Cumulative sum quality control for calibrated breast density measurements

John J. Heine^{a)} and Ke Cao^{b)}

Cancer Prevention and Control Division, Moffitt Cancer Center, 12902 Magnolia Drive, Tampa, Florida 33612

Craig Beam^{c)}

Division of Epidemiology and Biostatistics, School of Public Health, University of Illinois at Chicago, 1603 W. Taylor St., Chicago, Illinois 60612

(Received 15 May 2009; revised 22 September 2009; accepted for publication 23 September 2009; published 5 November 2009)

Purpose: Breast density is a significant breast cancer risk factor. Although various methods are used to estimate breast density, there is no standard measurement for this important factor. The authors are developing a breast density standardization method for use in full field digital mammography (FFDM). The approach calibrates for interpatient acquisition technique differences. The calibration produces a normalized breast density pixel value scale. The method relies on first generating a baseline (BL) calibration dataset, which required extensive phantom imaging. Standardizing prospective mammograms with calibration data generated in the past could introduce unanticipated error in the standardized output if the calibration dataset is no longer valid.

Methods: Sample points from the BL calibration dataset were imaged approximately biweekly over an extended timeframe. These serial samples were used to evaluate the BL dataset reproducibility and quantify the serial calibration accuracy. The cumulative sum (Cusum) quality control method was used to evaluate the serial sampling.

Results: There is considerable drift in the serial sample points from the BL calibration dataset that is x-ray beam dependent. Systematic deviation from the BL dataset caused significant calibration errors. This system drift was not captured with routine system quality control measures. Cusum analysis indicated that the drift is a sign of system wear and eventual x-ray tube failure.

Conclusions: The BL calibration dataset must be monitored and periodically updated, when necessary, to account for sustained system variations to maintain the calibration accuracy. © 2009 American Association of Physicists in Medicine. [DOI: [10.1118/1.3250842](https://doi.org/10.1118/1.3250842)]

Key words: cumulative sum, breast density, calibration, breast cancer risk, mammography

I. INTRODUCTION

Breast density is a significant breast cancer risk factor that is assessed from mammograms. The risk posed by breast density is approximately three to six times greater for women with extensive proportions of breast density compared to women with little breast density.^{1,2} Moreover, it is estimated that breast density in over half the breast may account for 1/3 of all breast cancers.³ Unfortunately, there is no accepted standard for measuring this important risk factor.⁴

Methods used to estimate breast density can be loosely categorized as either approaches that calibrate for acquisition technique differences before making the estimate⁵⁻⁹ or methods that make the estimation without adjusting for the acquisition. Noncalibration approaches include user-assisted computer methods,¹⁰⁻¹² fully automated statistical techniques,¹³ as well as observational estimations.^{14,15} A recent metaanalysis¹ provides a comprehensive overview of breast density risk associations by estimation technique. Few calibration methods have been validated using breast cancer as an endpoint. Although, recent studies^{16,17} found that calibration did not strengthen the risk association when compared with associations determined without adjusting for acquisition technique differences. The lack of a standard measurement for breast density may be one of the reasons this im-

portant risk factor has not been used in the clinical environment for risk assessments beyond preliminary investigations.^{15,18}

In our previous work,^{9,19} a method⁶ for calibrating full field digital mammography (FFDM) data was modified and evaluated without considering the serial reproducibility of the calibration dataset. The approach relied on generating acquisition dependent calibration curves with breast tissue equivalent phantom imaging. The calibration dataset generation required extensive phantom imaging to adequately sample the acquisition technique parameter space and detector field of view (FOV). Large-area phantoms were used that cover the majority of the detector FOV to account for spatial variations in the analysis. Calibration curves generated from the calibration dataset were well approximated as linear and were summarized by their regression parameters. The calibration uses the regression parameters, rather than the calibration curves. This initial dataset and calibration curves are referenced as the baseline (BL). The BL data analysis is in support of ongoing study to validate the calibration measure using patients' breast cancer status as the endpoint comparison. One prerequisite objective of this ongoing study is to determine the length of time that the BL dataset is valid before applying the method to patient images.

The cumulative sum (Cusum) technique, often referred to as Cusum charts, or Cusum charting, was used to analyze the initial serial samples and determine the Cusum parameters for prospective monitoring. The Cusum is a sequential technique originally developed for process control applications.^{20–22} The approach compounds previously measured samples to detect sustained shifts from the expected. As discussed here,²³ there have been many advancements to the Cusum theory since the early work. More recent work suggests Cusum applications may be useful for detecting disease outbreaks,^{24,25} monitoring various clinical related outcomes,^{26,27} investigating air pollution level variations,²⁸ and analyzing text for authorship.²⁹ Cusum analysis was also used for assessing serial change in regression relationships.^{30,31}

Cumulative sum analysis was applied to serial data to evaluate the BL dataset validity. The calibration application was used to make theoretical estimates of the acceptable serial drift from the BL dataset to initialize the Cusum process. The goal of this work is to develop a systematic quality control procedure to ensure calibration accuracy over extended timeframes.

II. METHODS

II.A. Imaging

Phantom imaging was performed with the General Electric (GE) Senographe 2000D FFDM system. This system is used for routine breast screening at the Moffitt Cancer Center, Tampa, Florida. The Senographe 2000D produces both raw and processed data. Raw data were used for this work. Image datasets are 1914×2294 pixels with 100 micron spatial resolution (pitch) and 14 bit/pixel dynamic range. This system has three target/filter combinations: molybdenum/molybdenum (Mo/Mo), molybdenum/rhodium (Mo/Rh), and rhodium/rhodium rhodium (Rh/Rh).

