# American College of Cardiology/ European Society of Cardiology International Study of Angiographic Data Compression Phase III

## Measurement of Image Quality Differences at Varying Levels of Data Compression

Rüdiger Brennecke, PHD,* Udo Bürgel, MS,* Rüdiger Simon, MD, FACC,† Gerd Rippin, MS,‡
Hans Peter Fritsch, MS,* Tim Becker, MS,† Steven E. Nissen, MD, FACC§
*Mainz and Kiel, Germany and Cleveland, Ohio*

**OBJECTIVES** We sought to investigate up to which level of Joint Photographic Experts Group (JPEG) data compression the perceived image quality and the detection of diagnostic features remain equivalent to the quality and detectability found in uncompressed coronary angiograms.

**BACKGROUND** Digital coronary angiograms represent an enormous amount of data and therefore require costly computerized communication and archiving systems. Earlier studies on the viability of medical image compression were not fully conclusive.

**METHODS** Twenty-one raters evaluated sets of 91 cine runs. Uncompressed and compressed versions of the images were presented side by side on one monitor, and image quality differences were assessed on a scale featuring six scores. In addition, the raters had to detect pre-defined clinical features. Compression ratios (CR) were 6:1, 10:1 and 16:1. Statistical evaluation was based on descriptive statistics and on the equivalence *t*-test.

**RESULTS** At the lowest CR (CR 6:1), there was already a small (15%) increase in assigning the aesthetic quality score indicating "quality difference is barely discernible—the images are equivalent." At CR 10:1 and CR 16:1, close to 10% and 55%, respectively, of the compressed images were rated to be "clearly degraded, but still adequate for clinical use" or worse. Concerning diagnostic features, at CR 10:1 and CR 16:1 the error rate was 9.6% and 13.1%, respectively, compared with 9% for the baseline error rate in uncompressed images.

**CONCLUSIONS** Compression at CR 6:1 provides equivalence with the original cine runs. If CR 16:1 were used, one would have to tolerate a significant increase in the diagnostic error rate over the baseline error rate. At CR 10:1, intermediate results were obtained. (J Am Coll Cardiol 2000; 35:1388–97) © 2000 by the American College of Cardiology

In coronary angiographic imaging, the replacement of the cine film by digital media and by computer networks is in rapid progress (1,2). In this process, it is a primary prerequisite to maintain or even surpass the image quality of the cine film. This requires the digital acquisition of high-resolution images, thus potentially resulting in costly data storage and networking systems. Lossy image compression methods reduce the amount of image data to be stored and transferred, by performing a reduction of details that are considered to be irrelevant. This reduction is achieved using digital computational techniques such as those defined by the Joint Photographic Experts Group (JPEG) in the JPEG standard (3). The irrelevancy criterion has been based on the inability of the human visual system to perceive certain details such as small compression errors at steep edges in digital television images. However, this concept does not exclude the fact that coronary angiograms being viewed on a medical workstation will suffer from a loss in subjectively perceived (aesthetic) image quality, or even from a loss in diagnostically relevant information, at higher levels of lossy compression.

Some previous studies on the viability of JPEG compression of coronary angiograms focused on the visual or quantitative detection of coronary lesions (4–7). Another group of studies asked the raters to assign image quality scores to images or cine runs shown at different levels of compression (8–11). These single-center studies varied significantly in the statistical methods applied and in the

From the *II Medical Clinic, Johannes-Gutenberg-University Mainz, Germany; the †Medical Clinic, Albrechts-University Kiel, Germany; the ‡Department of Medical Statistics, Johannes-Gutenberg-University Mainz, Germany; and the §Cleveland Clinic Foundation, Cleveland, Ohio.

Manuscript received June 1, 1999; accepted December 1, 1999.

**Abbreviations and Acronyms**

| | |
|---|---|
| ACC | = American College of Cardiology |
| CR | = compression ratio |
| DC | = display controls |
| ESC | = European Society of Cardiology |
| fps | = frames per second |
| GQ | = general image quality |
| JPEG | = Joint Photographic Experts Group (computer standard for digital images) |
| QA | = aesthetic image quality |
| QCA | = quantitative coronary angiography |
| QD | = diagnostic image quality |
| ROI | = region of interest |
| SD | = standard deviation |

selection of technical parameters such as image enhancement preceding compression (4,5) and the use of digital (4,9,10), as opposed to digitized (11), images. Moreover, Robinson (12) summarized a large number of previous studies and showed that, in many image-quality studies, the variation attributable to raters was larger than the variation proceeding from image quality. In order to avoid some of these limitations, this new study consisted of three methodologically different, but complementary, parts that were coordinated by a joint American College of Cardiology (ACC)/European Society of Cardiology (ESC) steering committee. This approach offered also a multicenter basis for the selection of raters and angiograms. Phase II investigated the quantitative effects of image compression on the results of quantitative coronary arteriography (QCA) and will be reported elsewhere (13). For the assessment of subjectively perceived image quality, two new study designs were developed. The study design of Phase I (clinical decision making) added consensus readings from an expert panel as a gold standard to the schemes of image quality assessment of earlier studies (14). For Phase III of the study described below, the consensus approach of Phase I was integrated with a simultaneous display of compressed and original images (9,15). This design offers a paired assessment of barely noticeable differences in image quality and thus eliminates most sources of rater variability.

