Cellular Oncology 26 (2004) 101–117
IOS Press

Quality assurance system using statistical process control: An implementation for image cytometry [1]

David Chiu [a,f], Martial Guillaud [a], Dennis Cox [b], Michele Follen [c,d,e,*] and Calum MacAulay [a]
[a] *Department of Cancer Imaging, British Columbia Cancer Agency, Vancouver, BC, Canada*
[b] *Department of Statistics, Rice University, Houston, TX, USA*
[c] *Department of Gynecology, The University of Texas M.D. Anderson Cancer Center, Houston, TX, USA*
[d] *Biomedical Engineering Center, The University of Texas M.D. Anderson Cancer Center, Houston, TX, USA*
[e] *The Department of Gynecology, Obstetrics and Reproductive Sciences, The University of Texas Health Science Center at Houston, Houston, TX, USA*
[f] *Perceptronix Medical Inc., Vancouver, BC, Canada*

Received 13 October 2003
Accepted 4 March 2004

Abstract. *Aims:* Optical technologies have shown some promise for improving the care of cervical neoplasia. We are currently evaluating fluorescence and reflectance spectroscopy and quantitative cyto-histopathology for cervical neoplasia screening and diagnosis. Here we describe the establishment and application of a quality assurance (QA) system for detecting system malfunctions and assessing the comparability of four image cytometers used in a multicenter clinical trial.

Methods: Our QA system involves three levels of evaluation based on the periodicity and complexity of the measurements. We implemented our QA system at three image cytometers at the British Columbia Cancer Agency and one at M.D. Anderson Cancer Center. The measurements or tasks were performed daily, monthly, and semi-annually. The current and voltage of the lamp, the calibration image characteristics, and the room temperature were checked daily. Long-term stability over time, short-term variability over time, and spatial response field uniformity were evaluated monthly. Camera linearity was measured semi-annually. Control charts based on statistical process control techniques were used to detect when the system did not perform optimally.

Results: Daily measurements have shown good consistency in room temperature, lamp and calibration behaviour. Monthly measurements have shown small coefficients of variation between and within the four devices. There have been greater differences between sessions than within sessions. Comparability among the four systems is reasonably good. Semi-annual measurements have shown stable camera linearity. QA events were detected using the QA system. Multiple examples of event detection leading to correction of system malfunction are described in this report.

Conclusions: QA programs are critical for ensuring data integrity and therefore for the conduct of multicenter clinical trials.

Keywords: Image cytometry, quality assurance, quality control, statistical process control, histopathology, quantitative, cytology, quantitative, technology assessment, cervical intraepithelial neoplasia, lesions, squamous intraepithelial

1. Introduction

The term "technology assessment" refers to the systematic evaluation of established or emerging technologies. Among the models of technology assessment that have been described, the Littenberg model [8,9]

[1] Contract grant sponsor: National Cancer Institute, Program Project Grant 3PO1-CA82710-04.
[*] Corresponding author: Dr Michele Follen, M.D. Anderson Cancer Center, 1515 Holcombe Blvd., Unit 193, Houston, TX 77030, USA. Tel.: +1 713 745 2564; Fax: +1 713 792 4856; E-mail: mfollen@mdanderson.org.

is well suited for the evaluation of emerging medical technologies. The model evaluates biologic plausibility, technical feasibility, clinical effectiveness or intermediate outcomes, patient outcomes, and societal outcomes. Biologic plausibility asks questions whether our current understanding of biology and pathology of the disease support the technology. Technical feasibility questions whether, at the present level of assessment, we can safely deliver the target technology to the patient. Clinical effectiveness or intermediate effect assess the effectiveness of the technology in a relevant population. Patient outcomes assess whether the technology improves the patient's health. Societal outcomes assess the cost and ethical implications of the technology.

Our current effort is to evaluate several optical technologies following the paradigm of Littenberg, including fluorescence and reflectance spectroscopy and quantitative cyto-histopathology as emerging technologies for the diagnosis of cervical neoplasia. In the current phase of our program project, quantitative cyto-histopathology is being evaluated for technical feasibility. One aspect of technical feasibility is showing that several image cytometers can perform consistently and reliably in different clinical settings for prolonged periods of time. A separate publication describes the comparability of the trained personnel who operate the equipment [4].

In this study, we test the hypothesis that measurements from the image cytometers made for quality assurance with three devices in the British Columbia Cancer Agency (BCCA) and one device at The University of Texas M.D. Anderson Cancer Center are comparable and reliable. These devices collect data from both cytological and histological cervical specimens collected in the program project. We have designed clinical trials for 800 diagnostic patients and 1000 screening patients; thus far, 1200 Papanicolaou smears and 2300 cervical biopsies have been reviewed by study pathologists. As part of the ongoing effort to further understand the biology of cervical neoplasia, 1200 histological and 1000 cytological specimens have undergone quantitative assessment. It is essential that a Quality Assurance (QA) system is in place to ensure a high level of reliability of the data collected. In this manuscript, we describe the establishment and application of our QA system, the comparability of the four image cytometers used in the trial, and how our QA system made possible adverse event detection, problem identification, and problem solving.

2. Materials and methods

Our project is utilizes three image cytometry devices at British Columbia Cancer Agency and one at M.D. Anderson Cancer Center. A critical part of such a multicenter clinical trial is the establishment and application of a QA system to ensure proper system performance and comparability across systems and sites.

Drawing on statistical process control (SPC), a technique used in industry [2,13], we identified the critical system parameters and determined the frequency with which they needed to be measured. Following, in part, the 1997 European Society for Analytical Cellular Pathology (ESACP) Consensus Report Part II [7], we measured lamp performance, long-term repeatability, short-term repeatability, field uniformity, and system linearity. Our QA system involves three levels of evaluation based on the periodicity and complexity of the measurements. The measurements or tasks were performed daily, monthly, and semi-annually. Daily, the current and voltage of the lamp, the calibration image characteristics, and the room temperature were checked; monthly, the long-term stability over time, short-term variability over time, and spatial response field uniformity were evaluated; and semi-annually, camera linearity was measured. Control charts based on SPC techniques were used to detect when the system did not perform correctly.

We identified the possible failure modes of the critical device functions and components and developed procedures to make measurements that could detect the failure modes, as suggested by our group and others [1,5–7,11,14,15]. Table 1 shows how the various measurements are related to specific failure modes. Figure 1 shows a simplified diagram of the image cytometer, its critical components and functions, and their possible failure modes.

Daily, the operating parameters of the cytometer are recorded and compared as part of system set-up. The parameters included: (1) visual check of the interference filter, numerical aperture of the condenser, and setting of the field iris; (2) software version control; (3) calibration image characteristics (mean, standard deviation, intensity histogram symmetry, and unimodality); and (4) illumination power (voltage and current). Figure 2 shows the daily QA checklist used. The Fig. 2 checklist is used to monitor whether the device is configured correctly. The results are recorded by the cytotechnicians and compared to their historical records. The daily comparisons are based on "rules of

Table 1
Correspondence between QA measurements and failure-mode detection

Periodicity	QA measurements	Possible failure modes
Daily	Lamp voltage and current (power) Calibration, background image standard deviation Shape of the intensity histogram of the background image (symmetry, unimodality)	• Improper cytometer setup • Lamp deterioration – illumination fluctuation • Dust accumulation on any lens surface • Gross oil accumulation on the objective or condenser lens • Change in camera photo-response
Monthly	Mean of the 3 mean values of *IOD* and *AREA* of the 42 circles collected by the 3 repeated acquisitions Range of the 3 mean values of *IOD* and *AREA* of the 42 circles collected by the 3 repeated acquisitions CV's of the *IOD* and *AREA* of the 42 objects from each acquisition	• Fine deposits on the objective and/or condenser lens • Focus drive deterioration – misstepping • Lamp deterioration – illumination fluctuation • CCD partial failure – fluctuation • CCD partial failure – uneven response
	Measurements of the 20% transmissions vs. 60% transmission	• Shift of the camera response offset • Shift of the camera response gain • Camera photo-response change, nonlinearity
Semi-annual	Photo-response curve	• CCD partial failure • Camera photo-response change • Camera photo-response non-linearity

thumb" rather than formal statistics to permit the user to quickly detect gross changes in the system.

