Performance Estimation of Diagnostic Tests for Cervical Precancer Based on Fluorescence Spectroscopy: Effects of Tissue Type, Sample Size, Population, and Signal-to-Noise Ratio

Urs Utzinger, E. Vanessa Trujillo, E. Neely Atkinson, Michele F. Mitchell, Scott B. Cantor, and Rebecca Richards-Kortum*

***Abstract*— Fluorescence spectroscopy may provide a cost-effective tool to improve precancer detection. We describe a method to estimate the diagnostic performance of classifiers based on optical spectra, and to explore the sensitivity of these estimations to factors affecting spectrometer cost. Fluorescence spectra were obtained at three excitation wavelengths in 92 patients with an abnormal Papanicolaou smear and 51 patients with no history of an abnormal smear. Bayesian classification rules were developed and evaluated at multiple misclassification costs. We explored the sensitivity of classifier performance to variations in tissue type, sample size, tested population, signal to noise ratio (SNR), and number of excitation and emission wavelengths. Sensitivity and specificity could be evaluated within ±7%. Minimal decrease in diagnostic performance is observed as SNR is reduced to 15, the number of excitation-emission wavelength combinations is reduced to 15 or the number of excitation wavelengths is reduced to one. Diagnostic performance is compromised when ultraviolet excitation is not included. Significant spectrometer cost reduction is possible without compromising diagnostic ability. Decision-analytic methods can be used to rate designs based on incremental cost-effectiveness.**

***Index Terms*—Cervix, cost-effectiveness analysis, diagnosis, fluorescence spectroscopy, precancer, signal-to-noise ratio (SNR).**

I. Introduction

NUMEROUS studies have demonstrated that techniques based on fluorescence spectroscopy have the potential to improve the detection of epithelial precancerous lesions in a variety of organ sites (for a recent review, see [1]–[3]). Optical diagnosis can be carried out automatically in real time, potentially reducing the need for clinical expertise, biopsies and follow-up visits. As medical costs continue to rise in the United States, opportunities to develop technologies which allow more efficient and less costly delivery of health care are of particular importance [4]. In developing cost-effective technologies, it is useful to explicitly consider the tradeoff between economic cost and diagnostic performance within a decision-analytic model [5]. Most of the literature describing new optical diagnostic methodologies have compared their performance to that of a gold standard (such as biopsy) and the standard of care. Although the potential economic impact of a few optical techniques have been examined [6], [7], these articles do not address how to incorporate cost-effectiveness as a design goal in the technology development phase. This paper describes a method to characterize the diagnostic performance of optical technologies which can be used in conjunction with economic models to develop cost-effective clinical systems.

The accuracy of a diagnostic test is frequently characterized using the metrics of correct classification rate, predictive value, sensitivity, and specificity [8]. Unlike correct classification rate and predictive value, specificity and sensitivity are theoretically independent of disease prevalence in the tested population [8]. Thus, sensitivity and specificity provide an important method to compare the performance of two diagnostic methods tested in different groups of patients. However, comparing the performance of two tests based on a single value of sensitivity and specificity reported for each can be misleading, because these quantities vary as the thresholds for a positive diagnosis are raised or lowered [9]. The diagnostic performance of a test can be fully characterized by reporting sensitivity versus (1-specificity) as the threshold is varied; the resulting curve is known as a receiver operator characteristic (ROC) curve [9]. Estimating the ROC curve of a new diagnostic method relative to an established standard of care provides a method of judging whether a new technology adds diagnostic value [10], [11]. To develop new technologies which are most cost-effective, a method of estimating the associated ROC curve as a function of the economic cost of the technology is required. This paper describes a method to estimate the diagnostic performance of classifiers based on

Manuscript received August 21, 1998; revised April 29, 1999. This work was supported by the Whitaker Foundation under a Cost-Reducing Health Care Technology Grant and by the National Institutes of Health (NIH) under Grant CA72650. *Asterisk indicates corresponding author.*

U. Utzinger and E. V. Trujillo are with the Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, TX 78712 USA.

E. N. Atkinson is with the Department of Biomathematics, University of Texas M. D. Anderson Cancer Center, Houston, TX 77030 USA.

M. F. Mitchell is with the Department of Gynecologic Oncology, University of Texas M. D. Anderson Cancer Center, Houston, TX 77030 USA.

S. B. Cantor is with the Department of Medical Specialties, University of Texas M. D. Anderson Cancer Center, Houston, TX 77030 USA.

*R. Richards-Kortum is with the Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, TX 78712 USA (e-mail: kortum@mail.utexas.edu).

Publisher Item Identifier S 0018-9294(99)07638-7.

optical spectra as a function of the economic cost of the spectrometer. The method is illustrated by estimating ROC curves for fluorescence detection of cervical precancer.

Despite the availability of Papanicolaou smear screening, cervical cancer and its precursors are important and costly health problems. Worldwide, cervical cancer is the second most common malignancy in women, and approximately 475 000 women are diagnosed each year with invasive cervical cancer [12]. In the United States, Kurman estimates that over $6 billion are spent annually in the evaluation and treatment of low grade precursor lesions [13]. Optical technologies have the potential to improve both the screening and detection of cervical cancer and its precursors; however, it is important to demonstrate the new health care technologies are both accurate and cost-effective.

We previously developed an algorithm for the diagnosis of low grade and high-grade cervical precancers based on fluorescence spectra at 337-, 380-, and 460-nm excitation using Bayesian classifiers [14]. Data were divided into a training set, used to develop the classification rules, and a validation set, used to test the rules. Sensitivity and specificity were calculated at a single threshold where the minimum number of samples were misclassified. Similar sensitivity and specificity were obtained when the classifier was applied to both the training and validation sets, and the technique demonstrated a similar sensitivity and improved specificity relative to colposcopy in expert hands [14]. In a subsequent economic analysis, we showed that this system had the potential to both increase the number of cases of high-grade cervical precancer identified as well as to reduce health care costs, with the potential for annual cost savings of over $600 million in the United States [6]. The magnitude of the projected savings was sensitive to the economic cost of the spectroscopy system used to measure data. In this paper we present a method to estimate the diagnostic performance of Bayesian classifiers developed from optical spectra, and to explore the sensitivity of these estimations to variations in tissue type, sample size, tested population and economic cost of the spectrometer. Decision-analytic methods [6] can then be used to rate various designs based on their incremental cost effectiveness.