## SUBJECTS AND METHODS

**Raters.** Twenty-one raters performed the task of image quality assessment. They came from 18 European centers in Belgium, France, Germany, Italy, The Netherlands and Sweden and from three centers in the U.S. All raters routinely perform diagnostic as well as interventional catheterizations. The mean value and standard deviation for their ages were 45 ± 8 years, and according to their self-assessment their annual volume was an average of 364 ± 181 diagnostic cases and 279 ± 92 interventional cases.

**Table 1.** Diagnostic Feature Type Within the 91 Cine Runs

| Feature | Number of Features | Number of Cine Runs With Feature |
|---|---|---|
| Filling defect | 15 | 15 |
| Dissection | 13 | 12 |
| Calcium | 54 | 30 |
| Collaterals | 8 | 7 |
| Complex lesion | 15 | 15 |
| Stent | 19 | 14 |
| Significant stenosis (>50%) | 76 | 48 |
| Mild stenosis (<50%) | 51 | 40 |
| Aneurism | 8 | 7 |

**Image selection.** The multicenter collection process for the cine runs and the definition of clinically relevant features (Table 1) by a consensus panel of experts have been described in detail in the article on Phase I (14). In the following we will use the term "images" as synonymous with "cine run." The same 100 images that had been selected for Phase I were also used in Phase III. For the assessment of compression-induced image-quality differences, Phase III presented the images in a side-by-side format that showed the relevant area (region of interest [ROI]) of each coronary angiogram simultaneously in both original and compressed formats on the same screen (dual display). The two ROIs could be presented side by side on one screen because the selected ROIs usually represented only half of the area of the original image. In nine cine runs, however, the ROI was so large or there was so much movement that is was impossible to create a dynamic dual display on one screen. This left 91 images for Phase III. Ten of these images were randomly selected for rater training. Additionally, six randomly selected uncompressed/uncompressed pairs were presented for the assessment of the raters' ability to consistently determine small quality differences. This protocol left 75 images for the main part of the study (i.e., for the assessment of quality differences in compressed/uncompressed pairs).

**Image compression, image enhancement and randomization.** Joint Photographic Experts Group image compression was performed using the default set of parameters, including the default quantization matrix (3). The digital raw images (stored without edge enhancement) were compressed at the three compression ratios (CRs) of 6:1, 10:1 and 16:1 by selecting for each image an appropriate JPEG quality factor. For the 75 images used in the main part of the study, the mean values and standard deviations (SDs) of the quality factors were 95.5 ± 1.0 at CR 6:1, 90.3 ± 1.9 at CR 10:1 and 80.8 ± 3.4 at CR 16:1. The 75 images were randomized into three image groups (A, B and C) of 25 images each. Each of the 21 raters was assigned into one of three rater groups (1, 2 and 3) according to the order of his or her inclusion into the study. Table 2 shows the assignment between the resulting three groups of raters, three

**Table 2.** Assignments Between Rater Groups

| Rater Group | CR 6:1 | CR 10:1 | CR 16:1 |
|---|---|---|---|
| 1 | A | B | C |
| 2 | B | C | A |
| 3 | C | A | B |

Assignments between rater groups (1, 2 and 3), image groups (A, B and C) and compression ratios (CR). Each rater group contains seven raters; each image group 25 images. For example, the images from image group C were presented to rater group 2 at a CR of 10:1.

**Table 3.** Definitions of Aesthetic (QA) and Diagnostic (QD) Image Quality Scores

| Quality Criteria | Score | Color |
|---|---|---|
| Difference in clinical decision making? | QD2 | ■ |
| Assessment more difficult or less certain? | QD1 | ▩ |
| One image is clearly degraded but adequate for clinical use | QA2 | ▒ |
| Quality difference is barely discernible: equivalent | QA1 | ░ |
| Quality difference is indiscernible for me | QA0 | □ |
| Compressed side better than original | ≤QA−1 | ▦ |

The last table column presents the corresponding gray pattern applied for the plots of distributions of these scores (see Fig. 1 to 4).

groups of images and three CRs. In the next step, the 75 images assigned to a rater were randomly reordered. This scheme ensured that each rater would see each compressed image only once (i.e., at only one compression level), that each of these images would be seen by the same number of raters and that the images and compression levels would be presented to the raters in different orders. Finally, each of the six additional uncompressed/uncompressed pairs mentioned above (CR 1:1) was randomly inserted twice. Merging these data with the fixed training set of 10 images resulted in a total of 97 cine runs per rater.