The monthly QA session involves measuring the integrated optical density (IOD) and area of the fields of circles on a standardized PRESS-PRO21 slide, which is described in detail below and illustrated in Appendix. We selected optical transmissions of 20% and 60% to cover the range of the photometric response of the cytometers. Three repeated measurements of 42 circles from each of the two transmissions are made. In the QA session, first, the daily procedure is performed to ensure the cytometer is correctly configured. If problems are discovered during the daily procedure, the system is adjusted before proceeding to the monthly measurements. Then, the set of fields of circles of 20% transmission are positioned under the optical path of the cytometer. The scene is manually focused. A thresholding algorithm is used to separate the fields from the background based on pixel intensity; 42 fields are manually selected for analysis. Autofocusing and edge relocation algorithms are applied to each field to precisely and automatically place the edge of the object along the contour of highest local gray-level gradient [13]. The digital gray-level images of these fields are stored in the gallery. This process is repeated three times. To eliminate sampling variation, the same 42 fields at the two transmissions are imaged all three times. After the three repeated acquisitions of the 20% targets are complete, the process is repeated for the 60% targets. The monthly QA session is summarized in Fig. 3.

Semi-annually, a full linearity assessment of the system is performed by measuring the photo-response (see Fig. 4 for details of how this is performed and Fig. 6 for results). This is implemented by measuring the camera response at different exposure times. A highly accurate digital clock controls the exposure time in the digital cameras. If the camera is linear, the measured intensity should be proportional to the exposure time. The exposure feature of the Xillix MicroImager 1400 was used to measure its linearity.

A presentation was given to the staff to inform them of the general ideas and goals of the QA system. They were then trained to perform the QA system tasks. In total, 10 people were trained.

2.1. Slide description

Following the results of the PRESS Project [11,12], the ESACP created a standardized reference slide for quantitative evaluation and calibration of image cytometers. This reference slide (the PRESS-PRO21slide) contains three main patterned strips: (1) dark circles and squares on a bright background, (2) bright circles and squares on a dark background, and (3) large transmission windows. Within all three strips, there are 11 transmission steps (100%, 90%, 80%, 70%, 60%, 50%, 40%, 30%, 20%, 10%, and 5%). Refer to Appendix for a detailed description. In our quality assurance system, we use only the dark circles on a bright background as they mimic quantitatively absorption stained cell nuclei.

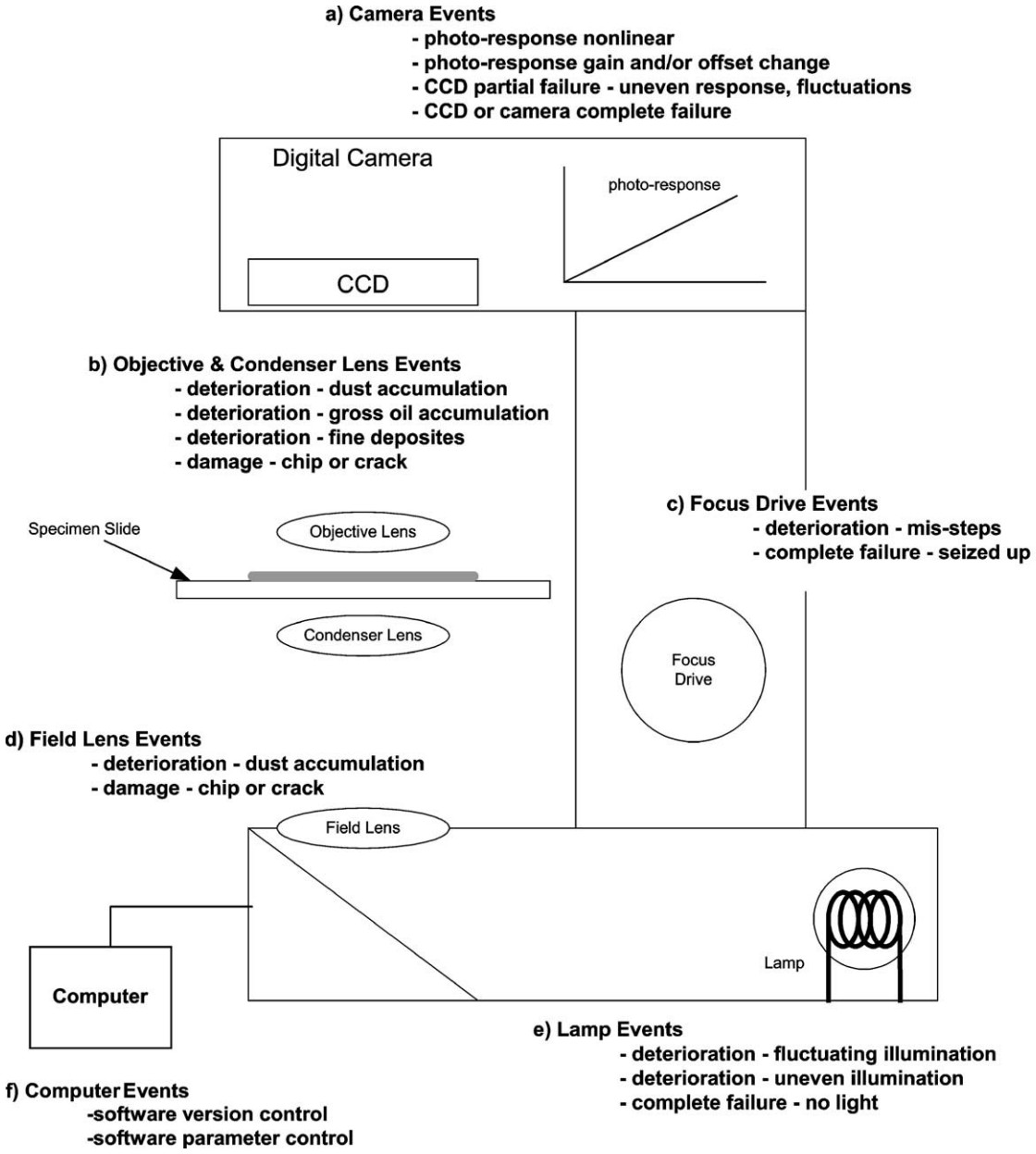


Fig. 1. The possible failure mode of the image cytometer. Listed are the most probable failures that can occur to the image cytometer.

To minimize the monthly workload, we decided to use only two transmissions (20% and 60%), which are relatively far apart and provide IOD's similar to those measured in nuclei of cells. The IOD of the 20% circles are similar to abnormal nuclei. The IOD of the 60% circles are close to the normal nuclei.

Based on previous experience with charged coupled device cameras and linearity measurements, we have determined that the cameras maintain linearity, but their gain and offset can vary with time. Therefore, by measuring two transmissions, it is also possible to differentiate between offset and gain failure. This minimizes the workload by not requiring measurements over the entire photo-response.

Because the four cytometers are located in two different cities (Vancouver and Houston), two different PRESS-PRO21 slides, one at each site, are used. A third slide was also used at BCCA after the origi-

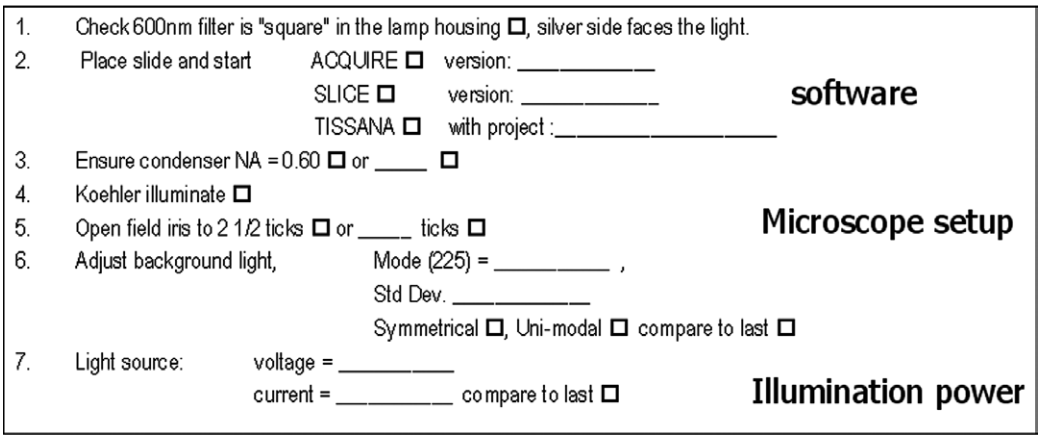


Fig. 2. Daily QA: sample daily system set-up checklist. This checklist is used to ensure that the software, microscope, and illumination power are set-up correctly.

nal was damaged. There are important issues related to slide reproducibility that will be discussed in the Results section.

2.2. Image analysis

Image analysis is performed using a modified version of the CytoSavant automated quantitative system (Oncometrics Imaging Corp., Richmond, British Columbia, Canada). This system uses a 12-bit double-correlated sampling MicroImager 1400 digital camera (pixels 6.8 μm^2) placed in the primary image plane of the microscope. This software was designed for semiautomatic cellular analysis. The system was designed to measure thionin–Feulgen-stained nuclei with a monochromatic light at a wavelength of 600 nm ($\pm$5 nm) using a $\times$20 0.75-NA Plan Apo objective lens (Omega Optical; Brattleboro, VT) [4,3,10].