Phantoms used for this study were manufactured by Computerized Imaging Reference Systems (CIRS, Norfolk, VA) consisting of either 100% glandular or 100% adipose breast tissue equivalent materials. These are 18×24 cm² rectangular slabs with either 1 or 2 cm heights. These 100% equivalent phantoms were used to construct a 4 cm height even mixture (50/50) phantom containing 50% of each tissue equivalent (by height) by adding 2 cm thickness phantoms for each tissue equivalent.

II.B. Calibration and serial sampling

The BL calibration dataset was acquired in October 2006. BL calibration curves were generated by measuring the response for a particular acquisition technique (for a given voltage setting, target/filter combination, and exposure) as a function of phantom height above the breast support surface for both tissue equivalents. Linear regression analysis was applied to the height dependent calibration curves. The BL dataset, related regression analyses, and calibration approach are described in detail here.⁹ For a given phantom arrange-

ment, the phantom height was used as a surrogate for the compressed breast height, which eliminated height variation as source of uncertainty in this analysis.

The calibration application maps an arbitrary image into a normalized attenuation representation defined as percent glandular (PG). The PG pixel-scale range is (0–100) corresponding to (0–100)% fibroglandular (glandular) tissue. For a given voltage and target/filter combination, the mapping takes this form

$$PG = Mx + B, \quad (1)$$

where x is the normalized logarithmic response for an arbitrary pixel value (PV) given by

$$x = \ln \left(\frac{PV}{mAs} \right). \quad (2)$$

In this expression, mAs is the system beam current readout value. Both M and B are functions of the x-ray tube voltage (kV) and target/filter combination. Because the calibration curves are approximated as linear, four regression parameters corresponding to the slope and logarithmic intercept for each tissue equivalent specify the calibration curves. Although Eq. (1) mapping is linear, M and B are nonlinear functions of the regression parameters, as well as functions of compressed breast height.

Sample points (sampled heights) from the BL dataset were acquired approximately biweekly with phantom imaging for these acquisition techniques: Mo/Mo 26 kV, Mo/Rh 28 kV, and Rh/Rh 30 kV. Serial sampling was initiated 2 weeks after the BL data collection. These sample points (a portion of the BL dataset) support detecting both serial changes due to the detector variation or changes in the beam energy. These samples were used to determine the control parameters for serial monitoring and for exploring methods of applying and interpreting the Cusum for quality control monitoring. A sampling grid, shown in Fig. 1, was established that covered the relevant portion of the detector.¹⁹ Each individual sector is $\frac{1}{2} \times \frac{1}{2}$ cm². Cusum analysis was used to track the serial stability of these eight sample sectors for roughly $2\frac{1}{2}$ years. The Mo/Mo and Mo/Rh curves were sampled at 2 and 6 cm heights, whereas the Rh/Rh curves were sampled at 3 and 6 cm heights. To limit the presentation, only Mo/Mo and Rh/Rh representative timelines were used as graphical examples.

The calibration response for a given sample (for a specific height sample) was given by Eq. (1) with mAs=160 for both the BL dataset and serial samples. Earlier work showed for a fixed target/filter combination and kV value, one mAs sample point was sufficient⁹ for generating calibration points for data acquired with arbitrary mAs values without additional phantom imaging. Equation (1) defined the logarithmic response with PV redefined as the average pixel value within a given sector.

II.C. Cumulative sum

The Cusum measure was modified to calibration application. Each sector of the sample grid was treated as an inde-

FIG. 1. The sampling grid to scale. The larger dark rectangle represents the detector. The smaller 800×1400 pixel interior rectangle defines the region where the grid samples were acquired and represents the area that is important for the calibration application. The inner rectangle is centered on the detector in the y-direction with a 1 cm margin on the left-hand side from the outside of the detector. The smaller sector (smaller squares) numbering is coded (dark light indicates No. 1–8): sector (1, 2, 3) correspond to the top, mid, and bottom samples on the left hand side, sector (4, 5) are the middle top and bottom samples, and sector (6, 7, 8) are on the right hand side from top to bottom, respectively. The individual sector size is $\frac{1}{2} \times \frac{1}{2}$ cm.

pendent entity and monitored as a function of height and tissue equivalent. The approach was formulated by collecting samples each of size N to determine whether the process drifts out of control (OC), where N is the number of pixels within a given sector ($50 \times 50 = 2500$ pixels). The sequential univariate Cusum metric for a given sector is given by

$$S_n = \sum_{i=1}^n (x_i - \mu_0) = S_{n-1} + U_n, \quad (3)$$

where x_i is the logarithmic response of the sample given by Eq. (2), $U_i = x_i - \mu_0$, and μ_0 is the in-control (IC) parameter. Often, another form is derived by normalizing each summand with the standard deviation of x_i . The above quantity is normally plotted against the sample time index n . While the chart remains IC, the partial sums will fluctuate randomly about zero. In our application we let $\mu_0 = x_b$, which was the value of x_i estimated from the BL dataset for each of the eight regions. This choice was based on our application designed to monitor shifts from the BL dataset. Equation (3) was customized by normalizing the summand with the BL standard giving

$$U_i = \frac{(x_i - x_b)}{|x_b|}. \quad (4)$$

Equation (4) was used with Eq. (3) to produce the Cusum giving a metric scaled to x_b units. The collection of eight regions for a given tissue equivalent and acquisition technique are generically referred to as a *process* below.

