Calibration standards for multicenter clinical trials of fluorescence spectroscopy for *in vivo* diagnosis

Nena M. Marín
University of Texas
Department of Biomedical Engineering
Austin, Texas 78712

Nicholas MacKinnon
Calum MacAulay
British Columbia Cancer Research Centre
Cancer Imaging Department
610 West 10th Avenue
Vancouver, British Columbia V5Z 1L3 Canada

Sung K. Chang
Massachusetts General Hospital
Wellman Center for Photomedicine
Boston, Massachusetts 02114

E. Neely Atkinson
University of Texas, M. D. Anderson Cancer Center
Department of Biomathematics
1515 Holcombe Boulevard
Houston, Texas 77030

Dennis Cox
Rice University
Department of Statistics
6100 South Main
Houston, Texas 77005

Dan Serachitopol
Brian Pikkula
University of Texas, M. D. Anderson Cancer Center
Center for Biomedical Engineering
Houston, Texas 77030

Michele Follen
University of Texas, M. D. Anderson Cancer Center
Center for Biomedical Engineering and
     Department of Gynecologic Oncology
Houston, Texas 77030
     and
The University of Texas Health Science
     Center in Houston
Department of Obstetrics, Gynecology, and
     Reproductive Sciences
Houston, Texas 77030

Rebecca Richards-Kortum
Rice University
Department of Bioengineering
6100 South Main
Houston, Texas 77005

Abstract. In the context of clinical trials, calibration protocols for optical instruments that ensure measurement accuracy and the ability to carry out meaningful comparisons of data acquired from multiple instruments are required. A series of calibration standards and procedures are presented to assess technical feasibility of optical devices for cervical precancer detection. Measurements of positive and negative standards, and tissue are made with two generations of research grade spectrometers. Calibration accuracy, ability of standards to correct and account for changes in experimental conditions, and device components are analyzed. The relative frequency of measured calibration standards is investigated retrospectively using statistical analysis of trends in instrument performance. Fluorescence measurements of standards and tissue made with completely different spectrometers show good agreement in intensity and lineshape. Frequency of wavelength calibration standards is increased to every 2 h to compensate for thermal drifts in grating mount. Variations in illumination energy detected between standards and patient measurements require probe redesign to allow for simultaneous acquisition of illumination power with every patient measurement. The use of frequent and well-characterized standards enables meaningful comparison of data from multiple devices and unambiguous interpretation of experiments among the biomedical optics community. © *2006 Society of Photo-Optical Instrumentation Engineers.* [DOI: 10.1117/1.2166389]

1 Introduction

Optical spectroscopy has the potential to probe the molecular and morphologic conformation of tissue.[1,2] As such, it provides a method to quantify biochemical and architectural changes in tissue that may provide early subtle indicators of disease or injury.[3–8] Moreover, unlike the removal of tissue for diagnostic biopsy, optical measurements are minimally invasive and can be analyzed in real time. The combination of reduced invasiveness, reduced patient discomfort, and immediate feedback to clinicians is a powerful incentive to adopt these methods.

Toward this end, optical measurement systems have been developed and tested to study the diagnostic ability of

Address all correspondence to Rebecca Richards-Kortum, Department of Bioengineering, Rice University, 6100 Main Street, Houston, TX 77005. Tel: 713–348–5869. Fax: 713–348–5877. E-mail: rkortum@rice.edu
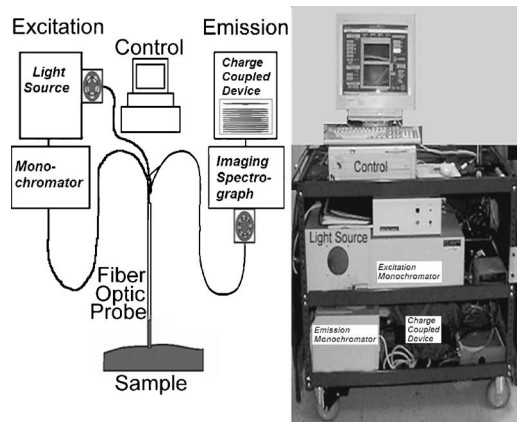
Fig. 1 Block diagram and photograph of the FastEEM2 system used to measure fluorescence at 16 excitation wavelengths and diffuse reflectance at four source-detector separations.

changes in absorption, scattering, and fluorescence properties of tissue.[9–14] Because of the high cost of clinical investigations, many of these studies have been limited in scope or in population size, and are generally classified as pilot studies. Trials involving larger numbers of patients are often sponsored by industry, typically with relatively narrow objectives. Our group has had the opportunity through a National Institutes of Health (NIH) program project grant to undertake large multicenter studies to investigate a range of optical methods for diagnosis of cervical neoplasia. In these studies, we used a model of technology assessment[15] to simultaneously test and refine emerging optical technologies for the detection of cervical cancer and its precursors.

In any complex optical measurement or imaging system, changes will occur in device components that can affect optical transfer function, signal transduction, or spatially dependent system response. An example of variation in optical transfer function is a change in filter characteristics due to photobleaching or thermal effects. An example of a transduction change is sensor background level change or increased noise due to electromagnetic interference (EMI) or breakdown in cables or connectors. Examples of spatially dependent factors are position shifting in optical mounts due to wear or thermal expansion. These effects may be combinatoric, compensatory, or synergistic. For example, a thermal effect could act on all three of these areas simultaneously. While these effects are not unanticipated, it is difficult to predict which factors or combination of factors will appear at any given time during a clinical study.

In these clinical studies, we developed a research-grade optical spectrometer to measure fluorescence emission spectra at up to 24 excitation wavelengths and reflectance spectra at up to 6 source-detector separations; we refer to this device as the FastEEM. Our goal is to fully integrate data from two generations of the FastEEM device and from multiple instruments at multiple sites.

Initially, we used a spectrometer that measures fluorescence at 16 excitation wavelengths and diffuse-reflectance at four source-detector separations[16] (FastEEM2). The system, as shown in Fig. 1, consists of four main components: (1) an arc lamp, stepper-motor driven monochromator and filter

wheel, which provides monochromatic and broadband excitation; (2) a fiber optic probe, which directs excitation light to the tissue and collects remitted fluorescence from one location and diffusely reflected light from four locations; (3) a filter wheel, imaging spectrograph, and CCD camera, which detects the spectrally resolved reflectance and fluorescence signals; and (4) quality control components ensuring wavelength, spectral sensitivity, and power calibration. The excitation monochromator position, filter wheel position, spectrograph grating position, CCD operation, and data acquisition are controlled using a personal computer.

Subsequently, we used a similar spectrometer that measures fluorescence over an extended range of 24 excitation wavelengths and diffuse reflectance at six source-detector separations (FastEEM3). The FastEEM3 system (Fig. 2) consists of four main components: (1) an arc lamp combined with filter wheels and a fiber optic coupling system; (2) a fiber optic probe that directs light onto tissue and collects fluorescence and elastically scattered light at 6 source-detector separations; (3) a filter wheel, spectrograph, and CCD camera that disperses and records the collected light spectrally and restricts excitation light from entering the analyzer; and (4) expanded quality control components ensuring wavelength, spectral sensitivity, and power calibration. The data acquisition process is fully automated and computer controlled. Figure 3 shows a diagram of the quality control components of FastEEM3. Optical standards with known responses (positive standards) and standards expected to have minimal optical response (negative standards) are an essential part of the device and should be constantly monitored.

Our program project grant has many goals, but a significant one is to try to establish paradigms that might assist academic and industrial groups to sponsor cost-effective clinical trials that will lead to the transfer of these promising technologies into clinical practice. A significant aspect of the program is to examine the calibration of photonic measurement systems and the quality of optical measurements to facilitate integration of data from multiple instruments and sites and ensure that diagnostic algorithms are transportable with high accuracy. Fluorescence- and reflectance-based point spectroscopy and imaging systems were successfully applied by many groups in many organ sites for noninvasive probing of tissue properties.[17–27] While calibration standards are widely used for preprocessing of the acquired data, methodologies for instrument characterization, validation, and radiometric traceability lack standardization. The National Institute of Standards and Technology (NIST) has recognized that a major factor inhibiting rapid growth is the inability to make comparable fluorescence intensity measurements across laboratories.[27] We developed statistical analysis methods and integrated these with engineering analysis with the aim of simplifying the design issues and helping ensure validity of data for future clinical trials involving medical photonics technologies. In this paper, we examine some of these methods in the context of the calibration and quality assurance of one part of our measurement set: tissue fluorescence properties.