II. Methods

1) Clinical System: The portable fluorimeter which was used to acquire cervical tissue fluorescence has been described in detail previously [14], and is briefly reviewed here. Two nitrogen pumped-dye lasers were used to provide illumination via a fiberoptic probe at three different excitation wavelengths: 337, 380, and 460 nm. The average transmitted pulse energies at 337-, 380-, and 460-nm excitation wavelengths were 12, 9, and 14 μJ, respectively. The laser characteristics were a 5–ns pulse duration and a repetition rate of 30 Hz. The proximal ends of the probe's emission collection fibers were imaged at the entrance slit of a polychromator coupled to an intensified diode array controlled by a multichannel analyzer. Using this system, fluorescence spectra were acquired from the cervix: ten spectra for ten consecutive pulses were acquired at 337-nm excitation; next, 50 spectra for 50 consecutive laser pulses were measured at 380-nm excitation and then an additional 50 spectra at 460-nm excitation. All spectra were corrected for the nonuniform spectral response of the detection system using correction factors obtained by recording the spectrum of an N.I.S.T. traceable calibrated tungsten ribbon filament lamp. Raw data were preprocessed and expressed in units relative to the peak fluorescence intensity of a Rhodamine 610 calibration standard. Spectra were recorded every 5 nm resulting in a total of 160 measured intensities per measurement site. The detection system was shot noise dominated.

2) Clinical Data: Clinical data were acquired in two settings: a referral setting, in which spectra were measured from a group of patients referred for colposcopy on the basis of an abnormal Papanicolaou smear (Study 1), and in a screening setting, in which spectra were measured from a group of patients with no history of an abnormal smear (Study 2). The details of each study have been previously reported [14], [15], but are briefly reviewed here.

Study 1: Clinical fluorescence spectra were measured from 374 cervical sites in a group of 92 nonpregnant patients referred to the colposcopy clinic of the University of Texas MD Anderson Cancer Center on the basis of abnormal cervical cytology. Informed consent was obtained from each patient who participated and the study was reviewed and approved by the Institutional Review Boards of the University of Texas, Austin and the University of Texas M. D. Anderson Cancer Center. After colposcopic examination of the cervix, but before tissue biopsy, fluorescence spectra were typically acquired from two colposcopically abnormal sites, two colposcopically normal squamous sites and one normal columnar site (if colposcopically visible) from each patient. The colposcopic examination includes the application of acetic acid to the cervix for approximately 2 min. Tissue biopsies were obtained only from abnormal sites identified by colposcopy and subsequently analyzed by the probe to comply with routine patient care procedures. All tissue biopsies were submitted for histologic examination by a panel of four board-certified pathologists and a consensus diagnosis was established using the Bethesda classification system. Samples were classified as normal squamous (SN) (188 sites), normal columnar (CN) (26 sites), metaplasia (20), inflammation (25), low grade squamous intraepithelial lesion (LG SIL) (45 sites), or high-grade squamous intraepithelial lesion (HG SIL) (70 sites).

Study 2: Clinical fluorescence spectra were measured from a group of 55 nonpregnant women with no history of an abnormal Papanicolaou smear. Informed consent was obtained from each patient who participated and the study was reviewed and approved by the Institutional Review Boards of the University of Texas, Austin and the University of Texas M. D. Anderson Cancer Center. Prior to spectroscopic measurements, Papanicolaou smears were collected with an endocervical brush and an Ayre's spatula. After the Papanicolaou smear was performed, acetic acid was applied to the cervix. Emission spectra were collected under colposcopic guidance from an average of three sites per patient. In the 51 women with a normal Papanicolaou smear, spectra were obtained from 103 squamous normal sites and 23 columnar normal sites.

TABLE I
OVERVIEW OF PARAMETERS VARIED DURING ALGORITHM DEVELOPMENT

Parameters studied	Excitation Wavelengths a=337, b=380, c=460 nm	Emission Wavelengths	Studies Included	Validation Technique	Algorithms Examined	S/N Ratio	Results
Size of Training Set	All	All	Study 1	1/2 training:1/2 validation 7/8 training:1/8 validation Jackknife	SN vs. SIL CN vs. SIL LGSIL vs. HGSIL	Maximum	Figure 1, Figure 2, Figure 3
Disease prevalence	All	All	Study 1 Study 1 and 2	1/2 training:1/2 validation	SN vs. SIL CN vs. SIL	Maximum	Figure 4
Signal to Noise Ratio	All	All	Study 1	7/8 training:1/8 validation	Normal vs. SIL	Maximum, 50, 25, 15, 10, 5, 2	Figure 6
Reduced Excitation Wavelengths	a, b, c, a&b, a&c, b&c, a&b&c	All	Study 1	7/8 training:1/8 validation	Normal vs. SIL Normal/LGSIL vs. HGSIL	Maximum, 25	Figure 7
Reduced Emission Wavelengths	a, b, c, a&b, a&c, b&c, a&b&c	Reduced	Study 1	7/8 training:1/8 validation	Normal vs. SIL Normal/LGSIL vs. HGSIL	Maximum, 25	Figure 8

3) Algorithm Development: The Bayesian algorithm previously described classifies cervical tissue using three rules: the first classifies samples as squamous normal (SN) or not based on normalized fluorescence spectra, the second classifies the remaining samples as columnar normals (CN) or not based on normalized, mean-scaled fluorescence spectra, and the third classifies the remaining samples as LG or HG SIL, based on normalized fluorescence spectra [14]. We developed each classification rule and evaluated its performance using separate training and validation sets using the method described in [14] with three important differences. First, data were divided into the training and validation sets with the constraint that spectra from all sites measured in a particular patient must be placed as a group in either the training or validation set. Previously, individual spectra were randomly assigned to either the training or validation set, so that not all spectra from a given patient were found in the same data set. Second, the dimension reduction using principal component analysis, which represents the first step of the algorithm development was previously carried out using both the training and validation sets, and the remaining steps were carried out using only the validation set. In this work, all steps, including dimension reduction, were carried out using only the training set. Finally, in order to study the influence of various parameters on the algorithm performance, its development was fully automated. Histograms of principal component (PC) distributions were fitted to gamma and normal distributions and results with least error were chosen. A subset of diagnostically relevant PC's were selected; the criteria to retain a PC was based on the statistical significance of the difference in means of the two tissue classes for that particular PC, in addition to the variance it accounted for. Only the first ten PC's, starting with the one that accounted for the most variance and going in descending order, were considered in the algorithm; PC's were included in the algorithm only if the difference in the means of the PC for the two tissue classes was statistically different below the $p = 0.10$ level.

4) Analysis of Algorithm Performance: The performance of each classification rule was assessed by calculating sensitivity and specificity as the cost of misclassification was varied [9]. An ROC curve for each data set was estimated from these data points using the method of Littenberg and Moses [16]. We explored the sensitivity of the ROC curve to variations in the tissue type, the size of the training set, the disease prevalence, the SNR of the data and the number of excitation and emission wavelengths for which data were recorded. Table I provides a summary of these studies, which are described in detail below.

The effect of the training set sample size on both the training and validation set ROC curves was explored for each classification rule. First, training sets were used which contained data from approximately half the patients, with the remaining patients assigned to the validation set. To make more efficient use of the available data, the technique of cross validation was also explored to increase the size of the training set [17]. Training and validation set ROC curves were estimated using cross validation under two alternatives: in the first, data from approximately one-eighth of the patients were held out at a time, in the second, data from a single patient were held out a time.