**Image display.** Edge enhancement was performed for all uncompressed and compressed images by computing for each pixel the mean value from a neighborhood of 5 × 5 pixels and subtracting this mean value from the unenhanced pixel value with a relative weighting of 0.7. The images were shown with one of two rates (4 and 12 frames per second [fps]) on the screen of a high luminance monitor (AWOS, Siemens Medical Systems, Forchheim, Germany). The rater was allowed to stop the cine display and move in single steps through the cine run. Two modes of display operation were available. In the first of these modes, the brightness and contrast controls had been fixed after optimization with the Society of Motion Picture and Television Engineers test pattern. This procedure was the same as in Phase I. In the second mode, raters were allowed to change these display controls (DCs). Six raters were randomly assigned to the latter mode (DC+ raters), while the remaining 15 raters used the fixed mode (DC− raters).

**Assessment of perceived image quality.** The raters were blinded regarding all properties of the images shown. They were guided through the assessment of image quality by a facilitator who was also blinded regarding the compression level of the images presented and regarding the side on which the compressed image appeared, while being informed about the consensus findings for the presence or absence of diagnostic features in each image. The facilitator recorded responses of the raters using a computerized form (Sun SPARCstation 2, Sun Microsystems; Palo Alto, California) for each image and each rater. Assessment of image quality was a two-step procedure. In the first principal step, the baseline image quality and the baseline detection rate for the diagnostic features were assessed. In this step, the rater was asked whether the general image quality (GQ) of the uncompressed image was adequate (GQ+) or inadequate (GQ−) for diagnostic work. This GQ score was based on the side of the screen that presented a better image quality.

Subsequently the rater checked which of the diagnostic features specified for Phase III of the study (Table 1) were visible on this side. These responses were entered into the form. Then, the rater was informed by the facilitator about the consensus findings for the features present in this image. If there were any differences between rater findings and consensus findings, the rater was given an option to change his or her opinion. A rater error was recorded only if the rater changed his or her opinion and agreed with the findings of the consensus panel. Thus, observed differences in findings were not automatically recorded as errors of the rater, and the reported error rate is lower than the true error rate. Error recording was image-specific, not feature-specific: if there was one change in feature detection, this was recorded as a false evaluation of the image, irrespective of other features in the same image that might have been detected correctly. Because some of the raters were not willing to discuss the often faint signs of calcification, we had to exclude this feature from the assessment of baseline variability. In the second principal step, the change in image quality attributable to compression was recorded. Accordingly, the rater was asked to assign a score to characterize the difference in perceived quality of the images (ROIs) seen on the two sides of the screen. Table 3 summarizes the definitions of the two groups of scores (aesthetically relevant differences [QAs] and diagnostically relevant differences [QDs]), and it shows for each score the corresponding graphical pattern applied in the diagrams in Results. The diagnostic scores QD1 and QD2 were assigned if one of several diagnostic features changed its appearance, even if all other features were detected correctly and easily. Note that if the rater scored the compressed ROI to be of better quality than the corresponding uncompressed side of the screen, the score QA-1 was later on assigned for statistical evaluation (irrespective of the quality score assigned by the rater).

**Statistical methods.** Overall rater response to the compression effects was assessed by applying descriptive statistics to the score distributions for aesthetic image quality (QA0 to QA2) and for diagnostic image quality (QD1, QD2) (for definitions, see Table 3).

The statistical tests focused on differences in diagnostic image quality (diagnostic scores QD1 and QD2). The dependence of the scores on general image quality (GQ+/GQ−) and on selection of display modes (DC+/DC−) was assessed by multiple logistic regression. For the GQ+/GQ− test, the independent variables were CR, rater and a binary variable indicating DC+/DC−. The interaction between GQ+/GQ− and DC+/DC− was not needed in the model. For the DC+/DC− test, the set of independent variables was CR, rater and GQ+/GQ−.

The main statistical test was the evaluation of the interrelationship between the diagnostic quality score QD2 (i.e., number of additional diagnostic errors resulting from compression) and the CR. Here, the relevant question is whether the distribution of diagnostic error rates found at a given CR is statistically equivalent to the baseline error rate distribution that was recorded during the assessment of the corresponding uncompressed images. The null hypothesis for this equivalence test (16,17) is that the two treatment means differ at least by an increment or tolerance limit "delta." Discrediting this null hypothesis proves the equivalence of the two response distributions for a given delta. In this study, delta characterizes the tolerance limit for a compression-induced increase in the diagnostic error rate over the baseline rate. The one-sided Student *t*-test for equivalence was used to generate a plot showing the significance of the test as a function of delta. From this plot, for a selected level of significance ($p = 0.05$) the corresponding delta was obtained.

## RESULTS

**Rater compliance with the quality scale.** In order to assess rater variability in these subjective image quality tests, statistical analysis was preceded by characterizing the compliance of each rater with the quality scale defined in Table 3. The test variable was the percentage of QA0 scores (i.e., "quality difference is indiscernible for me") assigned at CR 1:1 and at CR 16:1. The high-response rater was defined as an observer who assigned the score QA0 to less than 50% of the 12 image pairs with uncompressed/uncompressed ROIs (CR 1:1). The low-response rater was defined as assigning QA0 for more than 50% of the 25 compressed/uncompressed images with the highest CR (CR 16:1). It is well-documented that at this high CR a definite change in perceived image quality is usually detected in JPEG compressed angiograms (9,11). Table 4 summarizes the data on rater compliance with the quality scale. Two of the DC− raters were identified as low-response raters because they assigned QA0 scores for 92% (23/25) and 68% (17/25) of the images with CR 16:1. These two raters were excluded from the following evaluations, so that 13 of 15 raters in the DC− group, and all six raters in the DC+ group, remained. None of the raters had to be eliminated as a high-response rater (Table 4). Consequently, the analysis of rater compliance resulted in an increase of sensitivity for the following evaluations of compression effects.