2 Methods

The role of standards in optical instrumentation is twofold—calibration and quality assurance.
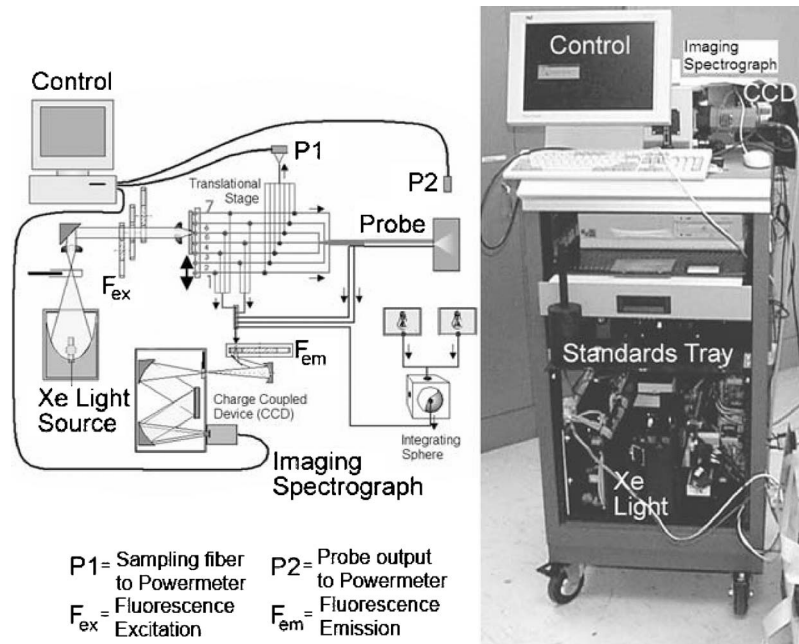
Fig. 2 Block diagram and photograph of FastEEM3 system used to measure fluorescence at 24 excitation wavelengths and diffuse reflectance at six source-detector separations.

2.1 Calibration

The fundamental role of standards is to provide a sound physical basis to map the measured signal to the true response of the sample being measured. All changes of energy originating from sources other than the measured sample are considered system-dependent signals that must be detected and corrected using the appropriate calibration standards. Standards also quantify unwanted background signal levels and facilitate background detection and removal. It is important to use both positive standards (standards with a known, strong optical signature) and negative standards (standards where no response is expected) to enable this calibration. Standards that bracket all data dimensions, including time, are required to ensure accurate translation from measured data to calibrated results. Careful consideration of the nature of standards and frequency of their measurement is necessary to guarantee that all dimensions are calibrated correctly.
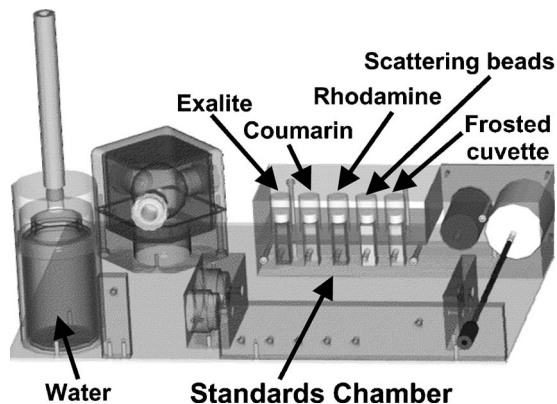
2.2 Quality Assurance

Standards are also used to track the device response as a function of time to mitigate various operational issues that may change over time. Standards provide a well-known blue print of expected (ideal) results. Measurement and analysis of data from appropriate standards with sufficient frequency can be used to easily identify any variation from this blue print; this can be used to correct previously acquired data and to modify instrumentation to prevent similar problems in future data acquisition.

The calibrated FastEEM data should become device independent and should be able to be combined for analysis with data acquired with other calibrated optical devices.[8,16–19] In this study, two generations of the FastEEM system were tested clinically at three different sites: the University of Texas (UT) M.D. Anderson Cancer Center in Houston (FastEEM2); the Lyndon B. Johnson Harris County Hospital District in Houston, Texas (FastEEM2); and the British Columbia Cancer Agency (BCCA) in Vancouver, Canada (FastEEM3). The Institutional Review Board of the UT at Austin, M.D. Anderson Cancer Center in Houston, and the BCCA in Vancouver approved the conduct of this study, and all subjects gave written, informed consent before study participation.

2.3 Measured Standards

It is critical to have a comprehensive set of standards in order to accommodate all data dimensions and sources of measurement variability like stages, filters, and shutters controlling integration times. With each device and at each clinical site, data were acquired from a series of standards as well as from patient sites. Tables 1 and 2 detail all positive and negative standards measured with each device.

Fig. 3 FastEEM3 calibration and performance validation standards.

Table 1 FastEEM2 calibration standards.

Standard	Type	Excitation (nm)	Emission (nm)	Frequency (Number of Measurements)
Mercury	Wavelength calibration	N/A	380–950	1/patient
Frosted cuvette	Negative	290–480	380–910	1/patient
Distilled water	Negative	290–480	380–910	1/patient
Rhodamine 610 2 mg/L solution in ethylene glycol	Positive	290–480	380–910	1/patient
Powermeter	Illumination energy calibration	N/A	N/A	1/patient
Tungsten	Intensity response calibration	N/A	N/A	9 total
Shuttered light source (background)	Negative	N/A	N/A	2/patient site
Shuttered camera (dark current)	Negative	N/A	N/A	3/patient site

These standards should be measured frequently in actual clinical settings prior to clinical trials to anticipate environmental issues, as well as throughout the entire clinical trial. Measurements in the clinic are often performed in small rooms with more than the usual number of personnel. This can sometimes result in temperature, humidity, and room light levels that differ substantially from the laboratory.

2.4 Positive Standards

The role of positive standards is to provide an input of known value; data measured from the positive standard are compared to measured output. The deviation of the measurement from the known provides correction factors for calibration to remove system-dependent responses. For the FastEEM, the correction factors calibrate for wavelength response, for variations in illumination energy as a function of time and wavelength, and for variations in emission wavelength sensitivity (optical transfer/transduction).

In selecting standards, we looked for those that were easy to obtain commercially and that covered the full UV VIS spectral range in both excitation and emission. Rhodamine, Exalite, and coumarin cover the full range of emission and excitation wavelength of interest and are easy to obtain. The choice of concentration was based on making them optically dilute. We also looked for solvents that were easy to work with and for standards with low photobleaching.

Positive standards measured with the FastEEM2 include: a HgAr calibration lamp as well as Hg lines from fluorescent room lights (later abandoned) for wavelength calibration purposes, a 4.2 μM solution of the organic fluorescent dye, rhodamine 610 (Exciton, Dayton, Ohio) in ethylene glycol to compensate for variations in overall collection throughput, manual measurements of illumination power using a power meter (Newport, Irvine, California, 818-UV) to compensate for variations in the intensity of the excitation light, and measurements of NIST traceable calibrated tungsten and deuterium lamps (550C and 45D, Optronic Laboratories Inc., Orlando, Florida) to correct for nonuniform spectral response of the detection system.

With the next-generation FastEEM3 system, we measured an expanded range of positive standards. These standards include: a HgAr calibration lamp (Ocean Optics) for wavelength calibration purposes, a 4.2 μM solution of the organic fluorescent dye rhodamine 610 (Exciton, Dayton, Ohio) in ethylene glycol, a 2.5 μM solution of the organic fluorescent dye coumarin 480 (Exciton, Dayton, Ohio) in ethylene glycol, and a 0.133 μM solution of the organic fluorescent dye Exalite 400E (Exciton, Dayton, Ohio) in ethylene glycol to compensate for variations in overall collection throughput. In addition, automated measurements of illumination power were performed using a power meter to compensate for variations in the intensity of the excitation light, and measurements of NIST traceable calibrated tungsten lamp (LS-1-CAL, Ocean Optics, Dunedin, Florida) were performed to correct for nonuniform spectral response of the detection system.

FastEEM3 power meter measurements are acquired at 10-ms intervals, beginning and ending approximately 30 ms before and after the shutter open and close time, providing an accurate measurement of actual illumination times and energy delivered. Two power meter measurements are acquired, and used to calculate the energy delivered to the tissue during data acquisition. The first measurement is made from a sampling

Table 2 FastEEM3 calibration standards.

Standard	Type	Excitation (nm)	Emission (nm)	Frequency (Number of Measurements)
HgAr	Wavelength calibration	N/A	257–805	1/day
Frosted cuvette	Negative	290–530	257–805*	1 every 2 h
Distilled water	Negative	290–530	257–805*	1 every 2 h
Rhodamine	Positive	290–530	257–805*	1 every 2 h
Coumarin 480	Positive	290–530	257–805*	1 every 2 h
Exalite 400E	Positive	290–530	257–805*	1 every 2 h
Powermeter	Illumination energy calibration	N/A	N/A	1/day
Powermeter (sampling fiber)	Illumination energy calibration	N/A	N/A	1 with every measurement
Tungsten	Intensity response calibration	N/A	N/A	1/day
Shuttered light source (background)	Negative	N/A	N/A	1 with every measurement
Shuttered camera (dark current)	Negative	N/A	N/A	1/day

optical fiber; the proximal tip of the sampling fiber is placed at the same location as the illumination fibers of the fiber optic probe and the distal tip is located at the power meter during tissue measurements. The second measurement is made from the distal tip of the fiber optic probe just prior to patient measurements. Figure 2 shows the location of the sampling fiber (labeled as P1) and the probe output (labeled as P2) in the FastEEM3 system diagram. The measurement from the sampling fiber is used to determine the optical power delivered during each tissue measurement, and the measurement from the probe output is used to calibrate the ratio of the sampling fiber measurement to the probe output.