2.3. Feature calculation

One hundred and three features are computed from the digitized images of each selected circle [3,13,14]. The IOD and area, which are two of these features, are used for the QA system.

2.4. Statistical analyses

The following description will focus on the analysis of the monthly data because it contains the most interesting information. At each transmission (20% and 60%) from each of the three acquisitions, features computed on a circle-by-circle basis are summarized by means and standard deviations to create the measurement-level features.

$$M\_OD_{42} = \frac{1}{42} \times \sum_{1}^{42} IOD,$$

$$S\_OD_{42} = \sqrt{\frac{1}{41} \times \sum_{1}^{42} (M\_OD_{42} - IOD)^2},$$

$$CV\_OD_{42} = \frac{S\_OD_{42}}{M\_OD_{42}} \times 100\%,$$

$$M\_Area_{42} = \frac{1}{42} \times \sum_{1}^{42} AREA,$$

$$S\_Area_{42} = \sqrt{\frac{1}{41} \times \sum_{1}^{42} (M\_Area_{42} - AREA)^2},$$

$$CV\_Area_{42} = \frac{S\_Area_{42}}{M\_Area_{42}} \times 100\%.$$

From these measurement-level features, means and standard deviations across the three repeat acquisitions are calculated to give session-level features.

$$M\_OD_{R3} = \frac{1}{3} \times \sum_{1}^{3} M\_OD_{42},$$

$$S\_OD_{R3} = \sqrt{\frac{1}{2} \times \sum_{1}^{3} (M\_OD_{R3} - M\_OD_{42})^2},$$

$$M\_Area_{R3} = \frac{1}{3} \times \sum_{1}^{3} M\_Area_{42},$$

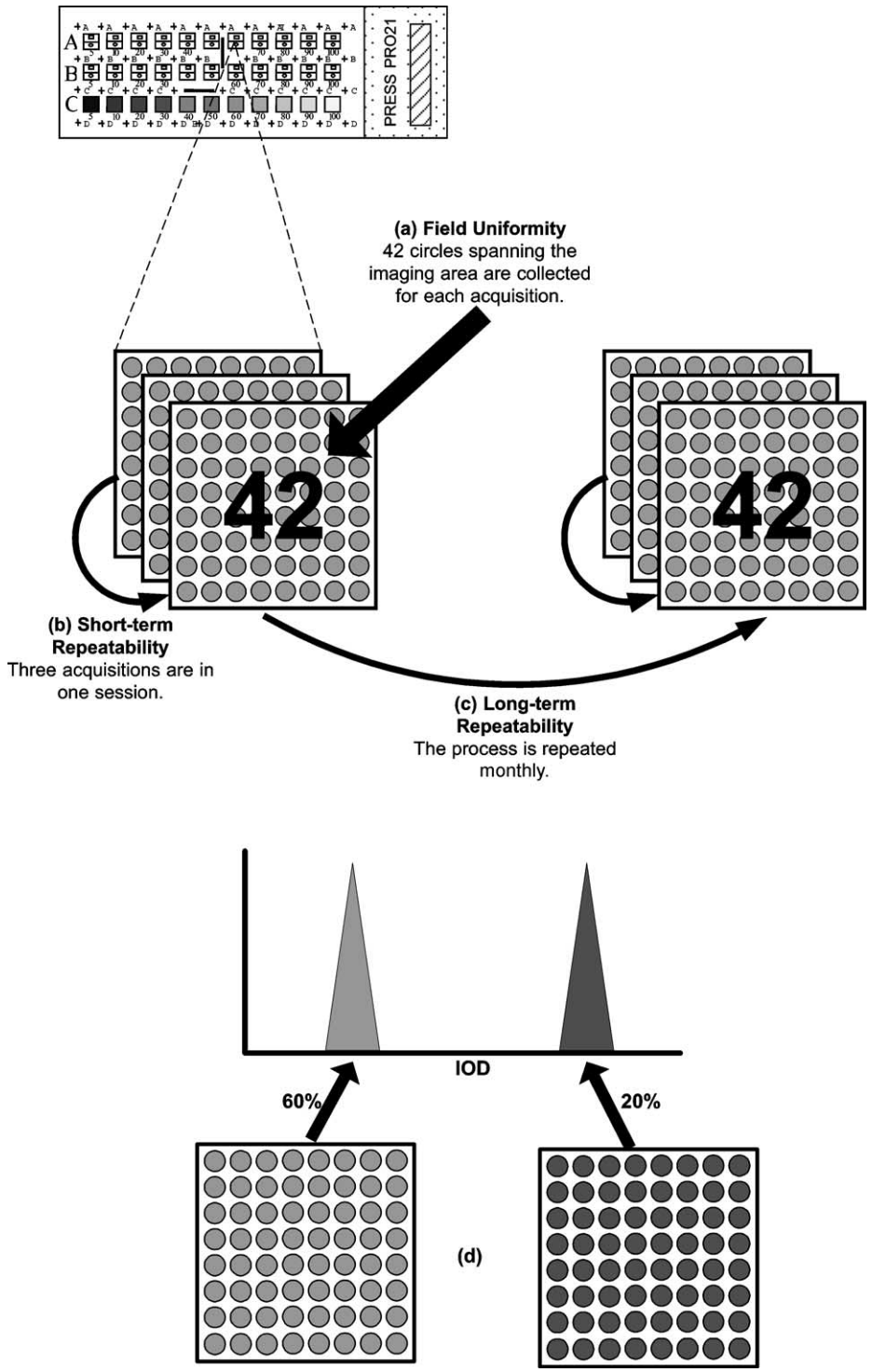
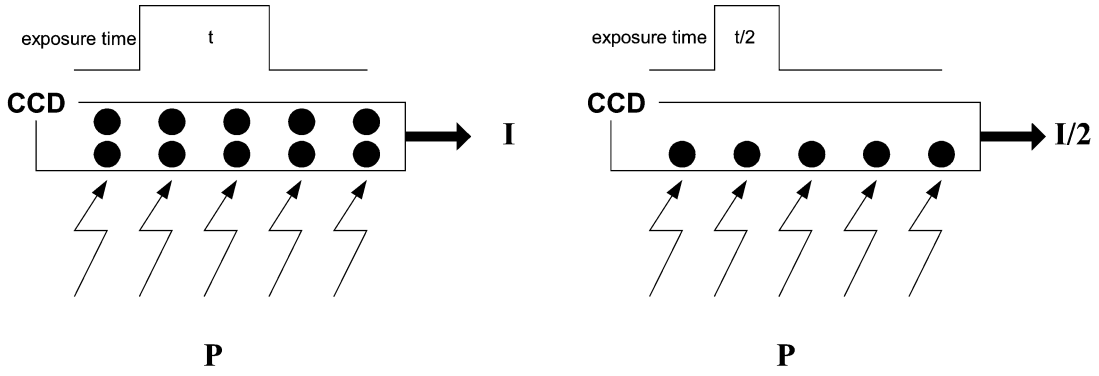


Fig. 3. Monthly QA: organization of the measurements for the monthly QA process. This illustration shows the three-tiered arrangement of the collected data. Each acquisition collects 42 circles that span the entire imaging area of the cytometer (a), which provides a measure of the (spatial) field uniformity. The acquisitions are repeated three times to provide a measure of the short-term (temporal) repeatability (b). The entire process is repeated every month to monitor the long-term (temporal) repeatability of the systems (c). Acquisitions are also made at two transmissions to provide some monitoring of the device's photo-response.



With a stable light level (P), the exposure time of the acquisition is changed to produce images with a proportionally changed intensity. The photo-response curve is a plot of the intensity versus exposure time.

a) Image on a slide with a clear scene
b) Koehler illuminate
c) Set the integration of the camera at 40ms.
d) Adjust the light intensity until the mode of the scene is 230 (out of a maximum of 255).
e) Decrease the integration time by 1 msec, and re-acquire the image.
f) Record the mode and the standard deviation at each integration time.
g) Plot the results as mode vs. integration time using the standard deviation as the y-error bars.

Fig. 4. Semi-annual QA: procedure for measuring the photo-response of an image cytometry system with these steps, the linearity of the photo-response system is confirmed. The diagram shows how, with the same illumination power, the amount of light detected by the CCD can be controlled by the exposure time. There should be a linear relation between the detected light power and exposure time.

$$S\_Area_{R3} = \sqrt{\frac{1}{2} \times \sum_{1}^{3} (M\_Area_{R3} - M\_Area_{42})^2},$$

$$M\_CVOD_{R3} = \frac{1}{3} \times \sum_{1}^{3} CV\_OD_{42},$$

$$S\_CVOD_{R3} = \sqrt{\frac{1}{2} \times \sum_{1}^{3} (M\_CVOD_{R3} - CV\_OD_{42})^2},$$

$$M\_CVArea_{R3} = \frac{1}{3} \times \sum_{1}^{3} CV\_Area_{42},$$

$$S\_CVArea_{R3} = \sqrt{1/2 \times \sum_{1}^{3} (M\_CVArea_{R3} - CV\_Area_{42})^2}.$$

The above calculations are performed on both the 20% and 60% transmission data and represent a set of monthly QA session data for one system.