**Table 4.** Rater Compliance With the Quality Scale

| | High Response Test: Percentage QA0 at CR 1:1 | Low Response Test: Percentage QA0 at CR 16:1 |
|---|---|---|
| Max. value | 100 | 92.0 |
| Min. value | 66.7 | 0 |
| Mean for DC+ raters | 87.5 | 17.3 |
| Mean for DC− raters | 83.3 | 10.0 |
| Overall mean | 86.1 | 15.2 |
| Standard deviation | 12.4 | 23.1 |

The ideal rater allocates QA0 in 100% of the images with CR 1:1 and in 0% of the images with CR 16:1.

**Secondary variables influencing compression effects.** Each of the six DC+ and 13 remaining DC− raters scored 75 compressed/uncompressed images and 12 uncompressed/uncompressed images (CR 1:1, 6:1, 10:1 or 16:1), resulting in 522 image evaluations for the DC+ and 1,131 evaluations for the DC− raters, or a total of 1,653 evaluations. For the DC+ group, 7.7% (40/522) of the evaluations scored QD1, and 4.6% (24/522) scored QD2. Multiple logistic regression showed that for the DC− raters the rates of assignment of diagnostic scores QD1 and QD2, with their mean values of 2.6% (30/1131) and 1.4% (16/1131),

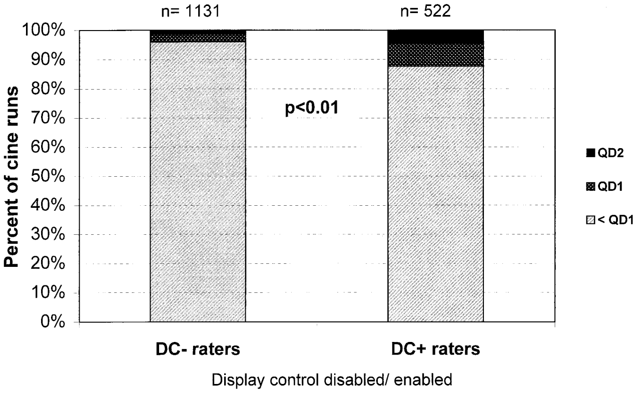

**Figure 1.** Allocation (percent of evaluations) of diagnostic quality scores (QD1, QD2; for score definition see Table 3) for DC+ (display controls enabled) and for DC− raters (display controls disabled).

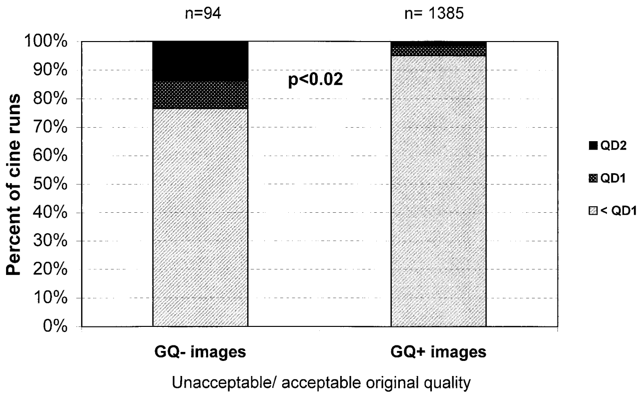

**Figure 2.** Allocation (percent of evaluations) of diagnostic quality scores (QD1, QD2; for score definition see Table 3) for GQ+ images (acceptable general image quality) and for GQ− images (general image quality not acceptable).

were significantly lower ($p < 0.01$) (see Fig. 1). Because this proved that the use of DCs tended to increase the sensitivity of the raters to adverse compression effects, all subsequent evaluations were performed separately for DC+ and DC− raters. This also improved comparability of results with Phase I of the study, which used DC− conditions exclusively.

The GQ+/GQ− score, that is the acceptability of the primary image quality of the ROI representing the original image, was not on the questionnaire for two of the raters during the starting phase of the study, reducing the total number of evaluations for this score to 1,479 (of 1,653 possible ratings). General quality was considered to be inadequate (GQ−) in 94 of these assessments. In this group of evaluations, QA scores were assigned to 76.6% (72/94) of the corresponding compressed images, QD1 was assigned to 9.6% (9/94) and QD2 to 13.8% (13/94). The corresponding numbers for the GQ+ group were 95% (1,316/1,385) with QA scores (i.e., with QA ≤ QA2), 3.3% (45/1,385) with QD1 scores and 1.7% (24/1,385) with QD2 scores. Figure 2 presents the distributions. The number of diagnostic scores assigned was significantly lower for the GQ+ group as shown by multiple logistic regression ($p < 0.02$). Therefore, lower GQ tended to increase the negative influence of lossy compression on QD.

