and collaborators at the George R. Harrison Spectroscopy Laboratory and Laser Biomedical Research Center at the Massachusetts Institute of Technology for their support in the preparation of this manuscript.

References

1. A. A. L. M. Velpeau, *Traite des Maladies du Sein et de la Region Mammaire*, Paris (1853). [Translated into English by M. Henry, *A Treatise on the Diseases of the Breast and Mammary Region*, printed for the Sydenham Society, London (1856)].
2. R. L. Virchow, *Disease, Life and Man; Selected Essays*, Stanford University Press, Stanford, CA (1958).
3. C. Ruge, ''Das Mikroskop in der Gynakologie und die Diagnostik,'' *Z. Geburtshilfe Gynakol* **20**, 178–205 (1890).
4. H. A. Azar, ''Arthur Purdy Stout (1885–1967): The man and the surgical pathologist,'' *Am. J. Surg. Pathol.* **8**, 301–307 (1984).
5. A. R. Moritz, ''Pathology and Legal Matters,'' in *Medicine in Cleveland and Cuyahoga County*: 12810-1976, K. L. Brown, Ed., pp. 188–230, The Academy of Medicine of Cleveland, Cleveland, OH (1977).
6. H. Humphreys, ''The surgeon-pathologists at P & S,'' *Prog. Surg. Pathol.* **1**, 1–3 (1980).
7. J. Rosai, *Guiding the Surgeon's Hand: The History of American Surgical Pathology*, Armed Forces Institute of Pathology, Washington, DC (1997).
8. F. Carson, *Histotechnology. A Self-Instructional Text*, American Society of Clinical Pathology Press, Chicago (1997).
9. R. Richards-Kortum, R. P. Rava, M. Fitzmaurice, L. Tong, N. B. Ratliff, J. R. Kramer, and M. S. Feld, ''A one layer model of laser-induced fluorescence of diagnosis of disease in human tissue: Applications to atherosclerosis,'' *IEEE Trans. Biomed. Eng.* **36**, 1222–1231 (1989).
10. J. J. Baraga, M. S. Feld, and R. P. Rava, ''*In situ* chemical analysis of biological tissue: Vibrational Raman spectroscopy of human atherosclerosis,'' *Proc. Natl. Acad. Sci. USA* **89**, 3473–3477 (1992).
11. R. Manoharan, J. J. Baraga, M. S. Feld, and R. P. Rava, ''Quantitative histochemical analysis of human artery using Raman spectroscopy,'' *J. Photochem. Photobiol., B* **16**, 211–233 (1992).
12. J. F. Brennan, T. J. Romer, R. S. Lees, A. M. Tercyak, J. R. Kramer, and M. S. Feld, ''Determination of human coronary artery composition by Raman spectroscopy,'' *Circulation* **96**, 99–105 (1997).
13. T. J. Romer, J. F. Brennan, M. Fitzmaurice, M. L. Feldstein, G. Deinum, J. L. Myles, J. R. Kramer, R. S. Lees, and M. S. Feld, ''Histopathology of human coronary atherosclerosis by quantifying its chemical composition with Raman spectroscopy,'' *Circulation* **97**, 878–885 (1998).
14. G. Deinum, D. Rodrigues, T. J. Romer, M. Fitzmaurice, J. R. Kramer, and M. S. Feld, ''Histological classification of Raman spectra of human coronary artery atherosclerosis using principal component analysis,'' *Appl. Spectrosc.* **53**, 938–942 (1992).
15. N. Iketa, C. MacAulay, S. Lam, J. Le Riche, P. Payne, D. Garner, C. Konaka, H. Kato, and B. Palcic, ''Malignancy associated changes in bronchial epithelial cells and clinical application as a biomarker,'' *Lung Cancer* **19**, 161–164 (1998).
16. S. Lam, T. Kennedy, M. Unger, Y. E. Miller, D. Gelmont, V. Rusch, B. Gipe, D. Howard, J. C. LeRiche, A. Coldman, and A. F. Gazdar, ''Localization of bronchial intraepithelial neoplastic lesions by fluorescence,'' *Chest* **113**, 696–702 (1998).
17. S. Lam and H. Shibuya, ''Early diagnosis of lung cancer,'' *Clin. Chest Med.* **20**, 53–61 (1999).
18. J. K. Dhingra, D. F. Perrault, Jr., K. McMillan, E. E. Rebeiz, S. Kabani, R. Manoharan, I. Itzkan, M. S. Feld, and S. M. Shapshay, ''Early diagnosis of upper aerodigestive tract cancer by autofluorescence,'' *Arch. Otolaryngol.* **122**, 1181–1186 (1996).