2.5 Negative Standards

The goal of measuring negative standards is the detection of light introduced into the illumination and measurement optical paths due to leakage, stray light, or component fluorescence or contamination that can result in measurement artifacts. FastEEM2 and FastEEM3 negative standards measured include: a measurement made with the illumination source shuttered and the probe placed on the sample (background), a measurement made with the camera shutter and the illumination shutter closed (dark current), and finally two measurements made under the same measurement acquisition parameters (integration time, illumination wavelength, etc.) conditions as tissue, but first with the probe placed in a bottle of distilled water and second with the probe placed in contact with the frosted quartz cuvette. The purpose of these last two measurements is to use a sample with no intrinsic fluorescence to provide a check for device related autofluorescence. The water sample results in only small specular reflection due to the refractive index difference at the probe tip and water interface with the remaining light going forward with minimal backscattering. The frosted cuvette is expected to diffusely backscatter the illumination light in a similar way to epithelial

tissue. This enables us to monitor device related autofluorescence, which could be backscattered by tissue and contaminate the measurement. Our criteria were to reject data if the tissue signal drops below 10 times the background fluorescence.[28] Device-related autofluorescence can sometimes be comparable to tissue signals if proper materials are not used in fiber optic probe construction or as a result of defects or damage accumulated over time.

2.6 Frequency of Standards Measurements

It is important to measure standards frequently enough to capture any drift in the performance of the measurement system over time. A typical clinic day with the FastEEM2 system would result in measurements from one to two patients. Prior to each patient measurement, a full set of positive and negative standards would be run. A typical day in clinic with the FastEEM3 would result in measurements from two to three patients. A full set of standards would follow the initial 20-min device warm up. Throughout the course of the day Rhodamine measurements would be repeated every 2 h. As we learned about system performance issues during the course of these trials, we adjusted the frequency with which standards were measured.

2.7 Analysis of Measured Standards

Measured standards were analyzed in three ways—first, to process data from positive standards and tissue to calculate fluorescence spectra in calibrated units; second, to verify the calibration process by comparing processed data from positive standards to expected results; and third, to identify and quantify any measurement artifacts, providing a means for data recovery under adverse events.

In the next section, we detail each step in the processing algorithm based on these standard measurements and the quality assessment tools developed to assess each type of
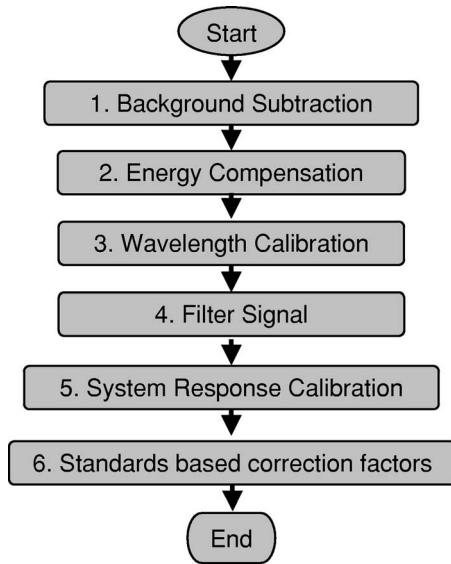
Start

1. Background Subtraction

2. Energy Compensation

3. Wavelength Calibration

4. Filter Signal

5. System Response Calibration

6. Standards based correction factors

End

Fig. 4 FastEEM data processing flow chart.

standard measurement. In each case, statistical analysis over the timeline of each collected standard was used to track the performance of the FastEEM system. Descriptive statistics including number of samples, arithmetic mean and standard deviation, median, minimum, and maximum were tabulated for the raw data, calibrated data, and calibration parameters. In addition, raw spectra were plotted over time, 10 measurements per figure to study consistency of spectral shape of measurements. These data enabled expert reviewers to identify changes in the instrument performance; these changes were then correlated with specific physical phenomena based on experts' background knowledge and events related to device maintenance or calibration recorded in the FastEEM operator's logbook.

2.8 Raw Data Processing Algorithm

Raw data recorded with the FastEEM are processed to yield system-independent data in a six-step process. Figure 4 illustrates the generalized processing algorithm flow chart applied to data acquired with both FastEEM systems.

2.8.1 Background subtraction

The raw data processing algorithm first subtracts dark current (FastEEM2) or background (FastEEM3) from the raw signal.

2.8.2 Power and exposure time normalization

NIST traceable optical power meter measurements are used to correct emission spectra acquired at each excitation wavelength for variations in the illumination energy. Power meter measurements provide the illumination optical power (FastEEM2 and 3) and integration time (FastEEM3) at each excitation wavelength. For data acquired with FastEEM2, the programmed exposure time was used as the integration time. For FastEEM3, the number of data points in power meter measurements acquired at 10-ms intervals (accounting approximately 30 ms before and after for shutter openning and closing time) provided an accurate measurement of actual illumination time. The spectra collected at each excitation

wavelength are then divided by the product of the illumination power and the integration time. The measured power values as a function of excitation wavelength were plotted over time throughout the trial and analyzed to identify changes or inconsistencies in the spectrum of illumination power.

2.8.3 Wavelength calibration

Wavelength calibration was performed using the spectra of Hg lines from room fluorescent lights (FastEEM2) and from a HgAr lamp (FastEEM2 and FastEEM3). Five known mercury peaks are detected in the measured spectra. A least-squares linear fit relating CCD pixel position to wavelength is performed at these five points. The slope and intercept of this fit is used to assign a wavelength value to each point in all spectra. For each measurement, the slope and intercept of this calibration are compared to historical measurement means (averages) to identify questionable measurements. Deviations from the mean can be associated with changes in the grating incident angle and camera realignment (magnification) changes.

2.8.4 Filtering

The calibrated, background-corrected signal is filtered using Savitzky-Golay smoothing in order to reduce unwanted noise.[29] For the fastEEM2 CCD width of 1600 pixels, the tissue background and dark current were filtered with a smoothing window half width of 80 and a first-order polynomial. The FastEEM2 tissue spectra were filtered with a smoothing window half width of 4 and a first-order polynomial. These parameters were selected by evaluating a range of values including higher order polynomials and comparing their effect on narrow peaks such as porphyrin and broad peaks such as NADH. For the FastEEM3, the parameters were chosen to avoid loss of spectral features while maintaining proportional relationship to the FastEEM2 CCD width and pixel size. The background and dark current were smoothed with a window half width of 80 and the tissue spectra were smoothed with a window half width of 8 times the pixel width of the CCD divided by 1600 pixels. First-order polynomial fitting was applied within the smoothing window.

2.8.5 System response calibration

When combining data acquired with multiple spectrometers and detectors, it is particularly critical that each system be corrected to account for the wavelength-dependent throughput of the system. Here, system response calibration is performed using a NIST traceable tungsten calibration lamp. System response correction factors were calculated for each system by dividing the known output spectrum of tungsten by the spectrum of tungsten measured with that system. Wavelength-calibrated data are then multiplied by these correction factors to correct for the optical transfer function of the system.

2.9 Additional Device-Dependent Corrections (Standards-based Correction)

When combining data acquired from multiple devices, it is important to use identical components and calibration standards. Given the length of the clinical trial, it may be impossible to replace a failed component with an identical counterpart due to discontinued production or manufacturers ceasing

production. Two additional corrections were necessary to compensate for device differences. The first correction was due to the differences in the components used for power measurements. Using the FastEEM3 light source, power measurements were made at each excitation with the FastEEM3 powermeter and the FastEEM2 powermeter. The powermeter correction factor calculated by dividing the FastEEM3 by the FastEEM2 power measurements was applied to the FastEEM2 data only. The second correction applied was due to system response differences that could not be captured by the tungsten-based correction factors already applied. Only nine tungsten measurements were made with the FastEEM2 device at the beginning of the trial. Before new measurements could be made at the end of the trial, the FastEEM2 detector failed. An additional correction was devised using the positive standards, which were measured later in the trial. Spectra at each excitation for Exalite, coumarin, and rhodamine were concatenated to cover the entire emission range in use. A standards-based system response correction factor was then computed by dividing the FastEEM3 standard spectra by the FastEEM2 standard spectra; the resulting correction factors were applied to the FastEEM2 tissue data.

2.10 Wavelength Calibration Validation

The positive standard of rhodamine was used to validate the accuracy of the wavelength calibration process. The emission maxima of this fluorophore occur at 580 nm independent of excitation wavelength. The peak emission wavelength of the processed rhodamine spectra was extracted at two excitation wavelengths (330 and 420 nm). Statistical analysis of peak locations between measurements at each excitation wavelength provided a tool to assess the consistency and accuracy of the wavelength calibration algorithm.