**Distributions of image quality scores.** Figure 3 shows distributions of the scores for aesthetic image quality (QA0 to QA2) and for diagnostic image quality (QD1, QD2, see Table 3) for the 13 DC− raters (pooled for all rater groups). Each of the raters saw 12 uncompressed/uncompressed image pairs, resulting in a total of 156 evaluations for CR 1:1, and he or she saw 75 compressed/uncompressed pairs,

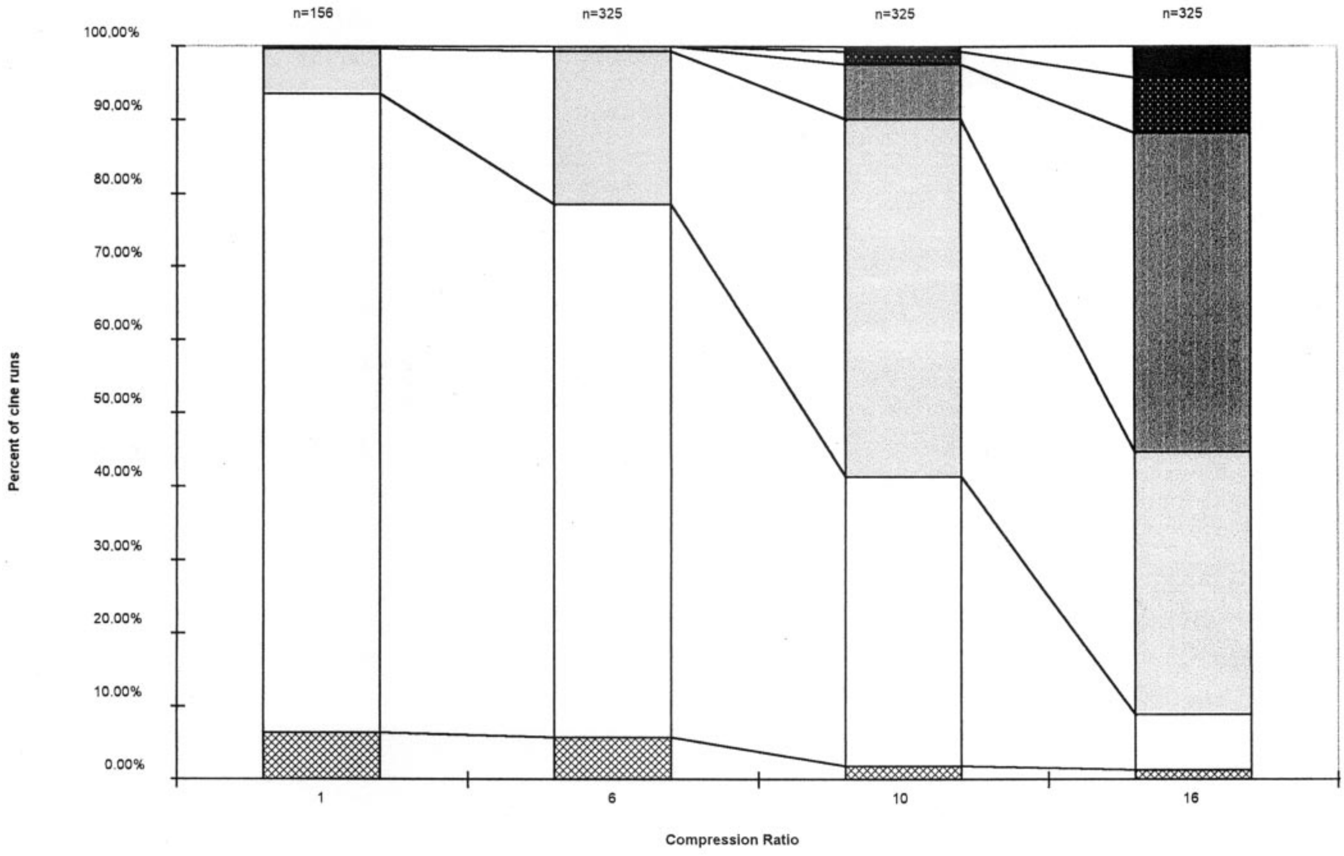

**Figure 3.** Allocation (percent of evaluations) of quality scores QA and QD in relation to compression ratios, DC− rater group (13 raters). The scores and the gray scale patterns used for the quality scores are defined in Table 3.