There are multiple methods for using the Cusum to detect OC behavior. These include a sliding V-mask or standard type I/II error bounds.³² The V-mask is a graphical overlay method that uses a sliding mask, resembling a sideways V, placed at every point along the Cusum trajectory to determine OC behavior.³³ The OC bound h (the height of the sideways V opening) and chart constant k (the slope of the V-mask leg) can be determined based on the critical deviation and timeframe to detect change. The V-mask is a cumbersome method for detecting sustained shifts. As an alternative, we used the decision interval (DI) Cusum, which is the algebraic equivalent of the sliding V-mask to avoid the overlay plotting.³² When a given chart is reasonably described as normal, it follows from Eqs. (3) and (4) that S_n is a zero mean Gaussian random variable with variance $= x_b^{-2} \sum_{i=1}^n \sigma_i^2$, where σ_i is the standard deviation of x_i , as long as a given chart remains IC. Theoretical considerations for the Cusum analysis and how it relates to better known statistical tests are addressed elsewhere.^{20,23,34} Equations (1)–(4) were used to analyze the serial samples, estimate the chart constant, and develop a method to identify OC behavior.

The DI Cusum was used to detect shifts. The DI expressions for detecting upward and downward shifts are given by

$$S_n^+ = \max(0, S_{n-1}^+ + U_n - k) \quad (5)$$

and

$$S_n^- = \min(0, S_{n-1}^- + U_n + k), \quad (6)$$

respectively, where the values of the zero order summands are initialized to zero for both cases. The above equations result in two serial plots. Positive and negative shifts can be detected when $S_n^+ > h$ and $S_n^- < -h$, respectively. The maximum and minimum operations return zero unless the DI argument at n grows larger than the critical value k (OC slope or chart constant), which gives a visible method for observing sustained drifts. To use the DI representation as a cue for detecting OC behavior, it is essential to first determine a suitable value for the chart constant.

II.D. Chart constant

Theoretical error analysis was used to estimate the chart constant k for a predetermined accuracy. The IC Cusum behaves as a one-dimensional random walk with statistically zero slope. A persistent deviation from the zero-slope indicates a sustained shift from the expected. Estimates of k were found by shifting the BL incrementally, applying the Eq. (1) calibration to the shifted BL for each tissue equivalent for a given acquisition technique and evaluating the variation from the known. The shifted BL was expressed as $x_d = x_b + d$, where d was an incremental shift. The height was set to 4 cm to evaluate M and B in Eq. (1). This height was used because it is bound by the height of the serial sample points. The value of x_b for region 2 was used for this analysis with no loss of generality.

FIG. 2. Deviation analysis. This gives the theoretical percent glandular deviation as a function of relative (fractional) logarithmic response (Log-response) shift from the baseline (positive shift) for the adipose tissue equivalent examples: molybdenum/molybdenum (solid), molybdenum/rhodium (cross), and rhodium/rhodium (diamond).

II.E. Calibration accuracy

The accuracy of standardizing prospective image data without considering the validity of the BL calibration data was evaluated by applying the Eq. (1) mapping to the 4 cm 50/50 phantom serial images for the Mo/Mo 26 kV and Rh/Rh 30 kV samples. The No. 2 region serial samples were used for demonstration purposes. These 50/50 serial samples were also monitored over the same (similar) timeframe as the BL samples but were not part of the BL dataset.

III. RESULTS

III.A. Chart constant estimation

Error analysis was performed for positive shifts from x_b because the resulting deviation was symmetric function about zero deviation. Figure 2 shows the theoretical PG variation as function of fractional shift from x_b for the three adipose examples. The shift axis was scaled to fractional deviation, $d \rightarrow d/x_b$, to match Eq. (4). The Rh/Rh variation (diamond) changed more rapidly than the other examples. Because the curves were linear and pass through zero, regression analysis was used to estimate the slopes. Table I

TABLE I. Rate of change per relative logarithmic response shift. This shows the fractional deviation for the three examples for both adipose and glandular equivalents for a phantom height=4.0 cm. The percent glandular (PG) error is given in the bottom row for each tissue equivalent example corresponding to a 1% positive shift in logarithmic response.

Adipose	Mo/Mo	Mo/Rh	Rh/Rh
Rate of change	-224.7	-316.5	-422.7
PG error	-2.24	-3.16	-4.22
Glandular			
Rate of change	-122.2	-217.7	-322.5
PG error	-1.22	-2.17	-3.22

shows the regression slopes for both the glandular and adipose examples. The slope is interpreted as the PG change per fractional logarithmic response (positive) shift relative to the BL. These quantities indicated (a) the error (absolute) increased as the beam energy increased and (b) the error increased more rapidly for the adipose equivalents. Although not shown, the error increased as the phantom height decreased while retaining the relative relationships shown in Table I. The Rh/Rh 30 kV adipose relation was the most sensitive to change for these examples. A 1% shift in the logarithmic response induced approximately 4%–5% error in PG as shown in Table I for this adipose example. Therefore, the Rh/Rh adipose findings were used to estimate k . Table I also gives the PG error for a 1% shift for the other examples as well. For the DI analysis, $k=0.01$. Alternatively, k could be selected as a function of the acquisition technique to give uniform accuracy. For the interacquisition comparisons discussed below, this static value of k was used to initialize the DI Cusum.

III.B. Decision interval Cusum

The DI demonstration was restricted to the Mo/Mo and Rh/Rh examples below to limit the presentation length. The independent axes were labeled in days. Although for the following data modeling analyses, the equally spaced Cusum index n was used as the independent variable unless stated. At this time the detection threshold h was not considered because the data were used to explore the Cusum behavior and estimate the OC bound. The Mo/Rh findings are cited without example. The x-ray tube failed (January 2009) and was replaced. After the x-ray tube replacement, the Cusum was restarted and reinitialized after acquiring a second calibration dataset that mirrored the BL dataset (a second duplicate BL dataset) at 819 days (3 February 2009). The serial sampling terminated at 944 days (10 June 2009) for these analyses. A second set of calibration parameters was generated with regression analysis and compared with the original BL calibration parameter set. Although not shown, the second BL regression parameter set was in very close agreement with the initial BL set described previously.⁹ Serial sampling was initiated again 12 days after the second BL data collection. After the x-ray tube replacement, Cusum charts included corresponding samples from the second BL dataset as the first sample point.