19. M. D. Wallace, L. T. Perelman, V. Backman, J. M. Crawford, M. Fitzmaurice, M. Seiler, K. Badizadegan, S. J. Shield, I. Itzkan, R. R. Dasari, J. Van Dam, and M. S. Feld, ''Endoscopic detection of dysplasia in patients with Barrett's esophagus using light scattering spectroscopy,'' *Gastroenterology* (submitted).
20. L. T. Perelman, V. Backman, M. Wallace, G. Zonios, R. Manoharan, A. Nusrat, S. Shields, M. Seiler, C. Lima, T. Hamano, I. Itzkan, J. Van Dam, J. M. Crawford, and M. S. Feld, ''Observation of periodic fine structure in reflectance from biological tissue: A new technique for measuring nuclear size distribution,'' *Phys. Rev. Lett.* **80**, 627–630 (1998).
21. R. Richards-Kortum, R. P. Rava, R. E. Petras, M. Fitzmaurice, M. Sivak, and M. S. Feld, ''Spectroscopic diagnosis of colonic dysplasia,'' *Photochem. Photobiol.* **53**, 777–786 (1991).
22. R. M. Cothren, M. R. Sivak, Jr., J. Van Dam, R. E. Petras, M. Fitzmaurice, J. M. Crawford, J. Wu, J. Brennan, R. Rava, R. Manoharan, and M. S. Feld, ''Detection of dysplasia at colonoscopy using laser-induced fluorescence: A blinded study,'' *Gastrointestinal Endosc.* **44**, 168–176 (1996).
23. M. Follen Mitchell, S. B. Cantor, N. Ramanujam, G. Tortolero-Lunai, and R. Richards-Kortum, ''Fluorescence spectroscopy for diagnosis of squamous intraepithelial lesions of the cervix,'' *Obstetrics Gynecol. (N.Y.)* **93**, 462–470 (1999).
24. N. Ramanujam, M. Follen Mitchell, A. Mahadevan, S. Thomsen, A. Malpica, T. Wright, N. Atkinson, and R. Richards-Kortum, ''Development of a multivariate statistical algorithm to analyze human cervical tissue fluorescence spectra acquired *in vivo*,'' *Lasers Surg. Med.* **19**, 46–62 (1996).
25. N. Ramanujam, M. Follen Mitchell, A. Mahadevan, S. Thomsen, A. Malpica, T. Wright, N. Atkinson, and R. Richards-Kortum, ''Spectroscopic diagnosis of cervical intraepithelial neoplasia (CIN) *in vivo* using laser-induced fluorescence at multiple excitation wavelengths,'' *Lasers Surg. Med.* **19**, 63–74 (1996).
26. R. Manoharan, K. Shafer, L. Perelman, J. Wu, K. Chen, G. Deinum,

M. Fitzmaurice, J. Myles, J. Crowe, R. R. Dasari, and M. S. Feld, ''Raman spectroscopy and fluorescence photon migration for breast cancer diagnosis and imaging,'' *Photochem. Photobiol.* **67**, 15–22 (1998).
27. B. J. Tromberg, O. Coquoz, J. B. Fishkin, T. Pham, E. R. Anderson, J. Butler, M. Cahn, J. D. Gross, V. Venugopalan, and D. Pham, ''Non-invasive measurements of breast tissue optical properties using frequency-domain photon migration,'' *Philos. Trans. R. Soc. Lond B Biol. Sci.* **352**, 661–668 (1997).
28. E. B. Hanlon, I. Itzkan, R. R. Dasari, M. S. Feld, R. J. Ferrante, A. C. McKee, D. Lathi, and N. W. Kowal, ''Near-infrared fluorescence spectroscopy detects Alzheimer's Disease *in vitro*,'' *Photochem. Photobiol.* **70**, 236–242 (1999).
29. I. J. Bigio and J. R. Mourant, ''Ultraviolet and visible spectroscopies for tissue diagnostics: Fluorescence and elastic-scattering spectroscopy,'' *Phys. Med. Biol.* **42**, 803–814 (1997).