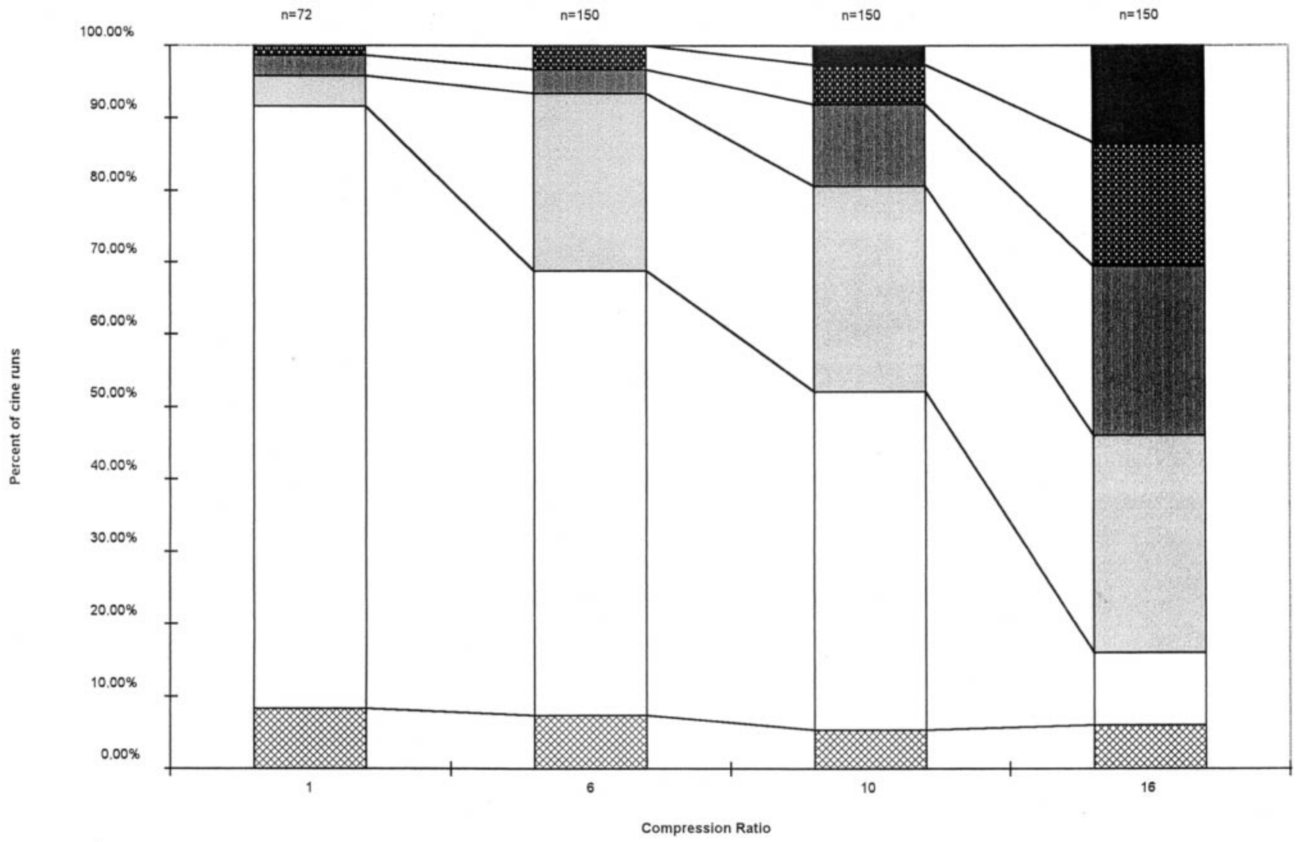

**Figure 4.** Allocation (percent of evaluations) of quality scores QA and QD in relation to compression ratios, DC+ rater group (6 raters). The scores and the gray scale patterns used for the quality scores are defined in Table 3.

resulting in a total of 325 evaluations for each of the CRs 6:1, 10:1 and 16:1.

Figure 4 presents the corresponding plot for the six DC+ raters, with a total of 72 evaluations for CR 1:1 and 150 evaluations for each of the CRs 6:1, 10:1 and 16:1. In both cases the percentages of scores representing higher-quality differences increase consistently with the increasing CR. The percentage of evaluations above the aesthetic threshold QA1 (i.e., for QA2, QD1 and QD2) are summarized in Tables 5 and 6.

**Comparison with the baseline detection rate.** The statistical significance of the compression effects on diagnostic scoring was assessed by the one-sided Student *t*-test for equivalence. This test compared the distribution of the rater-specific baseline rate of diagnostic errors with the corresponding distributions for total diagnostic errors at CR 10:1 and CR 16:1 (note that at CR 6:1 no diagnostic error resulting from compression was observed). These error distributions for the raters evaluating compressed images were obtained by merging the baseline errors (observed during the assessment of the uncompressed ROIs) with the additional QD2− errors that were reported for the ROIs showing the compressed images. Tables 7 and 8 list the means and the SDs of the error rates. The observed mean baseline error rate was 6.7% for the DC+ raters and 9% for the DC− raters.