Two example logarithmic response timelines are shown in Fig. 3. These timelines show general downward tendencies with a change in downward trend occurring between 449 and 500 days. The trajectory also tends to flatten between 575 and 800 days. The sudden jump occurring between 763 and 819 days shows when the x-ray tube failed and was replaced. The portion of the charts before the x-ray tube failure is discussed first (up to 763 days). Table II gives the averaged BL logarithmic response values corresponding with the serial samples for each height, acquisition technique, and tissue equivalent for comparison purposes.

Figures 4 and 5 show DI charts derived with Eq. (6) (downward drift detection only) for the Mo/Mo adipose and

FIG. 3. Logarithmic response timelines. These show the logarithmic response (LR) serial samples for the 6 cm adipose rhodium/rhodium (top) and molybdenum/molybdenum (bottom) timelines. The samples after the x-ray tube replacement start at day=819.

glandular samples. The corresponding Rh/Rh samples are shown in Figs. 6 and 7. Two time segments were used corresponding to before and after the x-ray tube failure/replacement. The DI control behavior for each process is summarized in Table III. In this table, a given process was labeled IC as long as all sample regions (all charts) of the process were IC over the time segment. The intervals show the time range from when at least one chart of a given process was OC to when all charts of the processes were OC. Many of the DI charts showed change in the downward drift rate (change point) between 449 and 463 days. There is a gap in the charts at this time, which is most likely coincidental. Gaps represent missing serial samples. This and other noticeable gaps, other than for the x-ray tube failure gap, are due to mammography system downtimes for reasons that would not be expected to influence logarithmic response values. These peripheral reasons include software installation or mammography workstation maintenance.

III.C. Standard Cusum

To evaluate the DI Cusum initialization, the standard Cusum was applied using the Eq. (4) normalization. Figures 8 and 9 show the standard charts for the Mo/Mo examples, which for the most part show downward trends. Because the

FIG. 4. Molybdenum/molybdenum adipose Decision interval (DI) Cusum charts. The top plots show the 2 cm charts for all regions and the bottom plot shows the 6 cm charts for all regions. The regions are denoted by these symbols (1) dash, (2) plus sign, (3) asterisk, (4) diamond, (5) triangle, (6) square, (7) multiplication, and (8) circle, respectively, which are also used in the following charts.

Mo/Mo 2 cm adipose charts appeared linear, linear regression was applied to each chart to assess the DI initialization. The Mo/Mo 2 cm charts showed change in the downward trend at the change point referenced above, as do the other charts. Because the Mo/Mo adipose charts appeared piecewise linear, they were split into two line segments at the change point (about 449–463 days) and fitted separately to assess the chart parameter choice. The first segment included the initial 449 days and the second segment included 463–762 days (up to the x-ray tube replacement). When there is a persistent shift of a fixed size (a step shift) starting at some point $m < n$, the Cusum will track about a linear path with a slope=step shift.³² The other Mo/Mo charts showed more variation in their downward behavior and were not suitable for linear regression analysis. We note that the slope derived with regression analysis is not the optimal slope estimator for the Cusum,³² but under these circumstances with high linear correlation, regression analysis was appropriate to assess the merits of the chart constant choice. The regression findings (slopes) are listed in Table IV. All of the 2 cm adipose absolute value slopes were less than 0.01 for the initial segment (S1), which agreed with the corresponding DI

TABLE II. The average baseline logarithmic response quantities. Each row (top) entry is the average value, $\langle x_b \rangle$, of the eight regions for a given process. The average percent difference, $100\% \times (\langle x_m \rangle - \langle x_b \rangle) / \langle x_b \rangle$, between the last serial sample ($n=47$) before the x-ray tube failure and the baseline (BL) is provided beneath the respective BL quantities.

	Mo/Mo		Mo/Rh		Rh/Rh	
	2 cm	6 cm	2 cm	6 cm	3 cm	6 cm
Adipose	3.966	1.669	4.364	2.320	4.329	2.993
	-1.22%	-2.91%	-1.08%	-2.75%	-2.27%	-4.00%
Glandular	3.274	-0.066	3.783	0.837	3.611	1.687
	-1.36%	-50.44%	-1.51%	-6.44%	-3.27%	-6.39%

FIG. 5. Molybdenum/molybdenum glandular decision interval Cusum charts.

FIG. 6. Rhodium/rhodium adipose decision interval Cusum charts.

charts indicating the drift was less than the chart constant (critical slope shift). In contrast, the absolute value slopes for the second segment (S2) for the most part were greater than 0.01 indicating OC behavior that agreed with DI charts for this interval. The linear correlation values indicated that the Cusum adipose plots were roughly linear, which suggests the shift was well approximated as a sudden persistent step change.

The Mo/Mo glandular 6 cm charts showed large shifts from the BL. We assume that these Mo/Mo 6 cm glandular samples are too noisy to be of analytical importance. To a lesser extent, the Mo/Rh 6 cm glandular samples were also noisy and may not be of much analytical importance. To summarize the variation from the BL, the averaged percent difference taken over all regions between the BL and the last serial sample is provided in Table II.