In this paper, two approaches to the analysis of the monthly QA data are presented: control charts that have been used for in-process monitoring and post-process analysis that has been conducted on the entire data set to provide further insight into system performance.

In-process monitoring relies on control charts to identify "events". Eight control charts are plotted for the monthly data for each system at each transmission: X-bar charts of the means: M_OD_{R3}, M_Area_{R3}, M_CVOD_{R3}, M_CVArea_{R3}; and s-charts of the standard deviation: S_OD_{R3}, S_Area_{R3}, S_CVOD_{R3}, S_CVArea_{R3}. This produces 64 charts for the four systems. The control line (CL), lower control limit (LCL), and upper control limit (UCL) are defined by a running average. There are two reasons for using this approach rather than the standard control chart methodology [9,12]. First, there was no prior reference data to calculate the boundaries. Since these were monthly measurements, it would take too long to collect a reliable reference data set of 10 to 25 months. Second, it became apparent after a few months that the within-session variance (based on three repeated scans) was not related to the between-session variance (based on

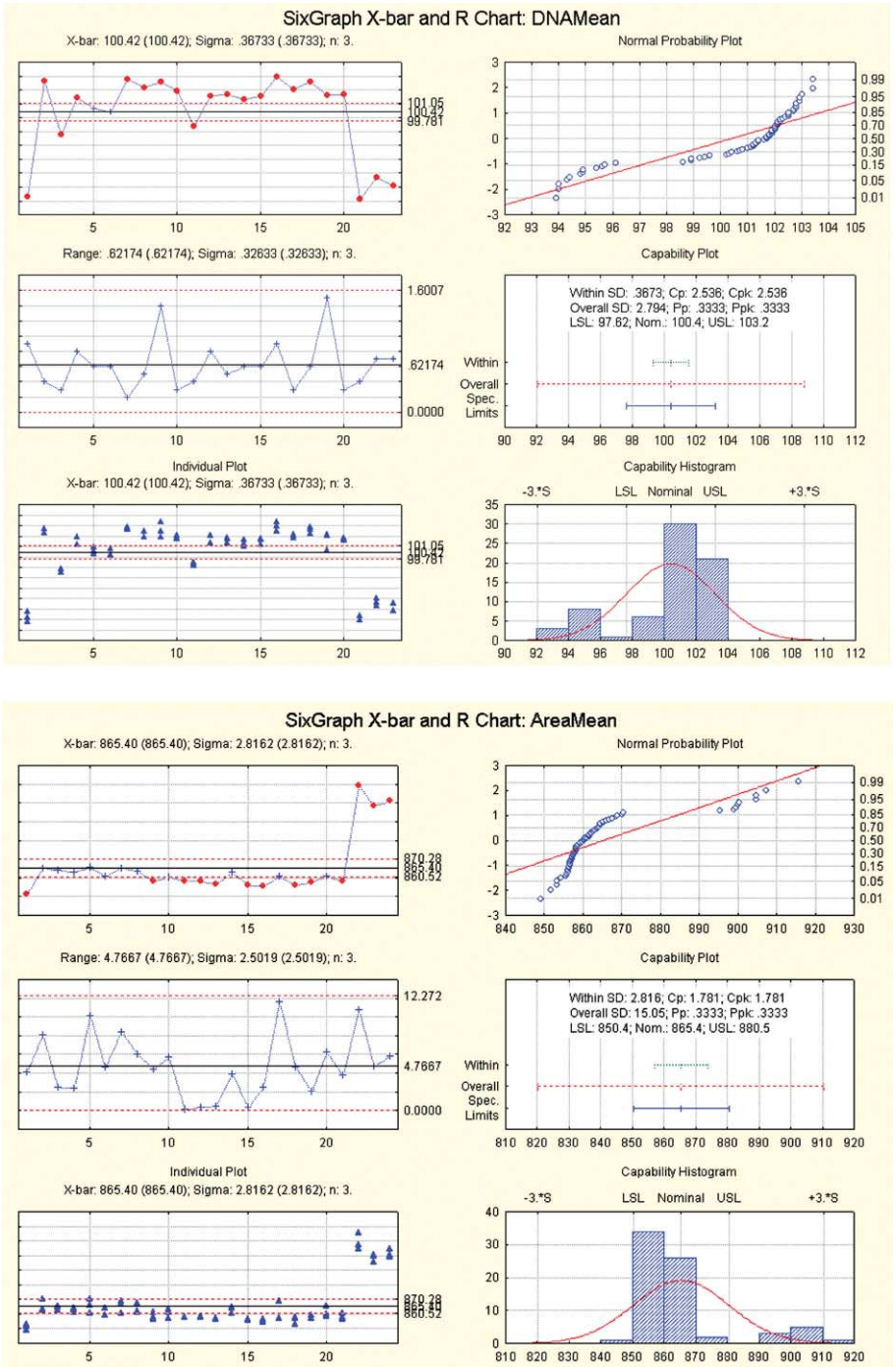


Fig. 5. Samples of the DNAMean and AREAMean data of 288 STATISTICA control charts created from the results of the monthly procedure for the OD and the AREA at the 20% and the 60% transmission of all four systems. These figures contain the X-bar chart, R-chart, individual plot, normal probability plot, capability plot, and the capability histograms.

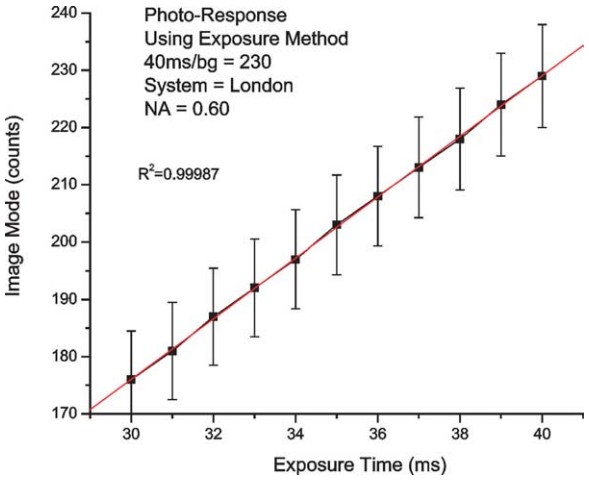


Fig. 6. Semi-annual measurement of camera linearity using the photo-response exposure method. This figure shows the photo-response of one cytometer. The x-axis is the integration time in milliseconds. The y-axis is the statistical mode of the clear scene imaged by the camera. The curve shows a definite linear behavior as noted by the correlation coefficient.

monthly average measurements) in the assumed manner. That is, the standard formulas that relate these two variances did not apply. A prototype program based on the running average approach was written to plot and analyze these charts.

The post-process analysis of the monthly session data set presented here was implemented with the SPC component of the STATISTICA software package (StatSoft, Inc. Tulsa, OK). We plotted "mean" X-bar charts and "range" (R)-charts instead of the "standard deviation" s-charts used in the in-process analysis. The upper and lower control limits were based on statistics calculated from the overall sample. Additionally, S-PLUS (Insightful Corporation, Seattle, WA) was used to perform univariate and multivariate analyses. S-PLUS is used to perform a one-way Analysis of Variance (ANOVA) to determine whether there was significantly more between-session variability than would be expected from the within session variability.

We also considered the use of a multivariate control chart, which incorporates four quantities (M_OD_{R3}, S_OD_{R3}, M_Area_{R3}, S_Area_{R3}) in a single plot. The rationale for this approach has been described [15]. A Mahalanobis distance is computed for each four-dimensional vector of observations from a session that measures the deviation from the mean across sessions "weighted" by the inverse of the covariance:

$$M_k = (x_k - xbar)' V^{-1} (x_k - xbar),$$

where M_k is the Mahalanobis distance squared, x_k is the four-dimensional vector (M_OD_{R3}, S_OD_{R3}, M_Area_{R3}, S_Area_{R3}) for the kth session, *xbar* is the four-dimensional vector for the means of the x's across sessions, and V is the 4×4 covariance matrix. The (i, j) entry of the matrix V is given by the formula

$$V_{ij} = (n-1)^{-1} \sum_k (x_{ik} - xbar_i)(x_{jk} - xbar_j),$$

where n is the number of sessions included, x_{ik} is the value of the ith variable for the kth session, $xbar_i$ is the mean of the ith variable, and the summation index k runs from 1 to n. Here, x_{1k} is the kth value of M_OD_{R3}, x_{2k} is the kth value of S_OD_{R3}, etc. The usual multivariate control charts plot T^2, which is a multiple of Mahalanobis distance.