These data were checked for their normal distribution (separately for the DC+ and for the DC− raters). This

**Table 5.** Summary of Results for the Six DC+ Raters

| Compression Ratio (CR) | ≥QA2 | ≥QD1 | QD2 |
|---|---|---|---|
| CR 1:1 | 5.6% (4/72) | 1.3% (1/72) | 0 |
| CR 6:1 | 6.7% (10/150) | 3.3% (5/150) | 0 |
| CR 10:1 | 19.3% (29/150) | 8.0% (12/150) | 2.6% (4/150) |
| CR 16:1 | 54.0% (81/150) | 30.7% (46/150) | 13.3% (20/150) |

Scoring QA2 and higher means a significant decrease in image quality.

**Table 6.** Summary of Results for the 13 DC− Raters (compare Table 5)

| Compression Ratio (CR) | ≥QA2 | ≥QD1 | QD2 |
|---|---|---|---|
| CR 1:1 | 0.6% (1/156) | 0 | 0 |
| CR 6:1 | 0.6% (2/325) | 0 | 0 |
| CR 10:1 | 9.9% (32/325) | 2.5% (8/325) | 0.6% (2/325) |
| CR 16:1 | 55.4% (180/325) | 1.7% (38/325) | 4.3% (14/325) |

**Table 7.** Comparison of Error Rates (Rater Numbers: 1 through 6)

| Rater Number | % Error at Baseline | % Total Error at CR 10:1 | % Total Error at CR 16:1 |
|---|---|---|---|
| 1 | 2.7 | 2.7 | 6.7 |
| 2 | 9.3 | 9.3 | 17.3 |
| 3 | 5.3 | 9.3 | 17.3 |
| 4 | 16.0 | 28.0 | 60.0 |
| 5 | 9.3 | 9.3 | 13.3 |
| 6 | 6.7 | 6.7 | 14.7 |
| Mean DC+ (excl. #4) | 6.7 | 7.5 | 13.9 |
| SD DC+ (excl. #4) | 2.8 | 2.9 | 4.4 |

Comparison of error rates for DC+ raters at baseline (CR 1:1) and total diagnostic error at compression ratios (CR) 6:1, 10:1 and 16:1. Rater 4 was excluded from the equivalence test as an outlier (see error rates at baseline and at CR 16). SD = standard deviation.

analysis showed that the error data from one of the raters of the DC+ group (Rater 4) were outliers caused by extreme rates for baseline errors (16%) and for total error rate at CR 16:1 (60%) (Table 7). Therefore these data were excluded from the *t*-test. Figure 5 presents the significance of the equivalence *t*-test as a function of the tolerance limit delta (see Statistical Methods) at CR 10:1 and CR 16:1 for the remaining five DC+ raters and for the 13 DC− raters. Table 9 summarizes the results of the equivalence tests. For the DC− raters at CR 10:1, for example, the error distributions measured at baseline and at CR 10:1 can be considered as equivalent at a significance level of 0.05 if one accepts a tolerance limit delta of 1.4% (i.e., an increase of the mean error rate from 9.0% [baseline] to 10.4% [CR 10:1]).

**Table 8.** Comparison of Error Rates (Rater Numbers: 7 through 21)

| Rater Number | % Error at Baseline | % Total Error at CR 10:1 | % Total Error at CR 16:1 |
|---|---|---|---|
| 7 | 10.7 | 10.7 | 10.7 |
| 8 | 12.0 | 12.0 | 12.0 |
| 9 | 2.7 | 2.7 | 2.7 |
| 10 | 9.3 | 13.3 | 9.3 |
| 11 | 1.3 | 1.3 | 9.3 |
| 12 | 8.0 | 8.0 | 16.0 |
| 13 | 6.7 | 10.7 | 18.7 |
| 14 | 6.7 | 6.7 | 22.7 |
| 15 | 6.7 | 6.7 | 6.7 |
| 16 | 8.0 | 8.0 | 8.0 |
| 17 | 2.7 | 2.7 | 2.7 |
| 18 | 4.0 | 4.0 | 8.0 |
| 19 | 22.7 | 22.7 | 30.7 |
| 20 | 10.7 | 10.7 | 10.7 |
| 21 | 10.7 | 10.7 | 10.7 |
| Mean DC− (excl. #9, 17) | 9.0 | 9.7 | 13.1 |
| StdDev DC− (excl. #9,17) | 4.8 | 4.9 | 6.6 |

Comparison of error rates for DC− raters at baseline (CR 1:1) and of total diagnostic error at compression ratios (CR) 6:1, 10:1 and 16:1. Raters 9 and 17 were excluded from the evaluations because of low rater compliance with the quality scale (compare Table 4).

## DISCUSSION

This article describes Phase III of the largest study to date on the clinical image quality attainable with lossy image data compression. The three-phase study assessed coronary angiograms that underwent compression according to the JPEG standard. It avoided limitations of earlier investigations by collecting more than 500 cine runs from systems that were manufactured by all major vendors of X-ray angiographic imaging equipment (14), by using directly digitized images without prior digital enhancement and by winning as observers more than 90 experienced angiographers and interventionalists from the U.S. and from many European countries. Moreover, all three phases evaluated the same images, but each applied its own independent methodology. The primary goal in Phase III of the study was the quantitative detection of differences in perceived (aesthetic) image quality that can be attributed to compression effects. The second goal was to find at which level of compression the diagnostic feature detection tasks could still be performed with an error rate that could be considered equivalent to the error rate found in uncompressed images.