In contrast, the Rh/Rh charts all have similar profiles. These charts have a nonlinear downward trend that appeared quadratic as shown in Figs. 10 and 11. A quadratic Cusum chart indicates a linear shift³² rather than step shift. To test for a quadratic trend, the Rh/Rh charts were fitted with a

quadratic model given by $S_n = c_0 + c_1n + c_2n^2$ for the initial time segments between 1 and 449 days. As an example, the 3 cm adipose Cusum quadratic model for region 2 is shown in Fig. 12. This model agreement is representative of the other Rh/Rh chart modeling results, which are not shown. The derivative of S_n provides an estimate of the linear change, which gives $D = c_1 + 2c_2n$. The averaged derivative coefficients for each process are listed in Table V with the associated standard deviations. The c_0 coefficients were ignored because they were close to zero and not relevant to this analysis. Using the averaged coefficients for a given process and evaluating the derivatives for $n=32$ (day=449) gives an estimate of the average relative deviation trend from the BL to the first change point. For the glandular derivatives: $D_3(32) \sim -0.031$ and $D_6(32) \sim -0.066$, where the subscript indicates the height. Likewise, for the adipose derivatives: $D_3(32) \sim -0.022$ and $D_6(32) \sim -0.038$. All Rh/Rh charts showed change beyond the critical value as indicated by the summaries in Table III. Solving the D equation for n with $D = -0.01$ and converting to days indicates it took approximately 169 days for the 3 cm adipose and 62 days for the

TABLE III. Decision interval Cusum control summary. The in-control (IC) behavior is summarized in two segments for before the x-ray tube failure (top) and after the replacement (bottom). Entries cited as intervals represent the range from when at least one chart of a given process was out of control (OC) to the day that all charts of the process were OC. The first segment is measured from the first serial sample (add 14 days to adjust to the baseline collection) and the second segment is measured from the day of the second baseline dataset acquisition. Isolated numerical entries indicate that all charts of a given process were OC in unison by the specified day. Processes that were IC over a given segment are labeled as such.

	Mo/Mo		Mo/Rh		Rh/Rh	
	2 cm	6 cm	2 cm	6 cm	3 cm	6 cm
Before x-ray tube failure						
Adipose	393–438	15–364	508	22–379	169–183	22–145
Glandular	393–438	15–225	424	92–225	134–145	36–78
After x-ray tube replacement						
Adipose	IC	12	IC	40–82	IC	IC
Glandular	99–113	12	IC	10–62	IC	IC

FIG. 7. Rhodium/rhodium glandular decision interval Cusum charts.

6 cm adipose charts to pass the critical value and move OC. Similarly, the Rh/Rh glandular charts took approximately 123 and 62 days for the 3 and 6 cm samples to move OC, respectively. These time intervals are in agreement with the corresponding DI charts indicating the chart constant is appropriate.

In the initial segment before the x-ray tube replacement, the Mo/Rh Cusum charts in form (not shown) more closely resembled the Mo/Mo form than the Rh/Rh form for a given process. The Mo/Rh charts did not exhibit the parabolic Rh/Rh characteristic. Although the charts appeared similar, the Mo/Rh adipose charts dropped about $\frac{1}{2}$ as rapidly as the corresponding Mo/Mo adipose charts at given time point, whereas the Mo/Rh glandular 2 cm charts were very similar to 2 cm glandular Mo/Mo charts. The Mo/Rh 6 cm glandular charts were similar with the corresponding 6 cm Mo/Mo glandular charts up to 449 days, but afterwards the Mo/Mo charts dropped considerably faster (by as much as factor of 10 in some cases). An outlier was removed from the Mo/Mo glandular samples at $n=14$ and replaced with the average value of adjacent samples. The Mo/Mo and Mo/Rh glandular 6 cm samples contain more noise due to the greater attenuation as indicated in Table II.

FIG. 8. Molybdenum/molybdenum adipose standard Cusum charts.

III.D. Calibration accuracy

The calibration accuracy was evaluated without accounting for serial variation from the BL dataset. The 50/50 phantoms for the Rh/Rh and Mo/Mo acquisition techniques were also sampled biweekly since the initiation of the serial study. Equation (1) was applied to the serial samples, whereby mapping them to the PG representation. Unlike the 100% equivalents, we have no BL measure for these 50/50 mixtures. Each mapping was split into two separate timelines for before and after the x-ray tube replacement. To show the influence, the four timeline segments were fitted with linear regression analysis. Figure 13 shows the mapped timelines and fitted line segments. The regression parameters are listed in Table VI.

IV. DISCUSSION

The work showed distinct patterns in the serial samples. The Rh/Rh charts showed an approximate characteristic quadratic form in comparison with the Mo/Mo and Mo/Rh charts before the x-ray tube failed. In general, the Rh/Rh charts were OC before the other charts. The Mo/Mo and Mo/Rh charts were more amorphous in comparison. There appears to be a change point in the Cusum trajectories before the tube replacement occurring about 449–463 days that is

TABLE IV. The molybdenum/molybdenum adipose 2 cm Cusum regression slopes (logarithmic response per Cusum index) for the first (top) and second (bottom) Cusum segments indicted by S1 and S2, respectively. The lower entries in each row represent the linear correlation coefficient for each fitted line.

Region	1	2	3	4	5	6	7	8
S1	−0.008	−0.009	−0.007	−0.006	−0.006	−0.004	−0.005	−0.003
	−0.99	−0.99	−0.99	−0.99	−0.99	−0.98	−0.99	−0.96
S2	−0.011	−0.011	−0.010	−0.010	−0.011	−0.011	−0.011	−0.009
	−0.99	−0.99	−0.99	−0.99	−0.99	−0.98	−0.99	−0.98

FIG. 9. Molybdenum/molybdenum glandular standard Cusum charts.

FIG. 11. Rhodium/rhodium glandular standard Cusum charts.

more pronounced in the Mo/Mo and Mo/Rh charts than in the Rh/Rh charts. In contrast after the x-ray tube replacement, the Rh/Rh processes remained IC, whereas the Mo/Mo processes for the most part showed OC behavior, as did the Mo/Rh 6 cm processes. In addition to the general OC behavior before the x-ray tube failure, both the Cusum chart variations and change point may be indications of system wear and eventual system failure.