3. Results

3.1. Establishment and application of the QA system

The QA system has been successfully implemented for all image cytometry systems at the BCCA and M.D. Anderson Cancer Center. Standard operating procedures have been specified, and the operators have been trained on their use. The QA system has been in use for over 22 months and continues to operate.

A checklist for the daily QA was created and results are recorded and assessed by the users continuously (Fig. 2). The user records the calibration image coefficient of variation (CV) and the microscope lamp power and compares them to the previous days values. To minimize time and effort, the data are recorded on paper and assessed manually. Data from the daily procedure of one system are summarized in Table 2. Little variation is seen in room temperature, calibration image coefficient of variation, lamp voltage, and lamp current in over 100 consecutive usage days. On three occasions, the results from daily procedure prompted the cytotechnician to seek assistance, which led to the discovery of "failures in progress" or events. These failures are explained in a later section.

For the monthly procedure (Fig. 3), the in-process analysis using control charts was helpful in detecting when a cytometer might not be working optimally. Events were flagged by the charts, which prompted the technician to look for possible faults. In some cases, the events were false alarms; in other cases, a failure mode was identified. Four true failures were detected. Further discussion is provided in a later section.

Table 2
Summary statistics of the daily QA data of the M. D. Anderson system

	Room temp.	Calibration image std. dev.	Calibration image CV	Lamp voltage	Lamp current
Mean	25.277	5.037	2.2654	4.3544	4.6465
N	101	102	101	102	102
Std. dev.	0.7708	0.2225	0.10036	0.03531	0.02428
Min	23.4	4.5	2.03	4.30	4.59
Max	27.4	5.6	2.50	4.46	4.69
Range	4.0	1.1	0.47	0.16	0.10
Variance	0.594	0.049	0.010	0.001	0.001

Table 3
Sample sizes (N), mean, CV within QC runs, and CV between QC runs for *IOD* and *AREA* at 20% optical transmission targets

Machine	N	Monthly QA measurement (20% transmission)							
		IOD				*AREA*			
		Mean	CV-within	CV-between	p-value	Mean	CV-within	CV-between	p-value
Gaz	54	411.1	0.32	1.08	0.0000	863.8	0.34	0.51	0.0000
London	48	394.4	0.18	1.13	0.0000	862.0	0.23	0.47	0.0000
Xtapa	57	391.0	0.16	0.82	0.0000	860.7	0.35	0.38	0.0005
MDA	51	388.1	0.22	0.80	0.0000	911.8	0.42	0.30	0.1571

Table 4
Sample sizes (N), mean, CV within QC runs, and CV between QC runs for *IOD* and *AREA* at 60% optical transmission targets

Machine	N	Monthly QA measurement (60% transmission)							
		IOD				*AREA*			
		Mean	CV-within	CV-between	p-value	Mean	CV-within	CV-between	p-value
Gaz	54	101.6	0.37	1.13	0.0000	822.7	0.37	0.49	0.0000
London	48	97.6	0.36	0.85	0.0000	822.1	0.43	0.53	0.0001
Xtapa	57	97.1	0.45	0.92	0.0000	816.6	0.64	0.58	0.0102
MDA	51	80.8	3.09	1.75	0.5189	839.0	0.37	0.38	0.0026

In the post-process analysis of the monthly data, which spans 16 to 19 months, STATISTICA was used to create 288 control charts of the *IOD* and *AREA* at the 20% and 60% transmissions of the four systems. Two samples of these charts are shown in Fig. 5. The sizes of the monthly data set vary because the QA system was implemented in a staggered fashion. We reviewed the X-bar chart, R-chart, individual plot, normal probability plot, capability plot, and capability histograms of the 288 charts. The majority of the charts showed that the data are normally distributed, suggesting that the long-term behavior may be caused by a gaussian distribution. In most cases, the within-session process variance is smaller than the between-session process variance. This difference is illustrated in Tables 3 and 4.

On a semiannual basis, the photo-response of the cytometer was measured (Fig. 6). The data was plotted, and the linearity was confirmed. These plots were compared to original plots made prior to the start of the QA system. The measurements showed that the systems' linearity did not change over the measurement period. See Fig. 6 for the photo-response of one of the cytometers. All cytometers exhibited similar linear behavior.

3.2. *Comparability and reliability of the four image cytometry systems*

Tables 3 and 4 show the means, CV within-sessions, CV between-sessions, and ANOVA p values for all four devices at the 20% and 60% transmission for both *IOD* and *AREA* from the monthly data. The within-session variance is the average of the variances from the three runs. The square root of this gives the standard deviation within-sessions, and dividing by the overall mean gives a (CV) within-sessions. The between-session variance is sample variance of the within-session means, and a between-session CV is calculated in the same manner. Note that most coefficients of variation are all small – less than 2%.

The null hypotheses of the ANOVAs presented in Table 3 are that all the means within sessions are equal for a given device. The p values show that in general there is statistically more variation between sessions than would be expected if the mean level were constant between sessions. Only two of the 16 p-values are greater than 0.05. Applying the multiple corrections procedure of Hochberg [12] shows that the 14 p-values below 0.05 remain significant even when we consider that there are 16 null hypotheses tested. We conclude that there is variation between sessions that is not the same as the variation within sessions, probably because of ambient conditions such as temperature, humidity, and other factors. As a consequence, we cannot use the within-session variance to establish control limits.

Figure 7 shows the X-bar charts of the M_OD_{R3} for the 20% transmission areas for the four systems. These charts show the long-term stability and the equivalency of the systems. The centerlines of these charts are the long-term means of the respective systems. The values of the centerlines of the four charts are acceptably close. The London system has a definite trend, but it is within the error bars. The centerlines of two of the three BCCA systems, Xtapa and London, are very close, most likely because they have the same model of camera, optics, and microscope. On the other hand, Gaz uses a different microscope, which may explain why its centerline is different from the other two BCCA systems. The component models of the M.D. Anderson Cancer Center system are the same as the BCCA Gaz system. M.D. Anderson Cancer Center does not appear to be very different. This may due to the fact that it was tested with a different reference slide. The issue of slide reproducibility is dis-

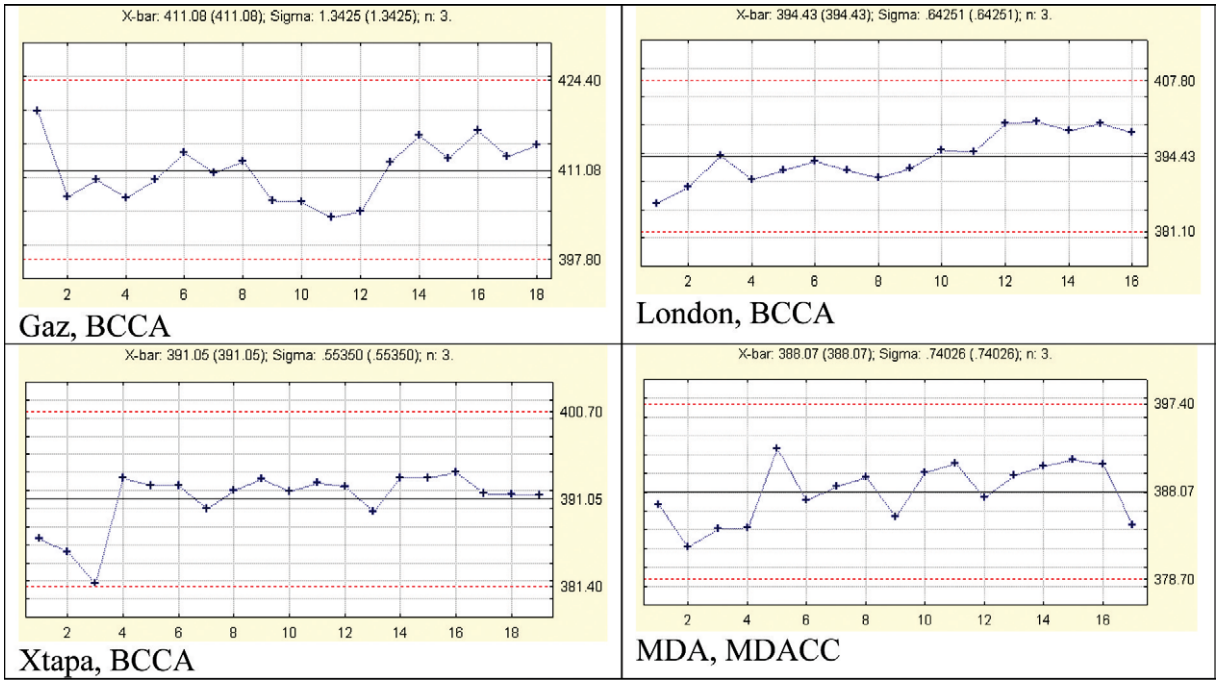


Fig. 7. Means of the three repeated means of the IOD of the 42 circles (M_OD_{R3}). This illustration shows the X-bar charts of the M_OD_{R3} for the 20% transmission areas for the four systems. These charts show the long-term stability and the equivalency of the systems. The centerline of each chart represents the mean of the M_OD_{R3} values of the respective system.