The effect of reducing the disease prevalence on the ROC curve estimate was explored by combining data from Studies 1 and 2. The classification rules for discrimination of SN and SIL tissues and CN and SIL tissues were redeveloped using the combined data. In this case, the training set contained data from approximately half of the Study 1 and half of the Study 2 patients and the validation set contained data from the remaining patients. Training and validation set ROC curves were estimated in the manner described above.

The three classification rules developed using cross validation were then used to estimate the performance of two composite algorithms, one to separate all normal samples from all SIL's and one to separate all normal and LG SIL samples from HG SIL's. Composite algorithm performance was estimated using either data from Study 1 only or the combined data from Studies 1 and 2. The ROC curves of the composite algorithm which separates normal samples from SIL's were compared to that of colposcopy in the referral setting, the standard of care for evaluation of an abnormal Papanicolaou smear [10].

Fluorescence spectra were collected using an expensive optical system consisting of laser excitation sources and an optical multichannel analyzer. In both studies, data had a high signal-to-noise ratio (SNR). The average SNR of these spectra ranged from 300 at 337-nm excitation to 600 at 460-nm excitation. We have identified a number of less expensive spectrometer designs which could be used to obtain tissue spectra. In order to select the one which is most cost effective, we must first predict the algorithm performance (in the form of an ROC curve) with each. In the alternative designs we wish to evaluate, spectrometer cost is lowered by reducing the target SNR or by reducing the number of emission or excitation wavelengths at which data are collected. In particular, reducing the number of measured emission wavelengths from the 160 in the original data set to only 10–15 can have a significant effect. Bandpass filters can be used instead of a spectrograph. Similarly, eliminating the need for UV excitation wavelengths (337 nm) can significantly reduce cost. The methodology we have developed to estimate algorithm performance permits exploration of these alternatives, without the need to collect additional data.

We first explored the effect of a reduction in the SNR of the fluorescence spectra by adding increasing amounts of random, Poisson noise to data from Study 1 in the validation set. The original spectra were scaled so that the maximum intensity was equivalent to the square of the target SNR. Then a new spectrum, with the desired SNR, was generated by feeding each intensity value individually into a Poisson random number generator (Statistical Toolbox, Matlab, Mathworks Inc.), whose mean was equal to the intensity of the scaled spectrum. The ROC curves for the SN versus SIL classification rule were estimated using the manner described previously; the training set contained data from half of the patients with the original SNR preserved, while the SNR of spectra in the validation set were varied.

The effect of reducing the number of excitation wavelengths was studied at two SNR's: the original and a SNR of 25. The two composite algorithms were redeveloped using the data from Study 1 for the seven possible combinations of one, two and all three excitation wavelengths. The area under the ROC curve (AUC) for the validation set was calculated at the original SNR and the reduced SNR. To explore the effect of simultaneously reducing the number of excitation and emission wavelengths, this analysis was repeated using a reduced set of emission wavelengths (Table II), selected because they were highly correlated to the PC's which were most diagnostically useful [14]. The mean emission intensity at the filter center wavelength was calculated, averaging over the bandpass width of the filter.

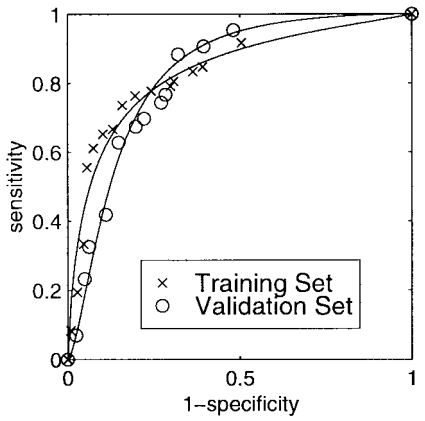
TABLE II
SET OF REDUCED EMISSION EXCITATION-EMISSION WAVELENGTH PAIRS AND THEIR CORRESPONDING FILTER BANDWIDTHS

Excitation Wavelength (nm)	Bandpass filter Wavelength (nm)	Bandpass filter bandwidth (nm)
337	420	10
337	460	20
337	570	60
337	580	60
380	460	20
380	570	60
380	580	10
380	600	10
380	640	10
460	510	10
460	580	10
460	600	10
460	620	10
460	640	10
460	660	10

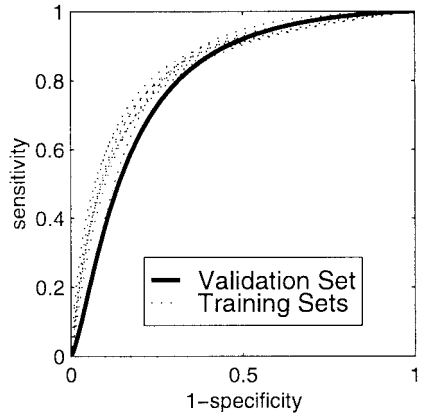
III. RESULTS

Figs. 1(a), 2(a), and 3(a) show the resulting ROC curves for each classification rule when applied to training and validation sets, each containing data from approximately half of the Study 1 patients. The training and validation set ROC curves for discriminating SN and SIL tissues [Fig. 1(a)] are approximately equal, indicating that the variance of the training set is sufficient to describe the data in the validation set and that the training set contains sufficient samples to develop an unbiased algorithm. The training set ROC curves are higher than those of the validation set for discrimination of CN and SIL tissues [Fig. 2(a)] and LG and HG SIL's [Fig. 3(a)]. This indicates that the training sets do not contain sufficient CN, LG, or HG samples to enable development of an unbiased algorithm.

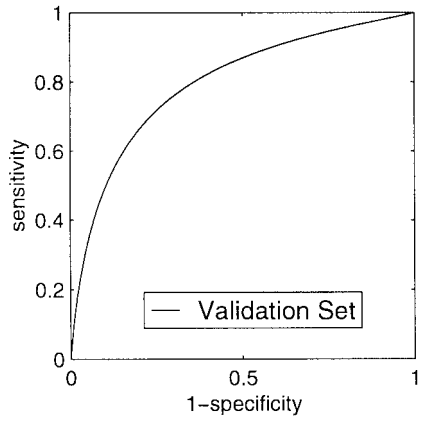
The ROC curves for each classification rule developed using cross validation with 1/8 of the patients held out at a time are shown in Figs. 1(b), 2(b), and 3(b). In each figure, the ROC curves are shown for the eight different training sets and the single validation set. In the case of the algorithm which separates SN and SIL tissues [Fig. 1(b)], the ROC curves of the eight training sets are similar to each other and to that of the validation set. The agreement between these nine curves can be used to estimate the uncertainty in the performance of the classification rule. At the point closest to ideal performance (the upper left-hand corner), the maximal variation in sensitivity and specificity is 7%. The eight training set ROC curves for discrimination of CN and SIL samples differ greatly [Fig. 2(b)], with maximal variation of 50% in sensitivity and specificity at the point closest to the gold standard, indicating that the training set still contains



(a)