Equation (3) Cusum representation was used to validate the chart constant choice and investigate the chart profiles. Both linear regression and curve fitting analyses showed the DI chart constant choice was valid. These modeling techniques were also used to describe the type of sustained serial change that may have occurred by modeling Eq. (3) charts as linear or quadratic. The Rh/Rh Cusum charts showed quadratic behavior indicating a linear change, whereas some Mo/Mo charts were better modeled as linear indicating a step change. These fitting techniques were not applied to find the best-fit model, in general, but were used to develop coarse descriptions of the change. For example, the quadratic trend may be an indication of continual system wear.

The work showed that the DI approach provided a visible cue for detecting when the system begins to drift past the acceptable tolerance. The present work did not explicitly address when a statistical decision should be made to accept the DI as a true change or not accept it because it represents a random fluctuation across the decision boundary. The statistical, decision-theoretic, aspect of the Cusum will be addressed in a subsequent paper so that the monitoring can be automated. The serial points represented 2 week intervals, approximately. Our patient accrual rate is approximately 1 patient/week. A persistent downward change, as detected by the DI Cusum, for two consecutive Rh/Rh samples, would introduce more than 5% deviation from the known when standardizing image regions corresponding to 100% adipose tissue. It follows that a continued monotonic trend detected for more than two consecutive serial points would be sufficient to trigger corrective actions, which provides a prescription for using the DI Cusum for visually detecting a change in the calibration application quality.

The calibrated 50/50 serial examples shown in Fig. 13 demonstrate the need for assessing the BL serial reproduc-

FIG. 10. Rhodium/rhodium adipose standard Cusum charts.

FIG. 12. Rhodium/rhodium region No. 2 Cusum chart. The chart (diamond) was fitted with quadratic polynomial (solid).

TABLE V. Rhodium/rhodium quadratic coefficients c_1 and $2 \times c_2$. This gives the coefficients for the derivative of the fitted second degree polynomial for the rhodium/rhodium Cusum charts averaged (Ave) over the eight regions: $D = c_1 + 2c_2n$, where n is the serial index. The data spans 1–449 days, corresponding to n spanning 1–32. The intraprocess standard deviations (SD) are also provided.

	3 cm adipose		6 cm adipose		3 cm glandular		6 cm glandular	
	Ave	SD	Ave	SD	Ave	SD	Ave	SD
c_1	−0.0028	0.0014	−0.0055	0.0035	−0.0022	0.0010	−0.0011	0.0037
$2 \times c_2$	−0.0006	0.0001	−0.0010	0.0001	−0.0009	0.0001	−0.0020	0.0001

ibility. In the segment before the tube replacement, the Rh/Rh mapped timeline reached as high as $PG \approx 65$, whereas the corresponding Mo/Mo timeline extended to $PG \approx 60$. The regression slopes listed in Table VI for before the tube failure show a steady increase in the Rh/Rh calibrated example, which is consistent with the logarithmic response decrease. Because, lower response implies greater attenuation that leads to greater PG. Likewise, the near zero slope for the Mo/Mo example is consistent with a step change in the logarithmic response. Applying the Cusum to the Rh/Rh calibrated PG serial representation resulted in a positive valued parabola (not shown) indicating the quadratic form holds across the mapping. The Rh/Rh 6 cm adipose DI charts indicate that corrective action was warranted at about day=60 in the segment before the x-ray tube failed. Hypothetically, action at this time point would have corrected the calibration before the error reached $PG=54$. In contrast, the Rh/Rh calibration example after restarting the Cusum, after the tube replacement, was closer to the ideal $PG=50$ value in comparison with the corresponding Mo/Mo samples as shown in Table VI. This Mo/Mo variation may be caused by the relevant 6 cm OC processes shown in Table III.

By assumption, the Cusum DI analysis indicated signs of system fault long before the failure occurred. To confirm this assumption, it will be necessary to characterize this system's serial behavior when operating under normal conditions without component failure for extended periods. All bi-weekly experimental imaging was performed after the system passed the built-in flat field quality diagnostic routine. Moreover, the system passed the half value beam quality assessments performed in October 2008. These examples show that the accuracy of BL calibration curves must be monitored, and the calibration data must be either regenerated or updated periodically to maintain accuracy. The clinical significance of these findings is not clear. It may be that there is greater latitude in the acceptable pixel value varia-

tion for clinical images in comparison with the acceptable variation when making calibrated measurements.

Large surface-area phantoms were used in this work and in our earlier developments. Similar calibration information can be derived with step-wedge (SW) phantoms. The choice of phantom is influenced by experimental objectives. The choice of phantom comes with tradeoffs. More calibration information can be derived from a single acquisition with SW phantoms because a range of tissue compositions and heights can be built into a single phantom.⁷ However, spatial localization is sacrificed with SW phantoms because of their geometry. Our approach supports investigating the calibration as a function of precise detector location, as demonstrated.

V. CONCLUSION

There are several unique aspects of this work. Although the FFDM systems may overcome many of the problems with film, this work showed that serial response reproducibility is not necessarily a given system characteristic. We have introduced a quality control measure for use in epidemiological studies and customized it to monitor the BL calibration dataset serial reproducibility. The acceptable deviation followed from evaluating the calibration model with error analysis. Cusum detection may have indicated pending system failure not captured with routine system diagnostics.

The work presented here shows the initial portion of an ongoing multiyear serial analysis. Determining the appropriate method for monitoring the calibration quality over extended timeframes represents one component of this ongoing calibration investigation. Future work includes monitoring many years of serial sampling to understand the limits of this particular FFDM system in reproducing quantitative measurements when the system is operating under normal conditions. Future work also includes developing automated

TABLE VI. 50/50 mixture percent glandular (PG) regression analysis. These regression parameters correspond to the plots shown in Fig. 13. The slope is estimated in PG/day units.