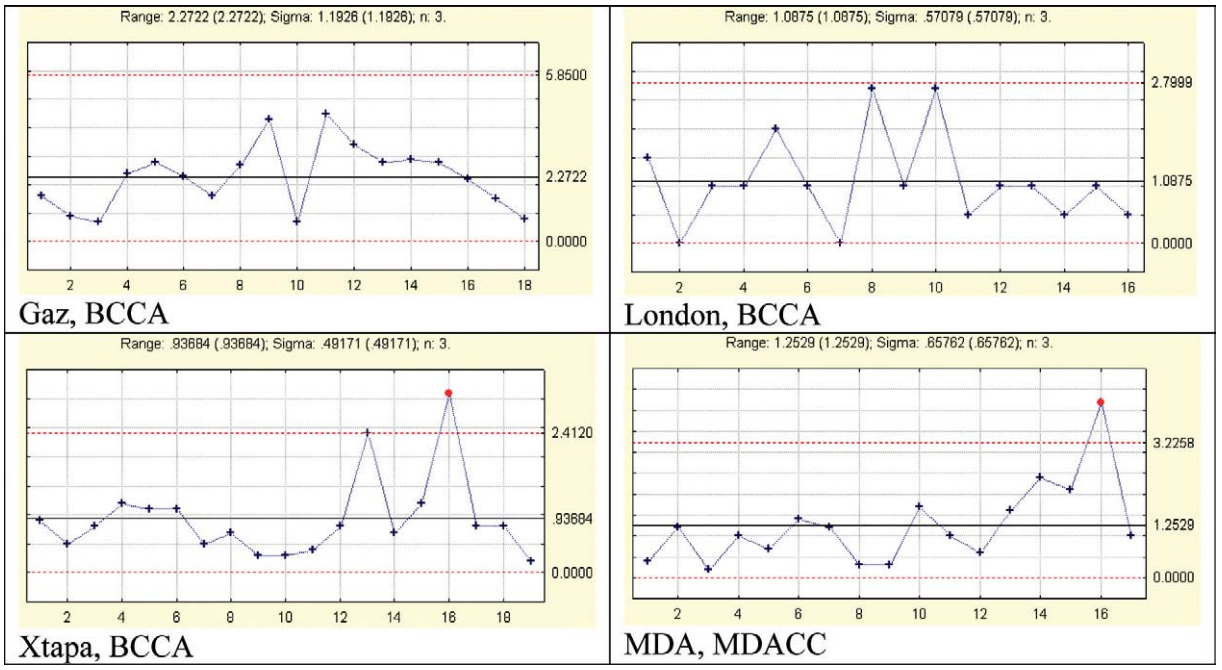


Fig. 8. Ranges of the three repeated means of the *IOD* of the 42 circles at 20% transmission. This illustration shows the R-charts of the range of the three repeated means of the *IOD* of the 42 circles for the 20% transmission areas for the four systems. The R-charts for this feature show the short-term acquisition repeatability. A small range indicates that the three repeated measures differ little; therefore, measurements can be made reliably over a short period of time. These charts reveal some events that required investigation.

cussed below. These graphs indicate that the devices are stable, consistent, and reliable over a long period of time.

Figure 8 shows the R-charts of the M_OD_{R3} for the 20% transmission areas of the four systems. The R-charts provide an indication of the short-term repeatability. A small range indicates that the three repeated measures differ little. These charts reveal some events that required investigation: one on Xtapa and one on M.D. Anderson Cancer Center. The resolution of these events is discussed below. Comparing the four centerlines shows that the mean range of the Gaz system is considerably larger than the other three. This may indicate a greater degree of mechanical wear in the focus mechanism of Gaz than the other three systems.

Figure 9 shows X-bar charts of the M_CVOD_{R3} for the 20% transmission targets for the four systems. This is a measurement of field uniformity, a small value indicating that the field uniformity is good and measurements can be made reliably across a given slide. According to the 1997 ESACP consensus report on diagnostic DNA image cytometry, Part II Specific Recommendations for Quality Assurance [6], maximum CV's of 3% for *IOD* and 2% for *AREA* are recommended. Therefore, the upper control limit of the charts in Fig. 9 has been set at 3%. For the M.D. Anderson Cancer Center system, there were three occasions where the CV was higher than 3%. Small deposits on the objective and or condenser lens were the causes.

Tables 5 and 6 show the mean coefficient of variation, the maximum coefficient of variation, and the number of times the acquisition was above accepted limits (an event occurred) at the 20% and 60% transmission, respectively. The mean *IOD* CV's vary between 1 and 3.5%, which is good. The CV's for *AREA* were all less that 2% at both 20% and 60% transmission. Recall that this measurement involves imaging 42 different circles arranged in a grid pattern spanning the camera field. The reproducibility of the 42 circles will affect the value of M_CVOD_{R3}. Another method of measuring M_CVOD_{R3} is to image the same cell in 40 different locations across the camera field of view by manually moving the cell to these locations thereby eliminating target reproducibility error. This test has been conducted on the three systems at BCCA and produced *IOD* CV's less than 3% and Area CV's less that 1%. Therefore, the discrepancy between the reference slide method and the single-cell method may indicate a slide reproducibility issue.

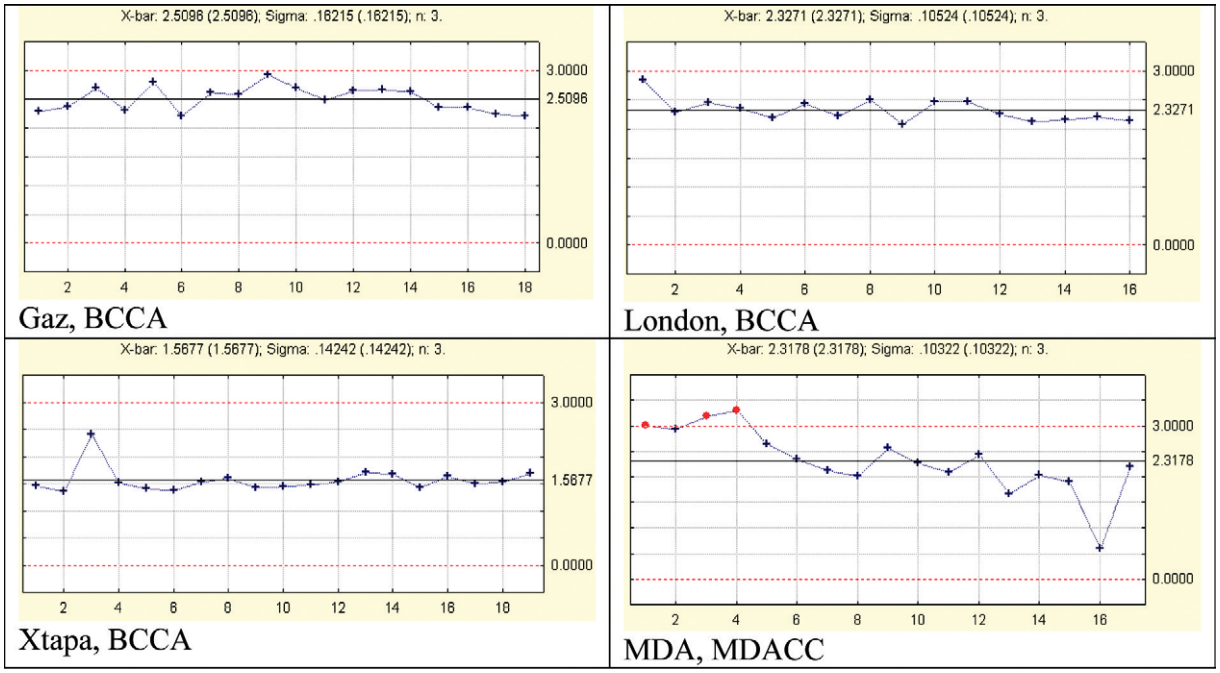


Fig. 9. Mean of the three repeated CVs of the IOD of the 42 circles (M_CVOD_{R3}) at 20% transmission. This illustration shows the X-bar charts of the M_CVOD_{R3} for the 20% transmission areas for the four systems. Because the M_CVOD_{R3} is a measure of the acquisition repeatability across the area imaged by the camera, a small value indicates that the field uniformity is good and measurements can be made reliably in space. An upper limit is set at 3% as recommended by the ESACP.