30. E. B. Hanlon, R. Manoharan, T. W. Koo, K. E. Shafer, J. T. Motz, M. Fitzmaurice, J. R. Kramer, I. Itzkan, R. R. Dasari, and M. S. Feld, ''Prospects for *in vivo* Raman spectroscopy,'' *Phys. Med. Biol.* **45**, R1–R59 (2000).
31. E. A. Murphy, ''The normal, and the perils of the sylleptic argument,'' *Perspect. Biol. Med.* **15**, 566–582 (1972).
32. *Systematized Nomenclature of Medicine: Microglossary for Surgical Pathology*, D. J. Rothwell Ed., College of American Pathologists, Skokie, IL (1980).
33. W. E. Stehbens, ''The pathogenesis of atherosclerosis: A critical evaluation of the evidence,'' *Cardiovasc. Pathol.* **6**, 123–153 (1997).
34. R. J. Schlemper, M. Itabashi, Y. Kato, K. J. Lewin, R. H. Riddell, T. Shimoda, P. Sipponen, M. Stolte, and H. Watanabe, ''Differences in the diagnostic criteria used by Japanese and Western pathologists to diagnose colorectal carcinoma,'' *Cancer (N.Y.)* **82**, 60–69 (1998).
35. O. N. Rambo, ''The limitations of histologic diagnosis,'' *Prog. Rad. Ther.* **2**, 215–224 (1962).
36. M. Alikhan, D. Rex, A. Khan, E. Rahmani, O. Cummings, and T. M. Ulbright, ''Variable pathologic interpretation of columnar lined esophagus by general pathologists in community practice,'' *Gastrointestinal Endosc.* **50**, 23–26 (1999).
37. R. H. Riddell, H. Goldman, D. E. Ransohoff, H. D. Appelman, C. M. Fenoglio, R. C. Haggitt, C. Ahren, P. Correa, S. R. Hamilton, B. C. Morson, S. C. Sommers, and J. H. Yardley, ''Dysplasia in inflammatory bowel disease: Standardization classification with provisional clinical implications,'' *Hum. Pathol.* **14**, 931–968 (1983).
38. C. Fenger, M. Bak, O. Kronorg, and H. Svanholm, ''Observer reproducibility in grading dysplasia in colorectal adenomas: Comparison between two different grading systems,'' *J. Clin. Pathol.* **43**, 320–324 (1990).
39. W. G. McCluggage, M. Y. Walsh, C. M. Thorton, P. W. Hamilton, A. Date, L. M. Caughley, and H. Bharucha, ''Inter and intra-observer variation in the histopathological reporting of cervical squamous intraepithelial lesions using a modified Bethesda grading system,'' *Br. J. Obstet. Gynecol.* **105**, 206–210 (1998).
40. L. M. Abbey, G. E. Kaugars, J. C. Gunsolley, J. C. Burns, D. G. Page, J. A. Svirsky, E. Eisenberg, D. J. Krutchkoff, and M. Cushing, ''Intraexaminer and interexaminer reliability in the diagnosis of oral epithelial dysplasia,'' *Oral Surg. Oral Med. Oral Pathol. Oral Radiol. Endod.* **80**, 188–191 (1995).
41. B. Richards, M. K. Parmar, C. K. Anderson, I. D. Ansell, K. Grigor, R. R. Hall, A. R. Morley, F. K. Mostofi, R. A. Risdon, and B. M. Uscinska, ''Interpretation of biopsies of 'normal' urothelium in patients with superficial bladder cancer. MRC Superficial Bladder Cancer Sub Group,'' *Br. J. Urol.* **67**, 369–375 (1991).
42. P. N. Furross, ''The use of digital images in pathology,'' *J. Pathol.* **183**, 253–263 (1997).
43. R. S. Galen and S. R. Gambino, *Beyond Normality: The Predictive Value and Efficiency of Medical Diagnoses*, Wiley and Sons, New York (1975).
44. N. Namais, M. G. McKenney, and L. C. Martin, ''Utility of diagnostic admission chemistry and coagulation profiles in trauma patients: A reappraisal of traditional practice,'' *J. Trauma: Inj., Infect., Crit., Care* **41**, 21–25 (1996).
45. K. Shafer (personal communication).
46. V. Backman (personal communication).
47. T. Bayes, ''An essay toward solving a problem in the doctrine of chance,'' *Philos. Trans. R. Soc. Lond B Biol. Sci.* **53**, 370–418 (1763).
48. J. Berkson, ''Limitations of the application of fourfold table analysis to hospital data,'' *Biometrics* **2**, 47 (1946).