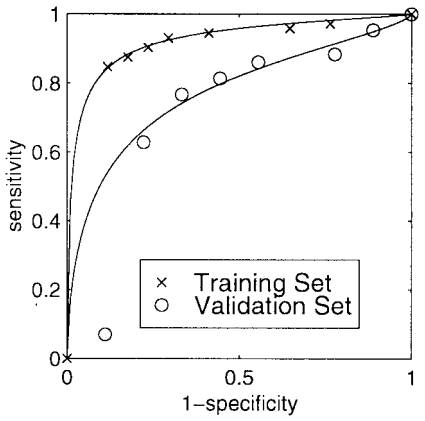


(b)

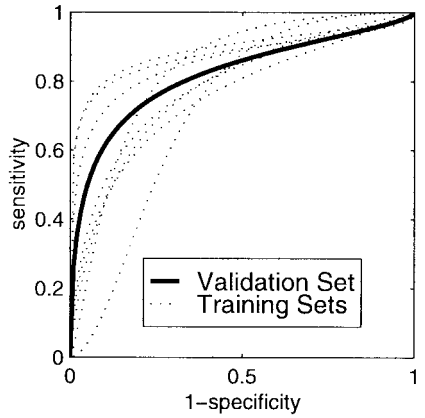


(c)

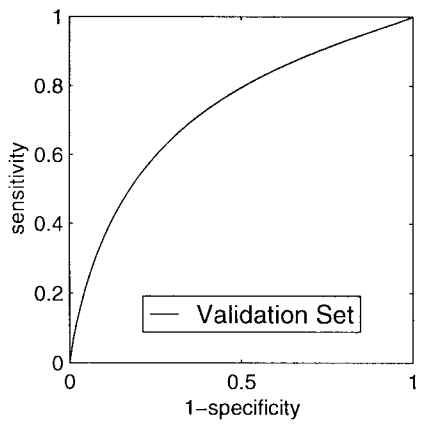
Fig. 1. Performance estimates for the classification rule which separates SN and SIL cervical tissues: (a) comparison of training and validation set ROC curves for the case where each set contained data from approximately half of the patients in Study 1, (b) comparison of the eight training and one validation set ROC curves generated using cross validation, where each training set contained data from approximately seven-eighths of the patients in Study 1 and the validation set contained all of the patients in Study 1, and (c) validation set ROC curve generated using the jackknife method (cross validation of one patient by the other patients).



(a)



(b)

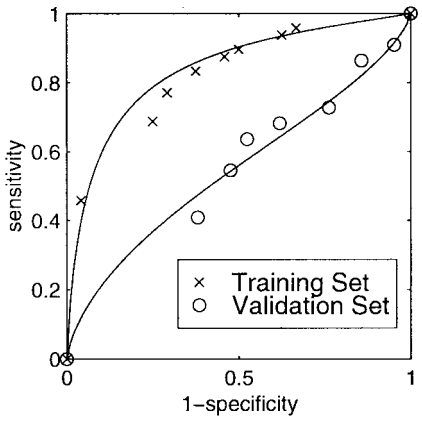


(c)

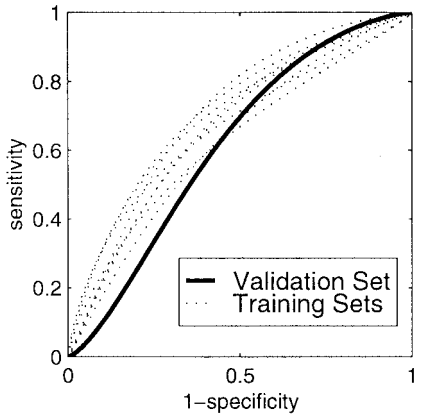
Fig. 2. Performance estimates for the classification rule which separates CN and SIL cervical tissues: (a) comparison of training and validation set ROC curves for the case where each set contained data from approximately half of the patients in Study 1, (b) comparison of the eight training and one validation set ROC curves generated using cross validation, where each training set contained data from approximately seven-eighths of the patients in Study 1 and the validation set contained all of the patients in Study 1, and (c) validation set ROC curve generated using the jackknife method.

insufficient samples to develop unbiased algorithms. The training set curves are much more similar for the classification of LG and HG SIL tissues (15% maximal variation in sensitivity and specificity), but the validation set curve shows a small decrease relative to that of the training set [Fig. 3(b)].

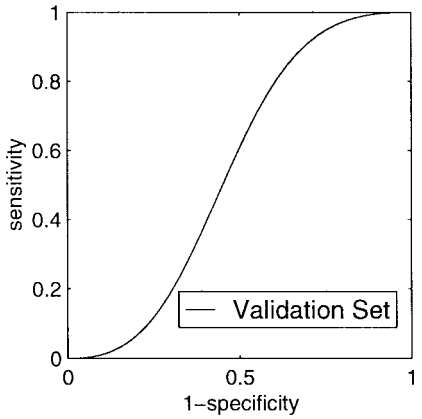
Figs. 1(c), 2 (c), and 3(c) show validation set results for cross validation with a single patient's data held out at a time. In all cases, good agreement was observed between all the training set ROC's (data not shown) and the validation set ROC, indicating sufficient sample size.



(a)



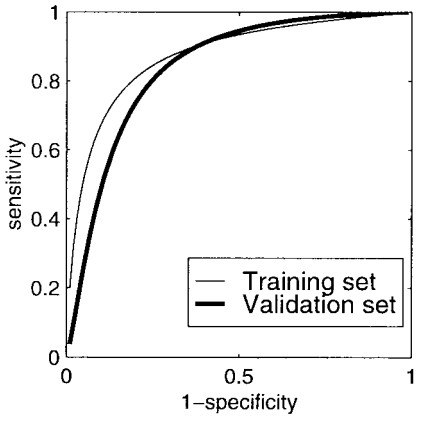
(b)



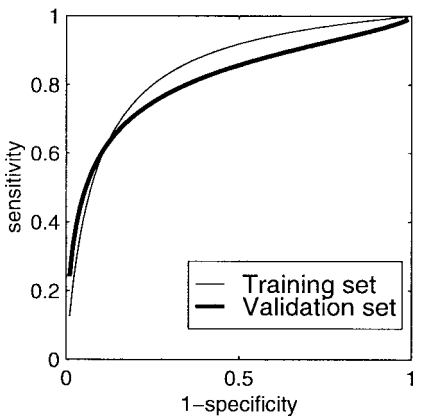
(c)

Fig. 3. Performance estimates for the classification rule which separates LG SIL and HG SIL cervical tissues: (a) comparison of training and validation set ROC curves for the case where each set contained data from approximately half of the patients in Study 1, (b) comparison of the eight training and one validation set ROC curves generated using cross validation, where each training set contained data from approximately seven-eighths of the patients in Study 1 and the validation set contained all of the patients in Study 1, and (c) validation set ROC curve generated using the jackknife method.