2.11 Wavelength-Dependent Throughput Calibration Validation

The accuracy of the wavelength-dependent throughput calibration was assessed by comparing the shape of the processed emission spectrum of the positive standard rhodamine to that measured with two other commercially available calibrated spectrometers (SPEX, Fluorolog II Spectrometer, Edison, New Jersey, and Photon Technologies International Quanta Master Model C Spectrofluorometer, Lawrenceville, New Jersey). To assess unexpected drifts in the wavelength-dependent system throughput, tungsten spectra were plotted over time.

2.12 Illumination Energy Calibration Validation

The processed emission spectra of the positive standard rhodamine were assembled into an excitation emission matrix and the excitation spectrum was extracted at 580-nm emission. The shape of the excitation spectrum was compared to that measured with two other commercially, available calibrated spectrometers (SPEX, Fluorolog II Spectrometer, Edison, New Jersey, and Photon Technologies International Quanta Master Model C Spectrofluorometer, Lawrenceville, New Jersey). The shape of the excitation spectrum was assessed over time throughout the trial to identify potential problems with the illumination energy calibration process.

2.13 Engineering Analysis Methods

The statistically derived calibration coefficients and standard spectra collected throughout the trial were analyzed to identify any instrumentation or processing problems, and to formulate a course of action to prevent future occurrences and correct affected measurements, which had already been collected. There is an ethical responsibility to the patients, the funding agencies, and all health care providers involved to use every patient measurement made. Most instrumentation problems can be corrected at the hardware and/or software level. It is important to correct hardware failures to prevent reoccurrence and/or acquisition of potentially unusable data. It is critical to devise algorithms based on physical models of cause and effect that can recover otherwise unusable data.

The raw and processed standards data collected throughout the trial were analyzed in two simple steps. First measurements are compared to ideal values. The spectral response of each positive standard is known. By comparing measured to ideal, automatic detection of deviation from the ideal can be implemented. This also allows for calculation of correction factors for affected measurements, which have already been collected. In the previous step, the symptom of the failure is identified. Correlating this symptom to a physical cause requires domain knowledge, well-documented device event logs, and sometimes experimentation. The extent of the affected measurements can be determined based on the frequency of occurrence of the symptom and validated based on the nature of the cause.

3 Results

3.1 Measured Standards and Their Frequency

Data were acquired from 1295 sites in 442 patients using the FastEEM2 device and from 788 sites in 322 patients using the FastEEM3 device. The frequency with which standards were measured was increased with the FastEEM3 as a direct result of experiences in the FastEEM2 study.

3.2 Raw Data Processing Algorithm

The data processing algorithm was implemented in Matlab (Mathworks, Inc., Natick, Massachusetts) Ver. 6.5 Rel. 13 on the Windows XP operating system. All data were stored and processed on portable fire-wire hard drives. Processing algorithm performance was limited by the file (i/o) access time. Raw data acquired from standards and tissue were processed to yield data in the form of excitation emission matrices (EEMs), encapsulated postscript files of processed EEM images in line plot and contour format, a comprehensive QC file with an image of every processed EEM in line plot and contour format, and processing logs including errors and lists indicating which standards data were used to process each tissue measurement. EEMs contain calibrated fluorescence intensity as a function excitation and emission wavelength.

Figure 5 shows the result of processing data acquired with both systems from the positive standard rhodamine at 420 nm excitation at each step in the processing algorithm. Figure 5(a) illustrates the effect of dark current and background subtraction between FastEEM2 and FastEEM3. Figure 5(b) shows the emission spectra after wavelength calibration, energy normalization, and smoothing with Savitzky-Golay filtering. Fig-

Step 1: Background Subtracted

Step4: λ-Cal,EnergyNorm, Filtered

Step5: After System Response
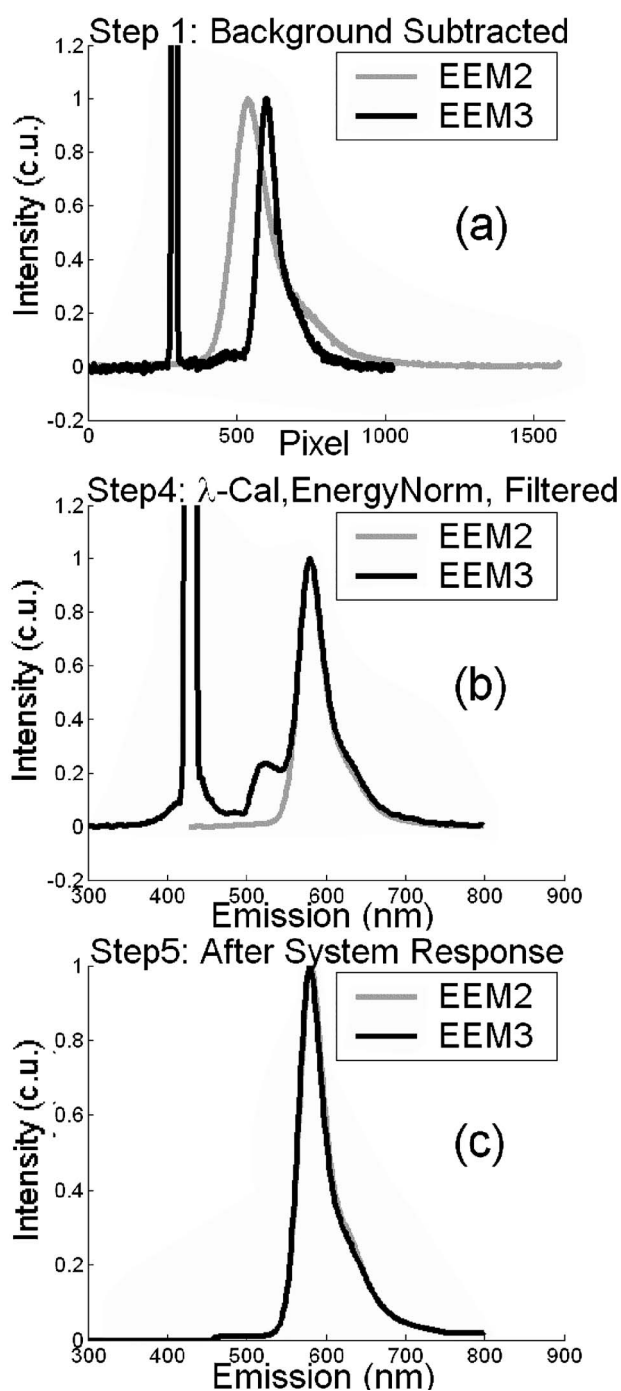
Fig. 5 Normalized rhodamine emission spectra from FastEEM2 and FastEEM3 at 420-nm excitation following various stages of data processing: (a) emission spectra after background subtraction; (b) emission spectra after wavelength calibration, energy normalization, and filtering; (c) final processed emission spectra (c.u. = calibrated units).

ure 5(c) shows the final processed emission spectra. Following processing, emission spectra have the same peak emission wavelength and spectral line shape; the small differences observed with this positive standard provide a measure of the calibration accuracy. We next describe how data processed in this manner were validated.

3.3 Wavelength Calibration Standard Validation

Figure 6 shows the consistency of the wavelength calibration derived from the HgAr measurements using the FastEEM3 device over the lifetime of the trial. Figure 6(a) shows ten consecutive HgAr spectra measured in March to April 2003. The first five mercury peaks in the measured spectrum (365, 404.7, 435.8, 546.1, and 578 nm) are used to derive a linear calibration between pixel and wavelength. Figure 6(b) shows the position of the pixel corresponding to each peak over the duration of the clinical trial. The pixel positions are relatively constant over time, except for three discrete steps, which correspond to changes in grating incidence angle made at the beginning of the study to improve efficiency and to a change in magnification due to camera realignment. Figures 6(c) and 6(d) shows the slope and intercept of the linear calibration coefficients as a function of time throughout the study. The changes in grating angle and magnification changes can be clearly seen.

3.4 Wavelength Calibration Algorithm Validation

Statistical analysis of the emission spectra of rhodamine showed that the peak emission wavelength did not occur at the same value for all excitation wavelengths within an EEM measurement as expected. Furthermore, the value of the peak emission wavelength shifted from one measurement to another [Fig. 7(a)]. Engineering analysis showed that these shifts were associated with thermal changes in the position of the spectrometer grating, which affected the wavelength calibration. These thermal changes shifted the spectrum slightly across the face of the CCD but did not affect the dispersion of the spectrometer; thus, only the intercept of the wavelength calibration and not the slope was affected. A particularly troubling thermal effect was a thermal expansion and contraction that was sourced to the optomechanical mounting of the grating in the Jobin Yvon Triax 320 spectrographs used in the FastEEM3 system. As the system warmed up by about $10\,^{\circ}\text{C}$ during operation, movement of the grating shifted spectral peaks by up to 11 pixels in the wavelength direction [Fig. 7(b)]. Additionally, a vertical shift of up to 6 pixels was also observed. This resulted in wavelength calibration errors of up to 4 nm. Because a HgAr calibration spectrum was obtained only once per day, this standard could not be used to correct for this thermal drift. Rhodamine spectra were collected approximately every 2 h. Monitoring the pixel location of the rhodamine peak provided a measure to correct for this thermal drift. Figure 7(c) shows the rhodamine spectra processed using this correction. Note that the wavelength position of the rhodamine peak is now constant.