	Before x-ray tube failure			After tube replacement		
	Slope	Intercept	Correlation	Slope	Intercept	Correlation
Rh/Rh	0.016	53.831	0.905	0.006	50.772	0.905
Mo/Mo	0.006	53.494	0.802	−0.006	56.136	−0.383

FIG. 13. 50/50 mixture serial calibration regression examples. These show the percent glandular mapping (diamond) applied to the rhodium/rhodium (top) and molybdenum/molybdenum (bottom) 50/50 mixtures. Regression lines are solid. The analysis was performed separately for the segments before and after the tube failure/replacement.

methods based on statistical decision theory to use the serial sampling to update the BL data when the DI Cusum is triggered rather than generating calibration data repeatedly. In typical applications of monitoring output production quality, the Cusum is restarted after correcting the production problem. Because our application does not permit modifying the mammography unit when the calibration samples deviate too far from the expected as determined by the Cusum detection, a method must be developed to use the serial samples to update/adjust the BL calibration data to improve the resulting calibration accuracy.

This work isolated the serial uncertainty from other sources of error, such as compression height artifacts.^{35,36} We are currently investigating these other sources of uncertainty as well. The serial samples were based on specific spatial sizes. These sizes were estimates of the intended calibration resolution that will be used for calibrating patient mammograms. More work is required to understand the appropriate spatial resolution used for calibrating patient images. Likewise, the height samples were also estimates. As shown in this work, some of the height samples are probably too noisy for analytical purposes. The sample heights will require some modifications in the future work. The work presented here applies to one specific mammography unit. Planned research includes extending the methods developed here to other mammography units for comparison purposes to assess the generality of our findings.

There is pretty much consistent agreement among researchers and clinicians that FFDM data have many advantages over film data. Our previous work⁹ showed that the effective attenuation coefficients generated from two similar FFDM units were in close agreement, whereas the logarithmic intercepts showed system related variability. This previous work in conjunction with these Cusum findings show that FFDM data still represent a challenge when attempting to force reproducible quantitative measures from systems that were not explicitly designed to produce quantitative out-

puts. If we assume that digital detectors are superior to film in reproduction capability, then the work using FFDM datasets may set the accuracy standard with all other influences being equal. If these assumptions hold, the variation in film data may prove difficult to standardize based on the ability (or lack of) of this FFDM to reproduce measures. This work shows that sources of calibration uncertainty must be understood and compensated prior to the calibration application.

ACKNOWLEDGMENTS

This work was supported by NCI Grant No. R01 CA114491. The authors would like to thank Autumn Smallwood and all of the staff at the Lifetime Cancer Screening Center at the Moffitt Cancer Center for their support of the work.

^{a)} Author to whom correspondence should be addressed. Electronic mail: john.heine@moffitt.org; telephone: 813-979-6719

^{b)} Electronic mail: ke.cao@moffitt.org; telephone: 813-745-6278

^{c)} Electronic mail: cbeam@uic.edu; telephone: 312-996-1589

¹V. A. McCormack and I. dos Santos Silva, "Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis," *Cancer Epidemiol. Biomarkers Prev.* **15**, 1159–1169 (2006).

²N. F. Boyd, H. Guo, L. J. Martin, L. Sun, J. Stone, E. Fishell, R. A. Jong, G. Hislop, A. Chiarelli, S. Minkin, and M. J. Yaffe, "Mammographic density and the risk and detection of breast cancer," *N. Engl. J. Med.* **356**, 227–236 (2007).

³N. F. Boyd, J. M. Rommens, K. Vogt, V. Lee, J. L. Hopper, M. J. Yaffe, and A. D. Paterson, "Mammographic breast density as an intermediate phenotype for breast cancer," *Lancet Oncol.* **6**, 798–808 (2005).

⁴J. Couzin, "Breast cancer: Dissecting a hidden breast cancer risk," *Science* **309**, 1664–1666 (2005).

⁵R. Highnam and M. Brady, *Mammographic Image Analysis* (Kluwer Academic Publishers, Boston, MA, 1999).

⁶J. Kaufhold, J. A. Thomas, J. W. Eberhard, C. E. Galbo, and D. E. Trotter, "A calibration approach to glandular tissue composition estimation in digital mammography," *Med. Phys.* **29**, 1867–1880 (2002).

⁷O. Pawluczyk, B. J. Augustine, M. J. Yaffe, D. Rico, J. Yang, G. E. Mawdsley, and N. F. Boyd, "A volumetric method for estimation of breast density on digitized screen-film mammograms," *Med. Phys.* **30**, 352–364 (2003).

⁸J. A. Shepherd, L. Herve, J. Landau, B. Fan, K. Kerlikowske, and S. R. Cummings, "Novel use of single x-ray absorptiometry for measuring breast density," *Technol. Cancer Res. Treat.* **4**, 173–182 (2005).

⁹J. J. Heine and J. A. Thomas, "Effective x-ray attenuation coefficient measurements from two full field digital mammography systems for data calibration applications," *Biomed. Eng. Online* **7**, 13 (2008).

¹⁰J. W. Byng, N. F. Boyd, E. Fishell, R. A. Jong, and M. J. Yaffe, "The quantitative analysis of mammographic densities," *Phys. Med. Biol.* **39**, 1629–1638 (1994).

¹¹N. F. Boyd, J. W. Byng, R. A. Jong, E. K. Fishell, L. E. Little, A. B. Miller, G. A. Lockwood, D. L. Tritchler, and M. J. Yaffe, "Quantitative classification of mammographic densities and breast cancer risk: results from the Canadian National Breast Screening Study," *J. Natl. Cancer Inst.* **87**, 670–675 (1995).