Fig. 4(a) shows the training and validation set ROC curves for the discrimination of SN and SIL tissues for the case where the training and validation sets each contain data from half of the Studies 1 and 2 patients. Fig. 4(b) shows the training and validation set ROC curves for the discrimination



(a)



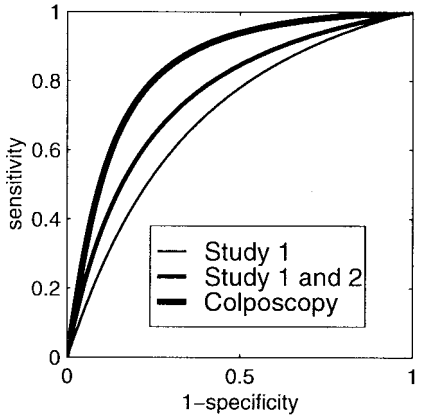
(b)

Fig. 4. (a) Performance estimates for the classification rule which separates SN and SIL cervical tissues. Comparison of training and validation set ROC curves for the case where each set contained data from approximately half of the patients in Studies 1 and 2. (b) Performance estimates for the classification rule which separates CN and SIL cervical tissues. Comparison of training and validation set ROC curves for the case where each set contained data from approximately half of the patients in Studies 1 and 2.

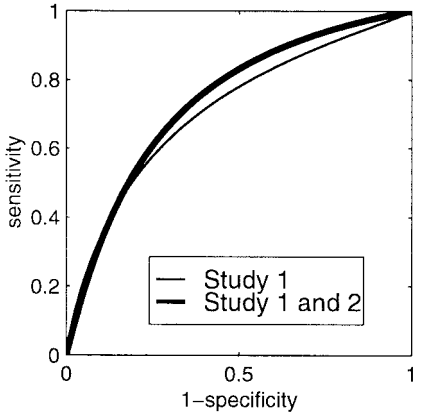
of CN and SIL tissues for the same data. In each case, the agreement between the training and validation sets is improved relative to that shown in Figs. 1(a) and 2(a), with the greatest improvement for the algorithm discriminating CN and SIL tissues. Furthermore, the validation set curves for the combined studies are shifted upward and to the left relative to those from Study 1.

The estimated ROC curves of the composite algorithms to separate SIL's from all tissues and HG SIL's from all tissues are shown in Fig. 5(a) and (b), respectively. In each case, composite algorithm performance is shown for the validation set for data from Study 1 alone and the data from Studies 1 and 2 combined. In Fig. 5(a), the estimated ROC curve for colposcopy in the referral setting [10] is also shown for reference.

Fig. 6 shows the effect of reducing the SNR of the validation set data on the classification rule which separates SN and SIL tissues in Study 1. As the SNR decreases, the ROC curves shifts downward and to the right, becoming a test with no discriminative ability (sensitivity + specificity = 1) when the SNR = 2. The algorithm performance is relatively insensitive



(a)



(b)

Fig. 5. (a) Validation set ROC curves for the composite algorithm which separates SIL's from other tissues. The composite algorithm is derived from the two classification rules which separate SN and SIL tissues and CN and SIL's tissues. Results are shown for algorithms: 1) derived from and applied to data from Study 1 alone using cross validation and 2) derived from a training set containing data from half the patients in Studies 1 and 2 and applied to a validation set containing data from the other half of the patients in Studies 1 and 2. The ROC curve for colposcopy in the referral setting is also shown [8]. (b) Validation set ROC curve for the composite algorithm which separates HG SIL's from other tissues. The composite algorithm is derived from the three classification rules which separate SN and SIL tissues, CN and SIL's tissues, and LG and HG SIL's. Results are shown for algorithms: 1) derived from and applied to data from Study 1 alone using cross validation are shown and 2) derived from a training set containing data from half the patients in Studies 1 and 2 and applied to a validation set containing data from the other half of the patients in Studies 1 and 2.

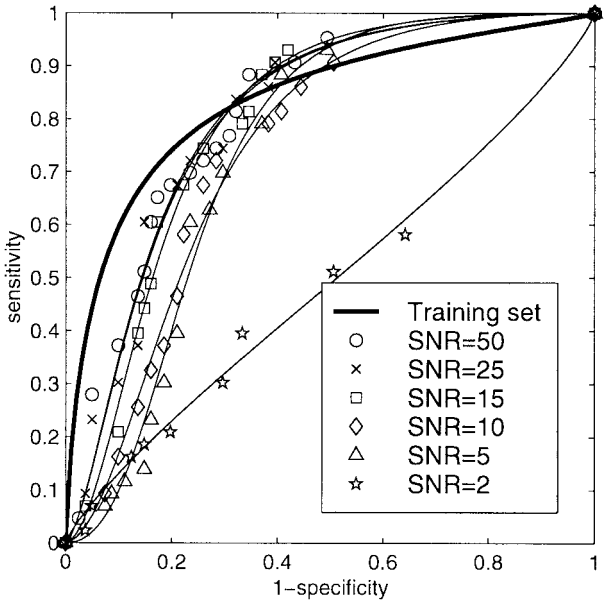


Fig. 6. Validation set ROC curves for the classification rule which separates SN and SIL tissues as a function of validation set SNR. Results are shown for an algorithm derived from and applied to data from Study 1 using cross validation.

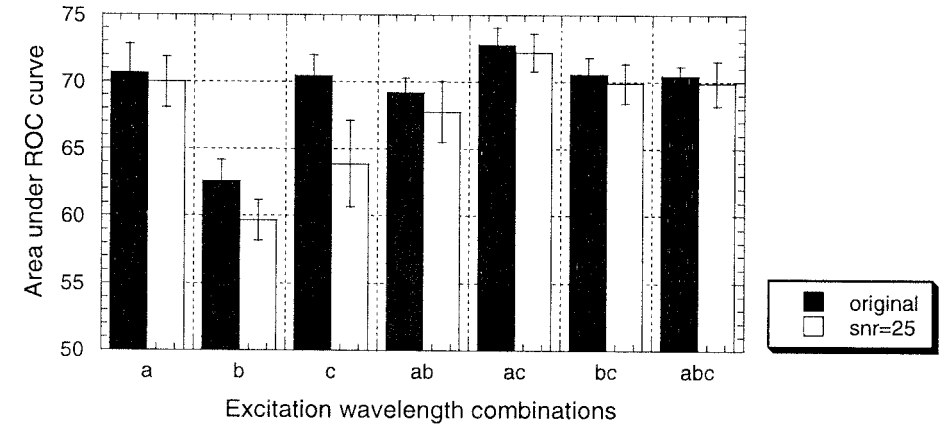
to large changes in data SNR, and performance does not drop substantially until a SNR of ten.