Table 5

Mean CV, max CV, and number of CV values above the accepted limits for *IOD* and *AREA* at 20% optical transmission targets

Machine	Monthly QA measurement (20% transmission)					
	IOD			*AREA*		
	Mean CV	Max CV	No. >3	Mean CV	Max CV	No. >2
Gaz	2.51	3.42	1	0.70	1.09	0
London	2.33	2.91	0	0.78	0.97	0
Xtapa	1.57	2.50	0	0.82	1.53	0
MDA	2.32	3.48	9	0.91	1.15	0

Table 6

Mean CV, max CV, and number of CV values above the accepted limits for *IOD* and *AREA* at 60% optical transmission targets

Machine	Monthly QA measurement (60% transmission)					
	IOD			*AREA*		
	Mean CV	Max CV	No. >3	Mean CV	Max CV	No. >2
Gaz	2.02	3.24	2	0.87	1.59	0
London	2.24	2.47	0	0.92	1.42	0
Xtapa	1.90	2.60	0	0.87	1.52	0
MDA	2.67	3.31	10	0.86	1.17	0

Table 7

CV values of the *IOD* and *AREA* of one HL60 cell measured at 40 locations spanning the entire camera field of view

Machine	*IOD*	*AREA*
Gaz	1.7%	0.4%
London	2.4%	0.5%
Xtapa	1.5%	0.4%

Multivariate/Quality control considers all measurements simultaneously to provide greater power to detect anomalies that are not apparent in any single variable. The multivariate analog of the X-bar chart is the T^2 chart, which plots squared Mahalanobis distance from the overall mean. Control charts for X-bar and T^2 are traditionally based on within-session variance and covariance. Because there is a greater between-session variability than would be indicated by the within-session variability, we cannot use the traditional approach for the T^2 chart, just as it was unusable for the X-bar charts. For X-bar charts, our control limits were established at $\pm 3\sigma$, where σ was estimated by the standard deviation across sessions, and the centerline of the charts was established at the long-term mean of the per session results, excluding those sessions where it was determined that an event had occurred. For the multivariate T^2 charts, we used the fact that the Mahalanobis distance computed from the population mean

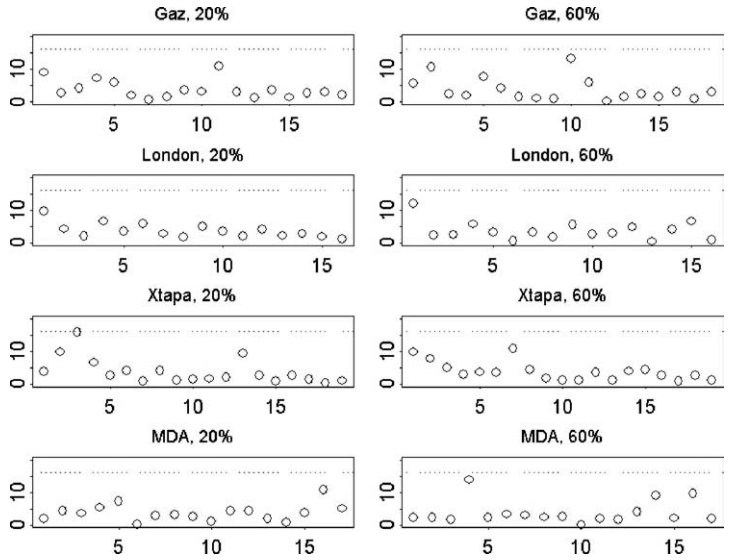


Fig. 10. Multivariate quality control charts. The vertical axis is the Mahalanobis distance from the mean computed for (M_OD_{R3}, S_OD_{R3}, M_Area_{R3}, S_Area_{R3}), which is dimensionless. The horizontal axis is the observation or session number. The UCL of 16.25 appears as a horizontal line in each plot. The plots are labeled by machine name and percentage transmission for the test circles.

and covariance has a chi-squared distribution with degrees of freedom equal to the dimensionality of the vector of observations (4 degrees of freedom for our case). We chose a quantile of this chi-squared distribution to match the false alarm rate corresponding to $\pm 3\sigma$ from the population mean. This false alarm rate is 0.0027.

The T^2 charts for each machine are shown in Fig. 10. Although we did not see any points in the multivariate control chart that were out of control (recall that we have removed points that were found to be out of control from the univariate analyses), there was one point very close to the UCL: the third observation for Xtapa. Examination of the values for this day showed that the mean M_OD_{R3} of 381.7 for that session was 2.91 standard deviations below the long-term mean M_OD_{R3} of 391.05, and the mean S_OD_{R3} of 9.24 was 3.6 standard deviations above the long-term mean of S_OD_{R3}, which was 6.13. Thus, although the point was not determined to be out of control, its large value of Mahalanobis distance suggests that there could be a problem, and by examining the particular values of the four measurements, we find that the optical density measurements are of concern.

3.3. *Event detection, problem identification, and false alarms*

One of the principle goals of the QA system is to detect major problems in the image cytometers before they impact data collection. Similarly, the impact of minor problems to data collection is minimized. Once an event is discovered, data acquisition is suspended until the event has been determined to be either a false alarm or a failure has been identified and rectified. The following are some examples of the events that were detected by the daily and monthly QA process.

There were three instances in which the daily process triggered events that were traced to real failures. In two cases, the lamp power (voltage and current) had changed considerably from the previous days. In the third case, the CV of the calibration image was considerably higher than the historical value, which indicated a worsening of the illumination uniformity. All three situations were traced to the degradation in the physical and electrical characteristics of the lamp filament. In these circumstances, it was desirable to replace the lamps before they failed completely, thereby avoiding a possible systematic error caused by a slow degradation in lamp performance.

Comparing the daily lamp voltage and current values also have indicated a situation where dust was occluding the optics of the cytometer. The dust adhering to the optics blocked the transmission of the light; therefore, more lamp power was required to establish the initial imaging light level. This has provided a sign of when to clean the systems.

The monthly QA process has helped to detect more subtle events. A few particles of dust or very small

drops of oil may not affect the parameters measured during the daily procedure, but they can have an effect on the images captured by the cytometer. On four occasions, small deposits of oil on the objective lens and the condenser lens were detected by events in the X-bar charts of the *IOD* and *AREA*. Generally, the 60% transmission measurements provided greater sensitivity to failures. The 20%T circles are very dark; their edges are highly contrasted to the bright background, which makes them easy to segment. Small deposits on the lenses may not drastically affect the optical density or the segmentation of the edge. On the other hand, the 60%T circles are light; their edges are not highly contrasted to the background. Any deposits will affect their *IOD* and edge segmentation thereby producing more drastic change in *IOD* and *AREA* values.

A limitation of using an external reference was evident when we had false alarms caused by improper and inadequate cleaning of the reference slide before measurements. If deposits such as watermarks were left over the targets, an event occurred in the control charts. Greater care in cleaning and handling the slide was implemented.

3.4. System re-qualification

At M.D. Anderson Cancer Center we were able to use the data collected from the daily and monthly procedures to re-qualify the system after it was transferred between facilities.

3.5. Reference slide reproducibility

At BCCA, a second reference slide was employed, at which time we were able to measure the reproducibility between two slides. Using the same monthly procedure conducted on all three cytometers, the data from the second slide had a difference of approximate 3% in the area at 20%T and 6% in the IOD at 60%T from the first slide. The last three points shown in both X-bar charts of Fig. 5 are the measurements from the second slide. This difference between the two slides is clearly illustrated in DNA & *AREA* measurements in Fig. 5. The difference is consistent across all three systems indicating that it is an issue of slide reproducibility. Even though slide reproducibility affects the utility of the QA system, there are methods that can be employed to mitigate this shortcoming. Cross measurements of all reference slides by all cytometers would provide the data required to do a proper cross-cytometer comparison. Another example is described in the discussion of the M_CVOD_{R3} results. A time-consuming and task-intensive method could be used to establish a baseline measurement before a less onerous procedure is employed using the reference slide as a surrogate tool. The issue of reference slide reproducibility is a critically important one that deserves further investigation with an appropriate hypothesis-driven study of sufficient size to answer remaining questions.