3.5 Intensity Calibration Validation

Tungsten standards were used to calculate system response correction factors. Consecutive tungsten spectra were plotted over time to verify the consistency of the wavelength-dependent system response throughout the trial. Figure 8(a) shows resulting rhodamine spectra at 420 nm excitation acquired with FastEEM2 and FastEEM3 after processing to correct for the system response, as well as spectra measured with the calibrated laboratory spectrofluorimeter (Spex Fluorolog II). The spectral shape agrees well in all cases. Figures 8(b) and 8(c) show similar data for two other organic fluorescent
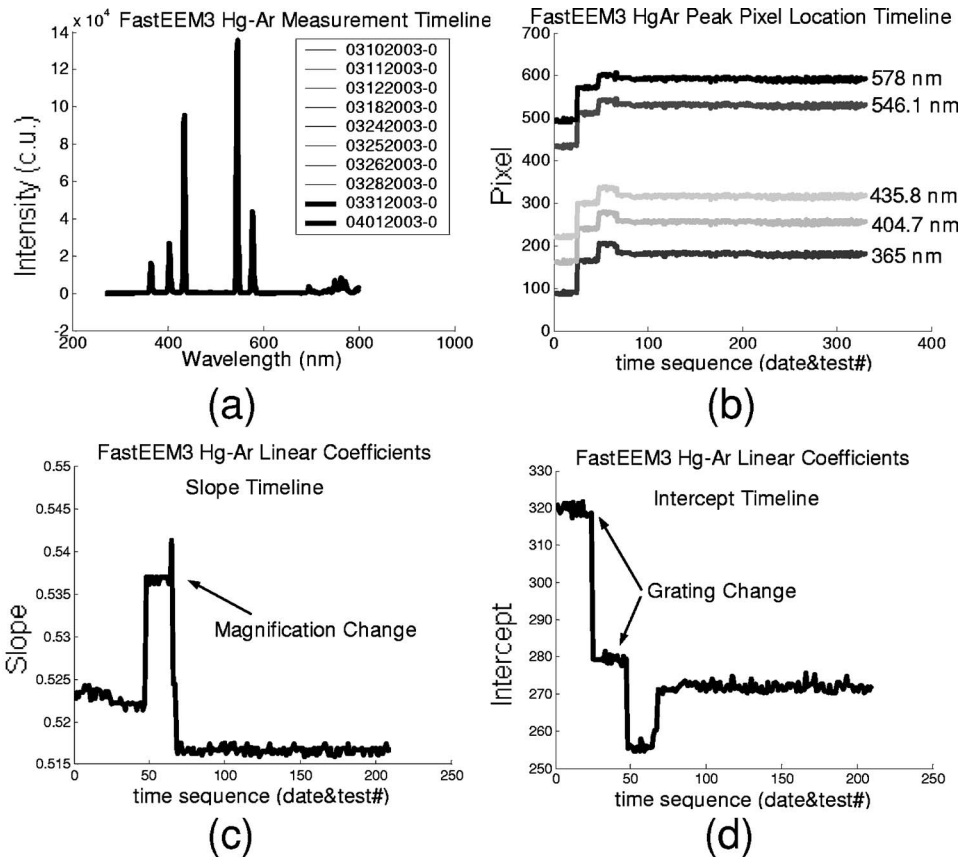
Fig. 6 Graphical tools developed to validate wavelength calibration of the FastEEM systems: (a) 10 consecutive, overlaid spectra of the HgAr lamp acquired with FastEEM3; (b) pixel locations of the five mercury peaks used for wavelength calibration as a function of time throughout the clinical trial; and (c) slope and (d) intercept of the linear relationship between pixel number and wavelength as a function of time throughout the clinical trial.

dyes, Exalite at 320-nm excitation [Fig. 8(b)] and coumarin at 400-nm excitation [Fig. 8(c)]. This confirms the calibration across a broad wavelength range.

3.6 Illumination Energy Calibration

To assess whether the illumination energy calibration was effective, we compared the average excitation spectra of the Rhodamine standard at 580-nm emission measured with both devices [Fig. 9(a)]. Figure 9(b) shows the ratio of the average Rhodamine excitation spectra for both systems. The ratio between FastEEM3 and FastEEM2 is close to 1 over the entire excitation wavelength range.

3.7 Final Validation

The true test of measurement device independence is whether patient data collected from a large and representative group of patients with different devices can be combined without loss of information. In this study, we measured fluorescence EEMs from 764 patients. We compared the average spectra of squamous normal tissue acquired with the two FastEEM systems. Figure 10 shows the average squamous normal tissue spectra for at 330-nm excitation [Fig. 10(a)] and 360-nm excitation [Fig. 10(b)] for both systems. The difference in starting wavelength is due to differences in excitation filter characteristics for both devices. The shape and intensity of both spectra are comparable at each excitation. The spectra shown in [Fig.

10(a)] are an average of acquired fluorescence at 330-nm excitation for each device without normalization. Error bars are very large (not shown) due to biological differences between patients like menopausal state that can amount to more than an order of magnitude.[30] Given the 10-fold patient-to-patient variation in fluorescence intensity, we feel that these results demonstrate excellent agreement.

4 Discussion

The standards protocol and calibration methods developed and tested here yields tissue spectra that can be compared across two entirely different instrument platforms. We investigated a number of additional measurements and data analysis tools to ensure that data were reliable and accurate, including stray light levels, device autofluorescence levels, wavelength calibration accuracy, and temporal reliability of standards measurements.