¹²J. W. Byng, M. J. Yaffe, R. A. Jong, R. S. Shumak, G. A. Lockwood, D. L. Tritchler, and N. F. Boyd, "Analysis of mammographic density and breast cancer risk from digitized mammograms," *Radiographics* **18**, 1587–1598 (1998).

¹³J. J. Heine, M. J. Carston, C. G. Scott, K. R. Brandt, F. F. Wu, V. S. Pankratz, T. A. Sellers, and C. M. Vachon, "An automated approach for estimation of breast density," *Cancer Epidemiol. Biomarkers Prev.* **17**, 3090–3097 (2008).

¹⁴G. Torres-Mejia, B. De Stavola, D. S. Allen, J. J. Perez-Gavilan, J. M. Ferreira, I. S. Fentiman, and I. Dos Santos Silva, "Mammographic features and subsequent risk of breast cancer: a comparison of qualitative and quantitative evaluations in the Guernsey prospective studies," *Cancer*

- Epidemiol. Biomarkers Prev.* **14**, 1052–1059 (2005).
- ¹⁵J. A. Tice, S. R. Cummings, E. Ziv, and K. Kerlikowske, “Mammographic breast density and the gail model for breast cancer risk prediction in a screening population,” *Breast Cancer Res. Treat.* **94**, 115–122 (2005).
 - ¹⁶J. Ding, R. Warren, I. Warsi, N. Day, D. Thompson, M. Brady, C. Tromans, R. Highnam, and D. Easton, “Evaluating the effectiveness of using standard mammogram form to predict breast cancer risk: case-control study,” *Cancer Epidemiol. Biomarkers Prev.* **17**, 1074–1081 (2008).
 - ¹⁷N. Boyd, L. Martin, A. Gunasekara, O. Melnichouk, G. Maudsley, C. Peressotti, M. Yaffe, and S. Minkin, “Mammographic density and breast cancer risk: evaluation of a novel method of measuring breast tissue volumes,” *Cancer Epidemiol. Biomarkers Prev.* **18**, 1754–1762 (2009).
 - ¹⁸J. Chen, D. Pee, R. Ayyagari, B. Graubard, C. Schairer, C. Byrne, J. Benichou, and M. H. Gail, “Projecting absolute invasive breast cancer risk in white women with a model that includes mammographic density,” *J. Natl. Cancer Inst.* **98**, 1215–1226 (2006).
 - ¹⁹J. J. Heine and M. Behera, “Effective x-ray attenuation measurements with full field digital mammography,” *Med. Phys.* **33**, 4350–4366 (2006).
 - ²⁰E. S. Page, “Continuous inspection schemes,” *Biometrika* **41**, 100–115 (1954).
 - ²¹E. S. Page, “Cumulative sum schemes using gauging,” *Technometrics* **4**, 97–109 (1962).
 - ²²E. S. Page, “Controlling the standard deviation by Cusums and warning lines,” *Technometrics* **5**, 307–315 (1963).
 - ²³J. D. Healy, “A note on multivariate CUSUM procedures,” *Technometrics* **29**, 409–412 (1987).
 - ²⁴R. Watkins, E. Serryn, V. Bert, W. Graeme, and P. Aileen, “Applying cusum-based methods for the detection of outbreaks of Ross River virus disease in Western Australia,” *BMC Medical Informatics and Decision Making* **8**, (2008).
 - ²⁵R. Adeyemi, “Retrospective analysis and design of Cusum control Chart for detection of outbreaks of infectious diseases,” *International Journal of Pure and Applied Sciences* **2**, 76–83 (2008).
 - ²⁶V. N. Naik, I. Devito, and S. H. Halpern, “Cusum analysis is a useful tool to assess resident proficiency at insertion of labour epidurals [L’analyse des sommes cumulees permet d’évaluer la maitrise, par les residents, de l’insertion du catheter epidural pour le travail obstetrical],” *Can. J. Anaesth.* **50**, 694–698 (2003).
 - ²⁷A. Bottle and P. Aylin, “Intelligent information: a national system for monitoring clinical performance,” *Health Serv Res.* **43**, 10–31 (2008).
 - ²⁸B. Barratt, R. Atkinson, H. Ross Anderson, S. Beevers, F. Kelly, I. Mudway, and P. Wilkinson, “Investigation into the use of the CUSUM technique in identifying changes in mean air pollution levels following introduction of a traffic management scheme,” *Atmos. Environ.* **41**, 1784–1791 (2007).
 - ²⁹J. M. Farrington, *Analysing for Authorship* (University of Wales Press, Cardiff, 1996).
 - ³⁰R. L. Brown, J. Durbin, and J. M. Evans, “Techniques for testing the constancy of regression relationships over time,” *J. R. Stat. Soc. Ser. B (Methodol.)* **37**, 149–192 (1975).
 - ³¹W. Ploberger and W. Krämer, “The CUSUM test with OLS residuals,” *Econometrica* **60**, 271–285 (1992).
 - ³²D. M. Hawkins and D. H. Olwell, *Cumulative Sum Charts and Charting for Quality Improvement* (Springer-Verlag, New York, 1997).
 - ³³NIST, “NIST/SEMATECH e-Handbook of Statistical Methods,” in *6.3.2.3. Cusum Control Charts* (2009).
 - ³⁴W. H. Woodall and M. M. Ncube, “Multivariate CUSUM quality-control procedures,” *Technometrics* **27**, 285–292 (1985).
 - ³⁵A. H. Tyson, G. E. Mawdsley, and M. J. Yaffe, “Measurement of compressed breast thickness by optical stereoscopic photogrammetry,” *Med. Phys.* **36**, 569–576 (2009).
 - ³⁶G. E. Mawdsley, A. H. Tyson, C. L. Peressotti, R. A. Jong, and M. J. Yaffe, “Accurate estimation of compressed breast thickness in mammography,” *Med. Phys.* **36**, 577–586 (2009).