Fig. 7 shows the effect of reducing the number of excitation wavelengths on the area under the validation set ROC curve (AUC) for Study 1. The two composite algorithms were developed and tested using cross validation with data from seven-eighths of the patients. The calculation was repeated for ten different randomly assigned groupings of patients into eight sets and the average AUC and the standard deviation were calculated. Fig. 7(a) shows the AUC for the SIL versus non-SIL algorithm when the complete emission spectra are used. At the highest SNR, the performance of all possible combinations, except the single excitation wavelength of 380 nm, performs within one standard deviation of all three excitation wavelengths. At a reduced SNR of 25, all combinations except the choices of 380-nm excitation alone or 460-nm excitation alone perform within a standard deviation of all three excitation wavelengths. Fig. 7(b) shows the AUC for the HG versus non-HG algorithm. At both SNR's, only the performance of 337-nm excitation alone, 337- and 380-nm excitation together and 337- and 460-nm excitation together, are within one standard deviation of all three excitation wavelengths.

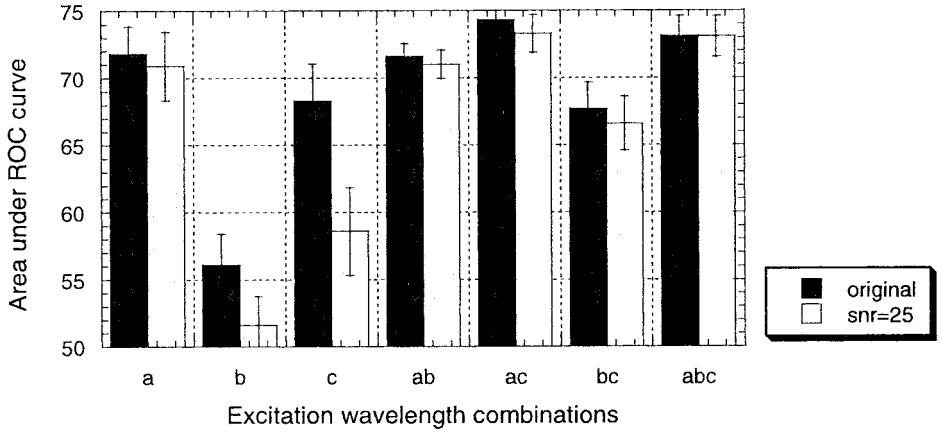
Fig. 8 shows similar results with a reduced set of emission wavelengths. Fig. 8(a) shows the performance of the SIL versus non-SIL composite algorithm; Fig. 8(b) shows the performance of the HG versus non-HG algorithm. For both composite algorithms and at both SNR's, the use of 337-nm excitation alone, the combination of 337- and 380-nm excitation and the combination of 337- and 460-nm excitation perform within one standard deviation of all three excitation wavelengths.

IV. Discussion and Conclusions

The performance of Bayesian classifiers based on fluorescence spectra obtained at 337-, 380- and 460-nm excitation compares well to that of colposcopy. In populations with a lower disease prevalence, the diagnostic ability of fluorescence is increased. At the excitation wavelengths examined in this study, the ability to discriminate SN and SIL tissues is greatest, and the composite algorithm performance is limited primarily by the relative difficulty in distinguishing CN and SIL tissues and LG and HG SIL's. These excitation wavelengths were selected on the basis of a pilot study [18], in which fluorescence excitation emission matrices (EEM's) were measured from ten paired biopsies from colposcopically normal and abnormal squamous epithelium from ten patients. EEM's were measured from 250- to 500-nm excitation and from 260- to 700-nm emission *in vitro*. Resulting data were



(a)



(b)

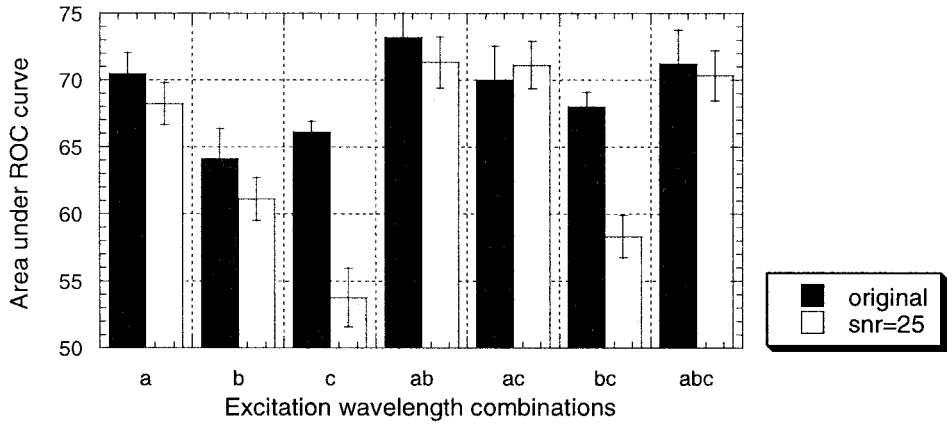
Fig. 7. Area under the ROC curve for: (a) the SIL versus non-SIL algorithm and (b) the HG versus non-HG algorithm for different combinations of excitation wavelengths ($a = 337$, $b = 380$, $c = 460$ nm). Results are shown for an algorithm derived from and applied to data from Study 1 using cross validation.

compared to determine optimal excitation wavelengths (three out of 26) for differentiating squamous normal samples and SIL's, and these wavelengths were used in subsequent clinical trials. Data were not obtained from columnar normal tissue and there were insufficient samples to compare LG and HG SIL's or inflammation and SIL's. Thus, incorporation of other excitation wavelengths may improve the performance of the composite algorithms and is the subject of further research.

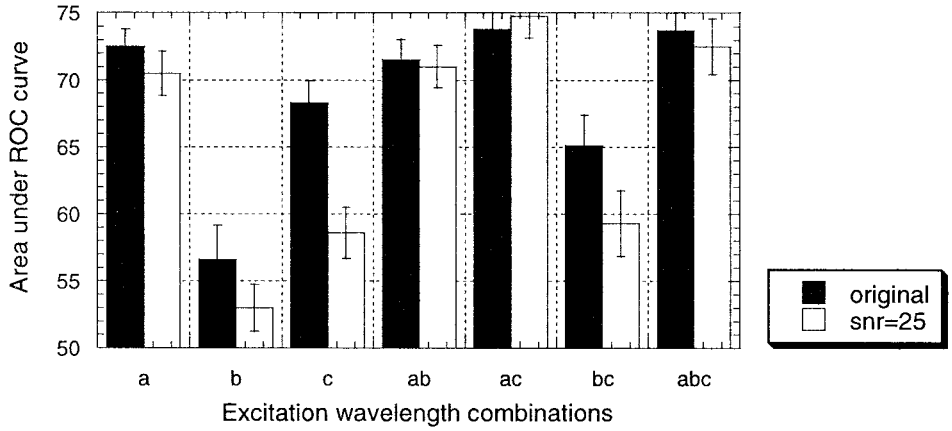
In order to interpret whether changes in the performance of the proposed algorithms are clinically significant, we performed a meta-analysis to estimate the ROC curves for a number of clinical tests used to diagnose cervical dysplasia, including the standards of care, repeat Papanicolaou smear and colposcopy, with emerging technologies such as cervicography, HPV testing, and fluorescence spectroscopy [10], [11]. We compared either the area under the ROC curve, or the point closest to the upper left-hand corner where sensitivity equals specificity (the Q point). The Q point for colposcopy is 77% $\pm$ 7%, while that for repeat Papanicolaou smear is 70% $\pm$ 2% [11]. Thus, a shift in sensitivity or specificity of between 2% and 7% is likely to be clinically significant. The area under the ROC curve for colposcopy is 0.84, that of the Papanicolaou smear is 0.76; other emerging technologies have AUC's which range from 0.71–0.75 [11]. Colposcopy is the standard of care for diagnosis of cervical dysplasia, thus a drop in AUC of (0.84–0.76) = 0.08 is clinically significant, and smaller drops may also be significant.