We did not find stray light to be a great problem during our measurements. One reason for this is that we limit the wavelength of our illumination using a cold mirror that only passes light between about 280 and 800 nm. This means that the light entering the spectrometer tends to be well directed by our optical components. Stray light is more of a problem when using the tungsten calibration source. For the tungsten source we have low intensity in the shorter wavelength and
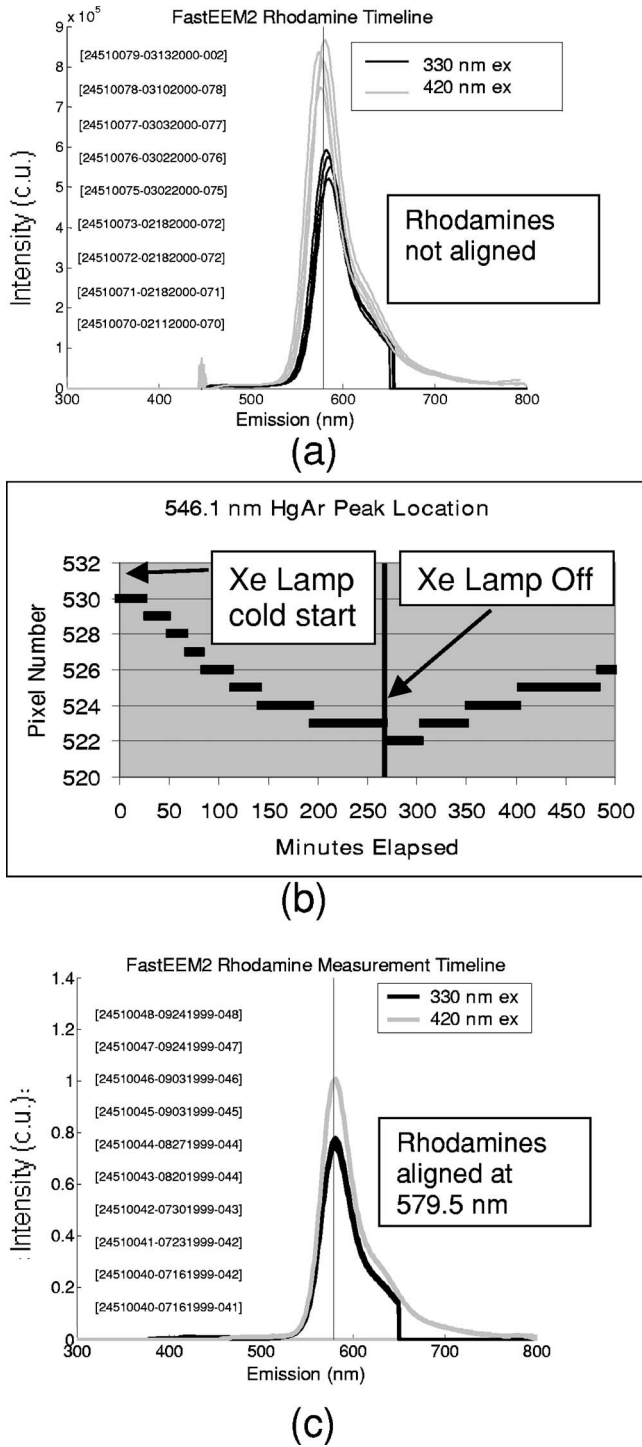
Fig. 7 (a) Ten consecutive emission spectra of rhodamine at 330- and 420-nm excitation processed with the initial wavelength calibration algorithm. Note that the peak emission wavelength shifts by approximately ±3 nm. (b) Pixel location of 546-nm mercury peak as a function of time following initial cold start of FastEEM3. The Xe arc lamp was turned off at 260 min. This effect was traced to thermal expansion and contraction of the grating mount. (c) Ten consecutive emission spectra of rhodamine at 330- and 420-nm excitation processed with the final wavelength calibration algorithm that corrects for temporal variations associated with thermal drift.
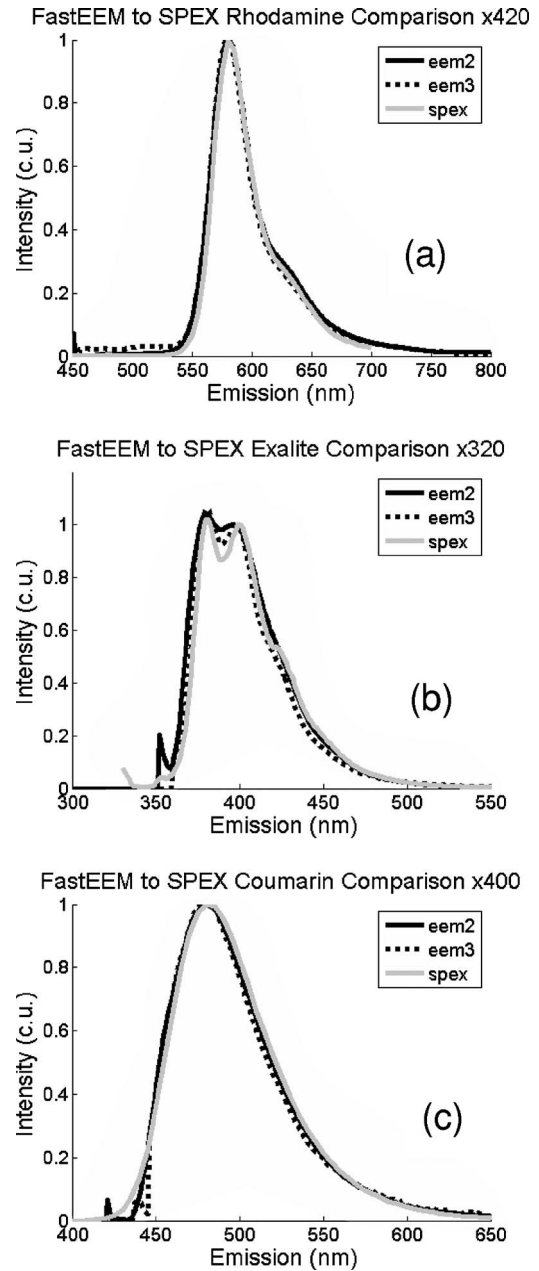
Fig. 8 Comparison of final processed emission spectra obtained with the FastEEM2, FastEEM3, and reference spectrometer for organic fluorescent dyes: (a) rhodamine at 420-nm excitation, (b) Exalite at 320-nm excitation, and (c) coumarin at 400-nm excitation.

high intensity in longer wavelength as well as a significant IR component that our optics are not optimized for. In this case, we saw some stray light and out-of-focus light at the bright longer wavelength end of the spectrum. It is difficult to disentangle the relative contribution of focus and stray light in this situation. Our approach was to take advantage of the fact that we used an array detector. We were able to examine the detector in areas between the signals from detection fibers as well as areas well away from the detection fibers to assess the relative contribution of stray light. We concluded that the contribution from stray light was relatively low with respect to the signal (390-nm excitation ratio of stray light to signal at
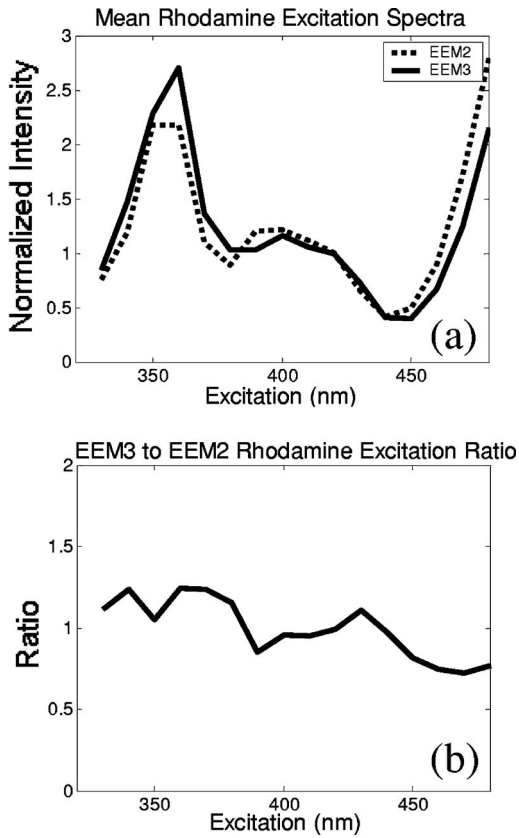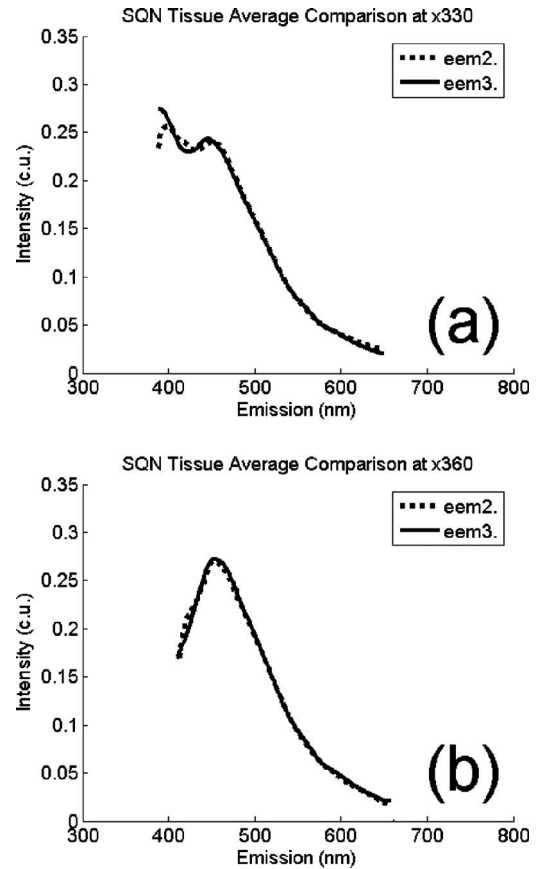
Fig. 9 (a) Average rhodamine excitation spectra at 580-nm emission measured with FastEEM2 and FastEEM3, where error bars represent one standard deviation, and (b) ratio of average rhodamine excitation spectra at 580-nm emission measured with FastEEM3 to that measured with FastEEM2.

Fig. 10 Average emission spectra of squamous normal cervical tissue measured *in vivo* from 435 sites in 206 patients with FastEEM2 and 167 sites in 98 patients with FastEEM3 at (a) 330- and (b) 360-nm excitation.

640-nm emission is 0.003915, at 800-nm emission is 0.00739) even for the worst-case situation.

Wavelength calibration of our spectrometer was relatively simple because of the long focal length and the relatively on-axis design of the system. We compared various orders in fitting the five HgAr peaks used for calibration and found that a simple linear fit was adequate to characterize the system. We also compared fitting with shorter and longer wavelength emission lines from the HgAr lamp and found that this gave no advantage over the linear fit of five peaks, so we elected to use the simpler method. This has been supported in our analysis of measurements over the full timeline of our study. During the FastEEM2 study, we used Hg lines from fluorescent room lights for wavelength calibration. This was later abandoned because fluorophors of room lights from different manufacturers have different spectral shapes that can overlap mercury lines and make detection difficult. It was found not to be a consistent, reliable source for detecting wavelength calibration peaks.

When we began our study, we did not expect the variation in wavelength accuracy of our system that we ultimately observed. We felt that one HgAr calibration per day would be sufficient. Our daily calibration was very reproducible; unfortunately our measurements seemed to show variation. Some of our patient's spectra also included porphyrin peaks. These are well known narrow peaks at 635 and 704 nm. We noticed variations in these peaks in our samples. We also noticed that these variations correlated with variations in rhodamine peak position, even though the rhodamine peak is not as narrow as the porphyrin peak. Since we had variation in peak position for this standard, which we measured regularly every 2 h, and this was confirmed on tissue measurements through the porphyrin peak, we ultimately performed an engineering analysis and determined we had a thermal drift problem with one of our ISA spectrometers. As a result of this we increased the frequency of measurements of our wavelength standard. We further compared the position of some of the sharper peaks of the xenon arc lamp used in our white light reflectance measurements, and found that they correlated as well. We purposely included a wide range of standards expecting some redundancy. We found that we used the standards in ways we had not anticipated when we began the study.