4. Discussion

Quantitative cyto-histopathology is a valuable tool for use in the medical field because it can replace qualitative assessment with quantitative measurement. In the developed world, this technological improvement also represents a cost savings; for the developing world, it may alleviate the problems stemming from a shortage of trained personnel and a lack of resources. Like all emerging technologies, quantitative cyto-histopathology should be subjected to rigorous technology assessment. Our successful implementation of the QA process described here shows that multiple systems can function comparably and reliably at multiple sites, which is a prerequisite for establishing the technical feasibility of quantitative cyto-histopathology. The methodology proposed here is practical, and it requires only a little extra time and resources to provide significant assurance that the systems are functioning at a high level.

Only 5 minutes per day of a cyto-histotechnician's time is required to perform the daily QA session, 30 minutes of a cyto-histotechnician's time for the data collection in the monthly QA session with an additional 10 minutes to run the analysis software, and 30 minutes of an optical engineer's time for the biannual QA session. Our QA process has resulted in several events that have led to the detection of problems but these problems have been corrected. This is a very small price to pay for the assurance that the systems are performing reliably and consistently for the success of a project involving multisite clinical trials.

There are two principle limitations with our proposed method. First, there are sources of variation that could indicate system malfunction that we are not measuring. The CytoSavant measures approximately 120 nuclear features of which only a few are directly captured by our QA measurements on the standardized slide. Furthermore, there is clearly a variation between

monthly QA sessions that exceed what would be predicted from the variation within monthly QA sessions, which leaves open questions about the sources of that variation and its possible effect on the clinical functioning of the CytoSavant. The issue of variation between operators has been discussed in a previous manuscript [4].

The second limitation is that with our primary QA sessions being performed on a monthly basis, there is the possibility that a system can be malfunctioning for an entire month before being discovered, requiring the remeasurement of numerous specimens and possible delays or reversals in patient diagnoses. The monthly frequency was chosen for practical considerations, especially of its application to a research project, but the frequency of major QA sessions should be reviewed if any routine clinical application is made of quantitative cyto-histopathology.

Some future work that can further improve the QA process, by providing greater efficiency and more detection power is under consideration. Incorporating automatic collection and analysis of daily quality control data in the acquisition software is being considered, so that the system can issue an automatic warning if it is not working correctly. Plans are underway to develop more sophisticated hardware to automatically perform the monthly check. Automation of the monthly procedure would permit it to be performed on a daily basis. We are considering using the EUROQUANT system developed by the EASCP, which provides "on line" access to QA on a daily basis [16].

Presently, the test slide provides reliable measurements of IOD and AREA. We and others are exploring the use of a more sophisticated slide that includes patterns mimicking nuclear texture. The assessment of the measurement of texture features would provide another level of quality assurance.

Acknowledgements

Special thanks go to the following individuals for their assistance and attention to detail in making the quality assurance system a success: Deanna Haskins, Jagoda Korbelik, Anita Carraro, Mark Cardeno, Tatiana Alexeenko, Maria-Teresa Arbelaez, Maria Valdizan-Garcia, Iouri Boiko, Nan Earle, and Trey Kell.

Appendix: The PRESS-PRO21 reference slide

For quantitative evaluation and calibration of your image analyzer, see Fig. 11.

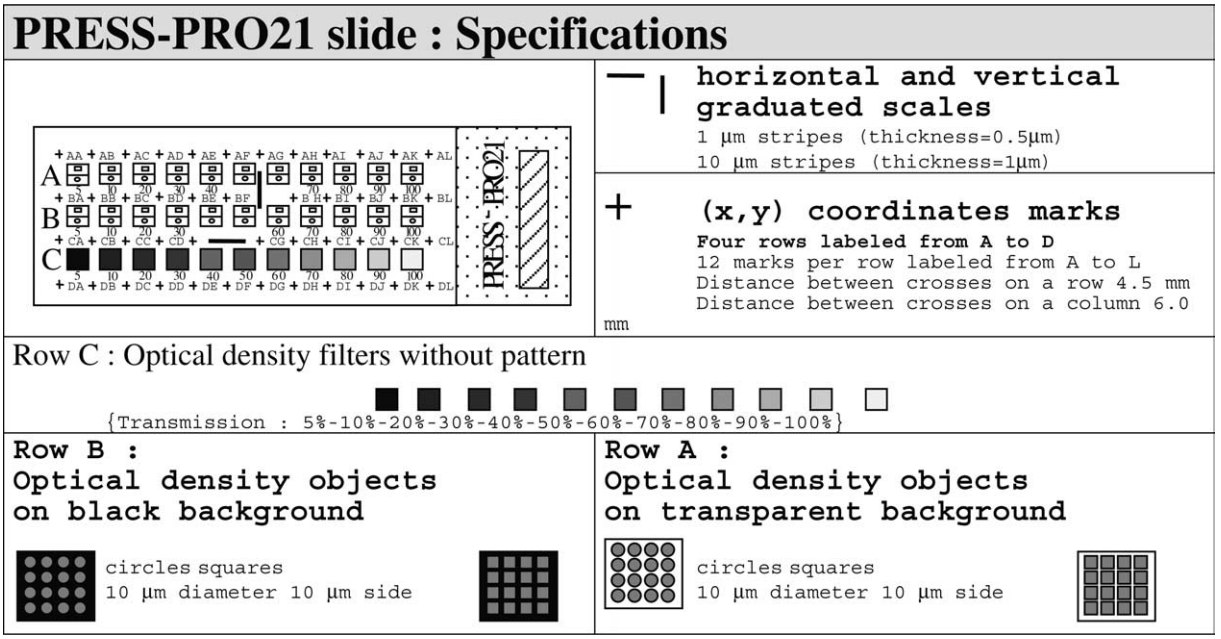


Fig. 11.

References

[1] A. Bocking, F. Giroud and A. Reith, Consensus report of the ESACP task force on standardization of diagnostic DNA image cytometry, *Anal. Cell. Pathol.* **8** (1995), 67–74; 17, 189–200.

[2] M. Chandra, *Statistical Quality Control*, CRC Press, New York, 2001.

[3] M. Doudkine, N. Poulin and B. Palcic, Nuclear texture measurements in image cytometry, *Pathologica* **87** (1995), 286–299.

[4] M. Guilland, D. Cox, A. Malpica, G. Staerkel, J. Matisic, D. Van Niekirk, K. Adler-Storthz, N. Poulin, M. Follen and C. MacAulay, Quantitative histopathological analysis of cervical intra-epithelial neoplasia sections: Methodological issues, Submitted to *Anal. Cell. Pathol.* (2003).

[5] G. Haroske, J.P.A. Baak, H. Danielsen, F. Giroud, A. Gschwendtner, M. Oberholzer, A. Reith, P. Spieler and A. Bocking, Fourth updated ESACP consensus report on diagnostic DNA image cytometry, *Analytical Image Cytometry* **23** (2001), *Analytical Image Cytology* 89–95.

[6] G. Haroske, F. Giroud, A. Reith and A. Bocking, Part I. Basic considerations and recommendations for preparation, measurement, and interpretation. The 1997 ESCAP consensus report on diagnostic DNA image cytometry, *Anal. Cell. Pathol.* (1998).

[7] G. Haroske, F. Giroud, A. Reith and A. Bocking, Part II. Specific recommendations for quality assurance. The 1997 ESCAP consensus report on diagnostic DNA image cytometry, *Anal. Cell. Pathol.* **17** (1998), 201–208.

[8] Institute of Medicine, Assessing medical technology, in: *Methods of Technology Assessment*, National Academy Press, 1985.

[9] B. Littenberg, Technology assessment in medicine, *Acad. Med.* **67** (1992), 424–428.

[10] C. MacAulay and B. Palcic, An edge relocation segmentation algorithm, *Anal. Quant. Cytol. Histol.* **12** (1990), 165–171.

[11] A. Marchevsky, T. Tomachoff and S. Lee, Quality assurance issues in DNA image cytometry, *Cytometry* **26** (1996), 101–107.

[12] NIST/SEMATECH e-Handbook of Statistical Methods; http://www.itl.nist.gov/div698/handbook, 2003.

[13] E. Ott, E. Schilling and D. Neubauer, *Process Quality Control: Troubleshooting and Interpretation of Data*, 3rd edn, McGraw-Hill, New York, 2000.

[14] M. Puech and F. Giroud, Standardization of DNA Quantification by Image Analysis, *Cytometry* **27** (1997), 21–25.

[15] F.B.J.M. Thunnissent, I.r G. Ellis and U. Jutting, Interlaboratoy comparison of DNA image analysis, *Analytical Cellular Pathology* **12** (1996), 13–24.

[16] K. Kayser, S. Blum, M. Beyer, G. Haroske, D. Kunze and W. Meyer, Routine DNA cytometry of benign and malignant pleural effusions by means of the remote quantitation server Euroquant: a prospective study, *J. Clin. Pathology* **53** (2000), 760–764.