The agreement between the training set and validation set ROC curves for fluorescence spectroscopy reported in this paper is very dependent on the size of the training set. If the training set is too small, overtraining results. As a result, the ROC curve of the validation set is shifted downward and to the right relative to that of the training set. When cross validation is used, the agreement between the various training set and validation set ROC curves provides a quantitative manner to assess the uncertainty associated with the sensitivity and specificity estimates. This work indicates that, at the excitation wavelengths presented here, data from 85 SN, 25 CN, and 120 SIL (50 LG and 70 HG) samples are required to estimate sensitivity and specificity within 7%, just within the variation which is likely to be clinically significant.

Combining data from two studies of patients with different disease prevalence showed an increased performance of the composite algorithms. Recent work [15] indicates that there may be differences in the fluorescence spectra from colposcopically normal cervical tissue in women with and without a history of an abnormal Papanicolaou smear at these excitation wavelengths.



(a)



(b)

Fig. 8. Performance of: (a) the SIL versus non-SIL algorithm and (b) the HG versus non-HG algorithm for different combinations of excitation wavelengths and the reduced emission wavelengths listed in Table II ($a = 337$, $b = 380$, $c = 460$ nm). Results are shown for an algorithm derived from and applied to data from Study 1 using cross validation.

All measurements reported by our laboratory to date have been performed with scientific grade equipment, which results in spectra with good SNR and high spectral resolution. This work shows that significant cost reduction is possible without compromising the diagnostic ability of the technique. Fig. 6 shows that as the SNR is reduced from of order 300 to 15, the drop in the area under the ROC curve is not clinically significant. Below this SNR, performance drops rapidly. Data acquisition electronics used to capture video signals with a dynamic range of 4 b to 8 b is sufficient to generate this SNR. Furthermore, significant reductions in the number of excitation and emission wavelengths do not significantly reduce diagnostic ability. Figs. 7 and 8 show, that as the number of emission wavelengths is reduced from 160 to 15 and as the number of excitation wavelengths is reduced from three to one, the changes in the area under the ROC curve are not clinically significant. Measuring fluorescence with one or two excitation wavelengths and 15 emission wavelengths can be accomplished with dielectric bandpass filters mounted in a filter wheel. This opens the opportunity to gain diagnostic information from fluorescence images of the whole cervix, because filters can be integrated into imaging optics. The use of UV excitation at 337 nm was found to be important; all combinations of one or two excitation wavelengths which performed similarly to the total of three included 337-nm excitation. Therefore cost reduction achieved by using illumination optical components made of plastic would result in poor algorithm performance. In summary, as the results presented in this paper show, the quality (and, thus, the cost) of the spectrometer can be reduced without affecting algorithm performance.

A method to estimate the expected SNR of various spectrometers for measuring cervical tissue was presented in [19], and this technique can be used to estimate ROC curves as a function of the economic cost of different spectrometers. Decision-analytic methods [6] can then be used to rate various designs based on their incremental cost-effectiveness, which will provide a rational guide for cost-effective instrument design for an emerging diagnostic technology. Our previous decision analysis [6] showed that a strategy based on a combination of fluorescence spectroscopy and colposcopy was both more effective and less expensive than the standard of care for diagnosis and treatment of cervical precancer, based on the outcome of dollars spent per case of high-grade precancer detected. However, it is difficult to assess accurately the costs of fluorescence spectroscopy because the technology is still in development. In estimating the cost of any diagnostic

technology, one must consider two components: 1) the cost of the diagnostic test itself and 2) the cost of physician time. As the costs of fluorescence spectroscopy were varied from 80% to 100% that of colposcopy, the results of our decision analysis stayed the same [6]. The second component, cost of physician time, depends directly on the degree to which the new technology has diffused into clinical practice. In the early phases of diffusion into clinical practice, we expect that physicians will perform the fluorescence spectroscopy based diagnostic test. As the technique becomes accepted widely, nurse practitioners or trained technicians can be expected to perform spectroscopy procedures. Our analysis [6] showed that the results were sensitive to the extent of technology diffusion. Thus, to provide a cost-effective diagnostic tool, the instrument costs of fluorescence spectroscopy devices must approach those of colposcopy and the instruments must be robust enough that they can be operated by less trained personnel.

The importance of considering such issues at an early stage is pointed out by the recent experience with improved tools for preparing Papanicolaou smears [20]. The ThinPrep test provides improvements in sensitivity, but at a higher cost, which could add $1 billion to health care costs annually in the United States without proven evidence that it saves more lives than a regular Papanicolaou smear [20]. Thus, for emerging technologies such as optical spectroscopy to be adopted will require evidence that they are both accurate and cost-effective. Our results here and in [6] show that, through careful spectrometer design, optical spectroscopy has the potential to be both more effective and less expensive than the current standard or care. This may provide a rare opportunity for patients, providers and payers to be on the same side of a new technology.

REFERENCES

[1] R. Richards-Kortum and E. Sevick-Muraca, "Quantitative optical spectroscopy for tissue diagnosis," *Ann. Rev. Phys. Chem.*, vol. 47, pp. 555–606, 1996.

[2] S. Nishioka, "Laser-induced fluorescence spectroscopy," *Gastrointestinal Endosc. Clin. North Amer.*, vol. 4, no. 2, pp. 313–326, Apr. 1994.

[3] G. A. Wagnières, W. M. Star, and B. C. Wilson, "*In vivo* fluorescence spectroscopy and imaging for oncological applications," *Photochem. Photobiol.*, vol. 68, no. 5, pp. 603–632, Nov. 1998.

[4] A. Laupacis, D. Feeny, A. S. Detsky, and P. X. Tugwell, "How attractive does a new technology have to be to warrant adoption and utilization?," *Tentative Guidelines for Using Clinical and Economic Evaluations, Can. Med. Assoc. J.*, vol. 146, no. 4, pp. 473–481, Feb. 1992.

[5] M. C. Weinstein and H. V. Fineberg, *Clinical Decision Analysis.* Philadelphia, PA: Saunders, 1980.

[6] S. B. Cantor, M. Mitchell, G. Tortolero-Luna, C. S. Bratka, D. Bodurka, and R. Richards-Kortum, "Cost-effectiveness analysis of the diagnosis and management of cervical squamous intraepithelial lesions," *Obstet. Gynecol.*, vol. 91, no. 2, pp. 270–277, Feb. 1998.