To validate the accuracy and precision of the final wavelength calibration procedure, we examined the temporal dependence of peaks in two of our positive standards. Figure 11 shows the precision of the wavelength calibration, plotting the coumarin peak emission wavelength at 400-nm excitation over the timeline of the study. The accuracy of the wavelength calibration was evaluated based on HgAr emission line at 696.543 nm. This emission line is not one of the five peaks used in the linear fit for calibration parameters. The calibrated
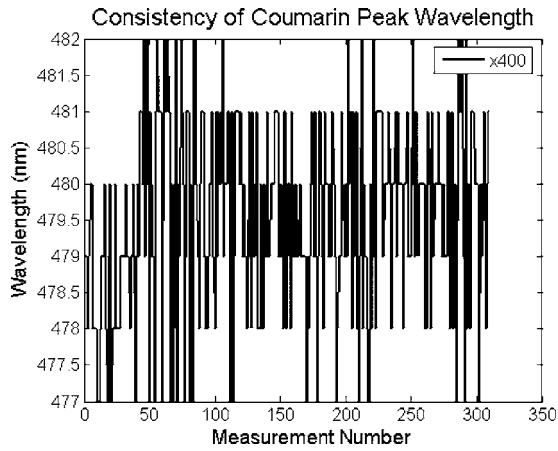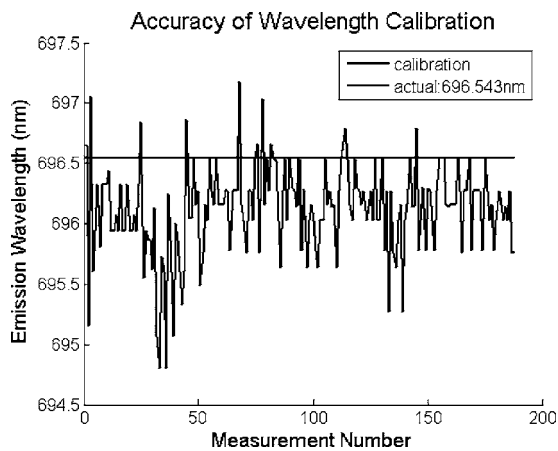
Fig. 11 Wavelength calibration validation by comparing the consistency over time of coumarin peak emission wavelength at 400-nm excitation.
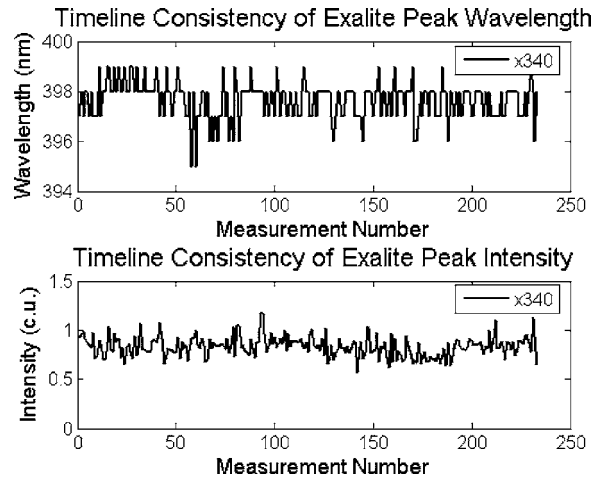

Fig. 13 Temporal consistency of positive standard Exalite demonstrated using peak location and peak fluorescence intensity at 340-nm excitation.

value of this emission line over the lifetime of the study was 696.12±0.38 nm (Fig. 12). The difference between actual and the calibrated is less than the 5-nm spectral resolution of the instrument. Future studies will focus on wavelength calibration standards accuracy, repeatability, and the effect of random drifts on data analysis.

The positive standards were examined to determine the reproducibility of both wavelength and intensity information. Calibration standards showed good consistency in wavelength and fluorescence intensity over the timeline of the study. Figure 13 demonstrates the timeline consistency of peak wavelength and peak intensity for the positive standard Exalite. The 2-nm variation in Exalite peak wavelength at 340-nm excitation was less than the 5-nm spectral resolution.

The negative standards were examined to determine the consistency of background autofluorescence levels. Figure 14 demonstrates the temporal consistency of fluorescence intensity for the frosted cuvette negative standard. Timeline consistency of water and frosted cuvette fluorescence intensity were monitored to detect probe autofluorescence.
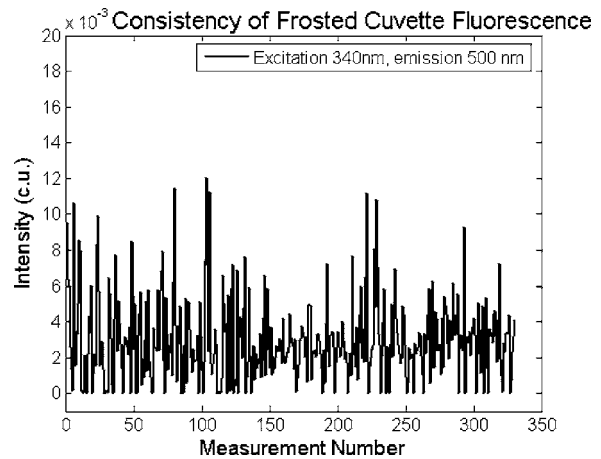
When choosing the frequency for measurements of standards we tried to balance our desire to have more frequent measurements with the constraints imposed by making measurements in a clinical setting where the schedules of patients and clinical staff are subject to exigencies. We wished to use the statistical power of the large number of patients and standards measurements in this study to develop an evidence-based rationale for type and frequency of calibration and performance verification standards measurements for optical measurements, as well as the feasibility of making such measurements in clinical settings. We measured both performance verification standards such as fluorescent laser dyes and calibration standards such as mercury-argon lamps and quartz tungsten halogen lamps with high frequency, but it was not our intent to adjust calibration so frequently. Rather these standards were to be used as performance verification standards, to see how well the calibration was holding up during measurements.

Fig. 12 Accuracy of wavelength calibration based on difference between the argon emission line at 696.543 nm and its calibrated mean value of 696.12±0.38 nm.

Fig. 14 Temporal consistency of negative frosted cuvette standard demonstrated using fluorescence intensity at 500-nm emission for 340-nm excitation.

It is difficult to arrive at a consensus for standardization of optical devices in calibration, validation, and performance. Our intent is not to establish what standards to use and how often to measure them, but to define a strategy for decision making on instrument calibration. Our suggested strategy is to determine dimensionality of the data and have calibration and performance verification for each dimension over time so that the investigators can demonstrate the validity of the measurements. We believe that the calibration strategy should be able to be defined to a great degree at the proposal stage for a clinical research project and a statement should be included as part of the clinical protocol. Discussion sessions at conferences should be the subject of this topic. Other research communities have been able to arrive at standardization guidelines successfully. The Minimal Information to Annotate a Microarray Experiment[31] (MIAME) has established the minimum information required to unambiguously interpret microarray data and to subsequently enable its independent verification and reproduction. Similarly, the American College of Radiology developed and published Mammography Quality Control Manuals[32] in 1990 and the Breast Imaging Reporting and Data System[33] (BIRADS) in 1993. BI-RADS is a quality assurance tool designed to standardize mammography reporting, reduce confusion in breast imaging interpretations, and facilitate outcome monitoring. The key elements are a lexicon of standardized terminology, a reporting organization and assessment structure, a coding system and a data collection structure. BI-RADS® is the product of a collaborative effort between members of various committees of the American College of Radiology with cooperation from the National Cancer Institute (NCI), the Centers for Disease Control and Prevention, the Food and Drug Administration, the American Medical Association, the American College of Surgeons, and the College of American Pathologists. Standardization in optical devices is needed to ensure portability and quality of data.

5 Conclusions

Clinical trials to assess technical feasibility of *in vivo* diagnosis using optical technologies face many instrumentation events, which can result in costly delays and unusable data. A plan to validate instrument performance can provide methods to correct for these measurement system issues and recover otherwise unusable data. It is important for clinical studies to incorporate a large number of standards, which are measured frequently and that adequately capture the range of optical parameters that can vary. Additionally these standards should be measured over long periods in actual clinical settings prior to commencement of clinical trials to anticipate environmental issues. A comprehensive set of standards supports quality control statistics that maximize consistency, completeness, and quality of the data. Furthermore, these standards provide viable options to protect study results from adverse events. In reducing these risks, the delivery and quality of clinical research is enhanced, by providing the ability to develop device independent algorithms. The biomedical optics community should adopt a consensus set of positive and negative performance standards to facilitate evaluation and comparison of data collected in different laboratories with different instruments.

References

1. N. Ramanujam, M. F. Mitchell, A. Mahadevan, S. Thomsen, E. Silva, and R. Richards-Kortum, "Fluorescence spectroscopy: a diagnostic tool for cervical intraepithelial neoplasia (CIN)," *Gynecol. Oncol.* **52**, 31–38 (1994).
2. G. M. Wagnieres, W. M. Star, and B. C. Wilson, "In vivo fluorescence spectroscopy and imaging for oncological applications," *Photochem. Photobiol.* **68**, 603–632 (1998).
3. N. Ramanujam, M. Follen-Mitchell, A. Mahadevan, S. Thomsen, A. Malpica, T. Wright, et al., "Development of a multivariate statistical algorithm to analyze human cervical tissue fluorescence spectra acquired *in vivo*," *Lasers Surg. Med.* **19**, 46–62 (1996).
4. R. J. Nordstrum, L. Burke, J. M. Niloff, and J. M. Myrtle, "Identification of cervical intraepithelial neoplasia (CIN) using UV-excited fluorescence and diffuse-reflectance tissue spectroscopy," *Lasers Surg. Med.* **29**, 118–127 (2001).
5. S. K. Chang, M. Y. Dawood, G. Staerkel, U. Utzinger, E. N. Atkinson, R. Richards-Kortum, and M. Follen, "Fluorescence spectroscopy for cervical precancer detection: Is there variance across the menstrual cycle?" *J. Biomed. Opt.* **7**(4), 595–602 (2002).
6. S. K. Chang, M. Follen, A. Malpica, U. Utzinger, G. Staerkel, D. Cox, E. N. Atkinson, C. MacAulay, and R. Richards-Kortum, "Optimal excitation wavelengths for discrimination of cervical neoplasia," *IEEE Trans. Biomed. Eng.* **49**, 1102–1111 (2002).
7. I. Georgakoudi, E. E. Sheets, M. G. Müller, V. Backman, C. P. Crum, K. Badizadegan, R. R. Dasari, and M. S. Feld, "Trimodal spectroscopy for the detection and characterization of cervical precancers in vivo," *Am. J. Obstet. Gynecol.* **186**, 374–382 (2002).
8. W. K. Huh, R. M. Cestero, F. A. Garcia, M. A. Gold, R. S. Guido, K. McIntyre-Seltman, D. M. Harper, L. Burke, S. T. Sum, R. F. Flewelling, and R. D. Alvarez, "Optical detection of high-grade cervical intraepithelial neoplasia in vivo: results of a 604-patient study," *Am. J. Obstet. Gynecol.* **190**, 1249–1257 (2004).
9. I. J. Bigio, T. R. Loree, and J. Mourant, "Spectroscopic diagnosis of bladder cancer with elastic light scattering," *Lasers Surg. Med.* **16**, 350–357 (1995).
10. B. Chance, "The use of intrinsic fluorescent signals for characterizing tissue metabolic states in health and disease," *Proc. SPIE* **2679**, 2–7 (1996).
11. J. R. Mourant, J. P. Freyer, A. H. Hielscher, A. A. Eick, D. Shen, and T. M. Johnson, "Mechanisms of light scattering from biological cells relevant to noninvasive optical-tissue diagnostics," *Appl. Opt.* **37**, 3586–3593 (1998).
12. G. Zonios, L. T. Perelman, V. Backman, R. Manoharan, M. Fitzmaurice, J. Van Dam, and M. S. Feld, "Diffuse reflectance spectroscopy of human adenomatous colon polyps in vivo," *Appl. Opt.* **38**, 6628–6637 (1999).
13. T. Collier, D. Arifler, A. Malpica, M. Follen, and R. Richards-Kortum, "Determination of epithelial tissue scattering coefficient using confocal microscopy," *IEEE J. Quantum Electron.* **9**(2), 307–313 (2003).
14. I. Pavlova, K. Sokolov, R. Drezek, A. Malpica, M. Follen, and R. Richards-Kortum, "Microanatomical and biochemical origins of normal and precancerous cervical autofluorescence using laser-scanning fluorescence confocal microscopy," *Photochem. Photobiol.* **77**, 550–555 (2003).
15. B. Littenberg, "Technology assessment in medicine," *Acad. Med.* **67**(7), 424–428 (1992).
16. A. F. Zuluaga, U. Utzinger, A. Durkin, H. Fuchs, A. Gillenwater, R. Jacob, B. Kemp, J. Fan, and R. Richards-Kortum, "Fluorescence excitation emission matrices of human tissue: a system for in vivo mea-

surement an method of data analysis," *Appl. Spectrosc.* **53**, 302–311 (1999).

17. J. F. Brennan, G. I. Zonios, T. D. Wang, R. P. Rava, G. B. Hayes, R. R. Dasari, and M. S. Feld, "Portable laser spectrofluorimeter system for in vivo human tissue fluorescence studies," *Appl. Spectrosc.* **47**, 2081–2086 (1993).

18. R. M. Cothren, G. B. Hayesm, J. R. Kramer, B. A. Sacks, C. Kittrell, and M. S. Feld, "A multifiber catheter with an optical shield for last angiosurgery," *Lasers Life Sci.* **1**, 1–12 (1986).

19. R. A. Zangaro, L. Silveira, R. Manoharan, G. Zonios, I. Itskan, R. Dasari, J. vanDam, and M. S. Feld, "Rapid multiexcitation fluorescence spectroscopy system for in-vivo tissue diagnosis," *Appl. Opt.* **35**, 5211–5219 (1996).

20. I. Georgakoudi, B. C. Jacobson, J. Van Dam, V. Backman, M. B. Wallace, M. G. Muller, Q. Zhang, K. Badizadegan, D. Sun, G. A. Thomas, L. T. Perelman, and M. S. Feld, "Fluorescence, reflectance, and light-scattering spectroscopy for evaluating dysplasia in patients with Barrett's esophagus," *Gastroenterology* **120**(7), 1620–1629 (2001).

21. N. Ramanujam, M. Follen Mitchell, A. Mahadevan-Jansen, S. L. Thomsen, G. Stearkel, A. Malpica, et al., "Cervical precancer detection using a multivariate statistical algorithm based on laser-induced fluorescence spectra at multiple excitation wavelengths," *Photochem. Photobiol.* **64**, 720–735 (1996).

22. M. G. Muller, I. Georgakoudi, Q. Zhang, J. Wu, and M. S. Feld, "Intrinsic fluorescence spectroscopy in turbid media: disentangling effects of scattering and absorption," *Appl. Opt.* **40**, 4633–4646 (2001).

23. J. C. Finlay, D. L. Conover, E. L. Hull, and T. H. Foster, "Porphyrin bleaching and PDT-induced spectral changes are irradiance dependent in ALA-sensitized normal rat skin in vivo," *Photochem. Photobiol.* **73**, 54–63 (2001).

24. T. J. Pfefer, D. Sharma, A. Agrawal, and L. S. Matchette, "Evaluation of fiber-optic based system for optical measurement of highly attenu-ating turbid media," *Proc. SPIE* **5691**, 163–171 (Mar. 2005).

25. Q. Liu and N. Ramanujam, "Experimental proof of the feasibility of using an angled fiber-optic probe for depth-sensitive fluorescence spectroscopy of turbid media," *Opt. Lett.* **29**(17), 2034–2036 (2004).

26. J. C. Finlay and T. H. Foster, "Recovery of hemoglobin oxygen saturation and intrinsic fluorescence with forward-adjoint model," *Appl. Opt.* **44**(10), 1917–1933 (2005).

27. A. K. Gaigalas, L. Li, O. Henderson, R. Vogt, J. Barr, G. Marti, J. Weaver, and A. Schwartz, "The development of fluorescence intensity standards," *J. Res. Natl. Inst. Stand. Technol.* **106**(2), 381–389 (2001).

28. B. C. Wilson and S. L. Jacques, "Optical reflectance and transmittance of tissues; principles and applications," *IEEE J. Quantum Electron.* **26**, 2186–2199 (1990).

29. S. J. Orfanidis, *Introduction to Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ (1996).

30. E. M. Gill, A. Malpica, R. Alford, A. Nath, M. Follen, R. Richards-Kortum, and N. Ramanujam, "Relationship between collagen autofluorescence of the human cervix and menopausal status," *Photochem. Photobiol.* **77**(6), 653–658 (2003).

31. A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, T. Gaasterland, P. Glenisson, C. P. Holstege, I. F. Kim, V. Markowitz, J. C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron, "Minimum information about a microarray experiment (MIAME)—toward standards for microarray data," *Nat. Genet.* **29**(4), 365–371 (2001).

32. R. E. Hendrick, L. W. Bassett, G. D. Dodd, et al., *Mammography Quality Control: Radiologist's Manual, Radiologic Technologist's Manual, Medical Physicist's Manual*, 4th ed., American College of Radiology, Reston, VA (1999).

33. D. B. Kopans, C. J. D'Orsi, D. D. Adler, et al., *Breast Imaging Reporting and Data System (BIRADS)*, 2nd ed., American College of Radiology, Reston, VA (1998).