[7] P. T. Wertlake, K. Francus, G. R. Newkirk, and G. P. Parham, "Effectiveness of the Papanicolaou smear and speculoscopy as compared with the Papanicolaou smear alone: A community-based clinical trial," *Obstet. Gynecol.*, vol. 90, no. 3, pp. 421–427, Sept. 1997.

[8] A. Albert and E. Harris, *Multi-Variate Analysis of Clinical Laboratory Data.* New York: Marcel Dekker, 1987.

[9] C. E. Metz, "Basic principles of ROC analysis," *Semin. Nucl. Med.*, vol. 8, no. 5, pp. 283–298, Oct. 1978.

[10] M. F. Mitchell, D. Schottenfeld, G. Tortolero-Luna, S. Cantor, and R. Richards-Kortum, "Colposcopy for the diagnosis of squamous intraepithelial lesions: A meta analysis," *Obstet. Gynecol.*, vol. 91, no. 4, pp. 626–631, Apr. 1998.

[11] M. F. Mitchell, S. Cantor, G. Tortolero-Luna, N. Ramanujam, and R. Richards-Kortum, "ROC curves for fluorescence spectroscopy for the diagnosis of squamous intra-epithelial lesions," *Obstet. Gynecol.*, vol. 93, pp. 462–470, 1999.

[12] P. Nieminen, M. Kallio, and M. Hakama, "The effect of mass screening on incidence and mortality of squamous and adenocarcinoma of cervix uteri," *Obstet. Gynecol.*, vol. 85, no. 6, pp. 1017–1021, June 1995.

[13] R. J. Kurman, D. E. Henson, A. L. Herbst, K. L. Noller, and M. H. Schiffman, "Interim guidelines for management of abnormal cervical cytology," in *Proc. The 1992 NCI Workshop, JAMA*, June 1994, vol. 271, no. 23, pp. 1866–1869.

[14] N. Ramanujam, M. F. Mitchell, A. Mahadevan-Jansen, S. L. Thomsen, G. Staerkel, A. Malpica, T. Wright, N. Atkinson, and R. Richards-Kortum, "Cervical precancer detection using a multivariate statistical algorithm based on laser-induced fluorescence spectra at multiple excitation wavelengths," *Photochem. Photobiol.*, vol. 64, no. 4, pp. 720–735, 1996.

[15] C. Brookner, U. Utzinger, G. Staerkel, R. Richards-Kortum, and M. F. Mitchell, "Cervical fluorescence of normal women," *Lasers Surg. Med.*, vol. 24, pp. 29–37, 1999.

[16] B. Littenberg and L. E. Moses, "Estimating diagnostic accuracy from multiple conflicting reports: A new meta-analytic method," *Med. Decision Making*, vol. 13, no. 4, pp. 313–321, Oct./Nov./Dec. 1993.

[17] P. A. Lachenbruch, *Discriminant Analysis.* New York: Hafner, 1975.

[18] A. Mahadevan, M. F. Mitchell, E. Silva, S. Thomsen, and R. Richards-Kortum, "Study of the fluorescent properties of normal and neoplastic human cervical tissue," *Lasers Surg. Med.*, vol. 13, no. 6, pp. 647–655, 1993.

[19] E. V. Trujillo, D. Sandison, U. Utzinger, N. Ramanujam, M. F. Mitchell, and R. Richards-Kortum, "Method to determine tissue fluorescence efficiency *in vivo* and predict signal-to-noise ratios of spectrometers," *Appl. Spectroscopy*, 1998, in press.

[20] *Improved Cancer Screening Costs More—Insurers skeptical on better Papanicolaou test*, Assoc. Press. Boston, MA: Boston Globe, Nov. 19, 1998.

Urs Utzinger received the M.S. degree in mechanical engineering in 1989 from the Swiss Federal Institute of Technology (ETH), Zürich Switzerland. He worked on spectroscopic guided laser coronary angioplasty at the Institute of Biomedical Engineering and Medical Informatics at the ETH and the University Hospital Zürich and received the Ph.D. degree in 1995.

He is a Research Associate at the University of Texas, Austin, in the Spectroscopy Laboratory of the Department of Electrical and Computer Engineering. Since 1995, he has been developing spectroscopic techniques and instrumentation for the diagnosis of epithelial lined tissue at the University of Texas and the M. D. Anderson Cancer Center, Houston, TX.

E. Vanessa Trujillo received the B.S. degree in electrical engineering from the University of Texas, Austin, in 1995, where she worked on tissue fusion research. She designed cost-effective systems for the detection of cervical precancer using fluorescence spectroscopy and received the M.S. degree in electrical engineering in 1997.

She is an Applications Engineer at National Instruments, Austin, TX.

E. Neely Atkinson received the B.A. degree in English from Rice University, Houston, TX, in 1975. He received the M.A. and Ph.D. degrees in mathematical sciences from Rice University in 1981.

He is currently an Associate Professor of biomathematics at the University of Texas, M. D. Anderson Cancer Center, Houston, TX. His research interests include computational statistics, regression analysis, computer science, and modeling of cancer processes.

Michele F. Mitchell received the B.A. degree from the University of Michigan, Ann Arbor, in 1975, the M.D. degree from the University of Michigan Medical School in 1980, and the M.S. degree in clinical research design from the University of Michigan in 1989.

She is an Associate Professor of Gynecologic Oncology at the University of Texas M. D. Anderson Cancer Center. She is the Director of Colposcopy at the M. D. Anderson Cancer Center and the Division Director of Gynecologic Oncology at the UT Health Science Center–Lyndon Baines Johnson Hospital, Houston, TX. Her research interests include the use of optical spectroscopy and imaging for detection of cervical precancer and treatment of preinvasive cervical neoplasia with chemo-preventive agents.

Scott B. Cantor received the B.S. degree in applied mathematics from Yale University, New Haven, CT, in 1981 and the Ph.D. degree in decision sciences from Harvard University, Cambridge, MA. in 1991.

He is an Assistant Professor of medicine in the section of general internal medicine in the Department of Medical Specialties at the University of Texas M. D. Anderson Cancer Center, Houston. He is a Clinical Decision Analyst. His research focuses on theoretical issues concerning cost-effectiveness analysis and diagnostic testing, and on the application of decision analysis for the screening and diagnosis of cancer.

Rebecca Richards-Kortum received the B.S. degree with highest distinction in physics and mathematics from the University of Nebraska, Lincoln, in 1985, the M.S. degree in physics and the Ph.D. degree in medical physics from the Massachusetts Institute of Technology (MIT), Cambridge, in 1987 and 1990, respectively.

In 1990, she joined the faculty at the University of Texas, Austin, where she has established collaborative research projects with the University of Texas M. D. Anderson Cancer Center using optical spectroscopy to detect precancer and cancer of the uterine cervix, oral cavity, and ovary. Her research goals range from describing the basic physics of light propagation in scattering tissues to developing hardware systems for the collection and analysis in the clinical setting.