**Reduction of rater variation.** Rater variation is a severe source of error in all studies on perceived image quality (12). Phase III attempted to detect especially subtle changes in image quality. In order to reduce the rater variations accordingly, rater training and two tests for rater consistency with the quality scale were applied. The most specific step for reduction of rater variation was the side-by-side comparison of the quality of compressed and uncompressed images that allowed the study to pose all the primary questions to the raters in terms of perceived differences between two images being viewed at the same time. This type of paired evaluation is capable of canceling most of the side effects interfering with the effects of the CR. Finally, for the diagnostic scoring tasks a consensus panel rating that had established a standard for lesion detection was applied.

**QA.** The clinical viability of lossy compression is related primarily to the correctness of diagnostic decision making, which will be discussed later, although changes in QA may also determine the acceptability of a compression method. These qualitative changes in image quality resulting from compression are presented in Figures 3 and 4. For both rater groups, the percentages of scores representing higher aesthetic quality differences increased systematically with higher CRs. For the DC− group and the lowest CR (CR 6:1), there was already a decrease of about 15% in the ratings assigned to the quality score QA0 ("quality difference is indiscernible for me"). Instead, the score QA1 ("quality difference is barely discernible—the image information is

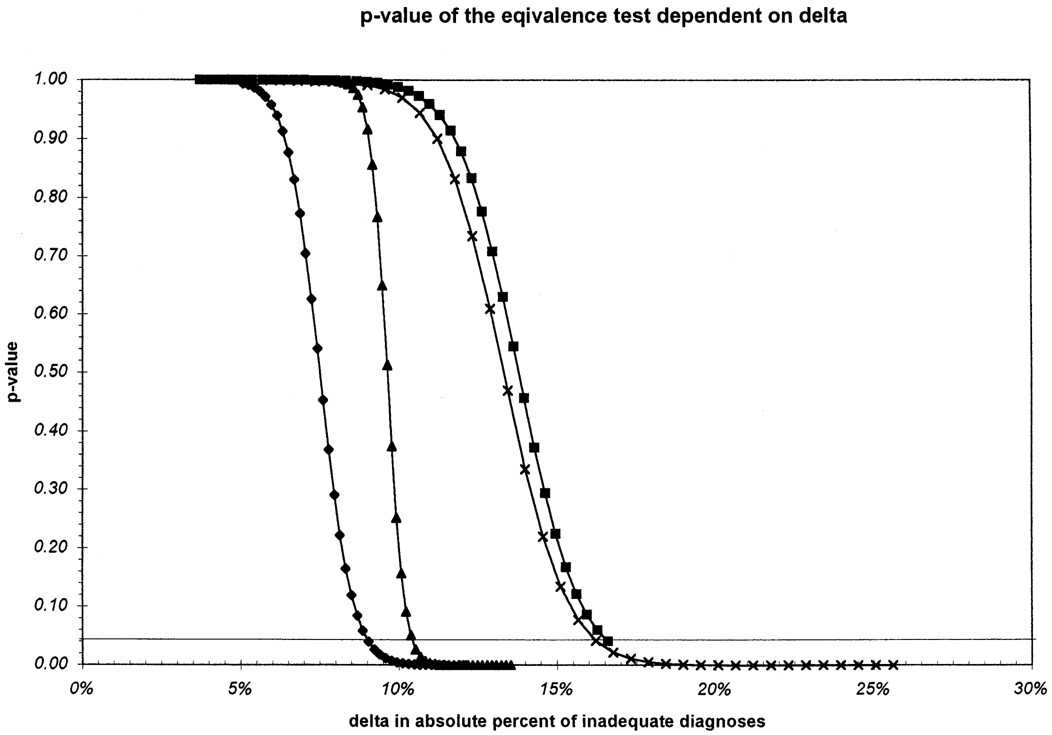

**Figure 5.** Plot demonstrating the significance (p-value) of Student *t*-test for equivalence as a function of the tolerance limit delta for the total diagnostic error in the DC+ and the DC− rater groups at CR 10:1 and CR 16:1. **Solid diamond** = DC+ for CR 10:1; **solid square** = DC+ for CR 16:1; **solid triangle** = DC− for CR 10:1; × DC− for CR 16:1. The intercepts between these functions and the horizontal reference line ($p = 0.05$) define the delta that has to be accepted if one wants to consider the error distributions at baseline (CR 1:1) and at CR 10:1 or CR 16:1 as equal with $p = 0.05$. For results compare Table 9.

equivalent") was given. Thus, lossy compression tends to degrade the QA even at the lowest CR applied, although according to the definition of QA1 the image information remains equivalent. At CR 10:1, close to 10% of the compressed images were rated to be "clearly degraded, but still adequate for clinical use" (score QA2) or worse—0.6% of these scores being already in the range of diagnostic quality changes (DC− raters). So, although we see no reason to discourage the use of images with CR 6:1, the higher rate of change in QA at CR 10:1 may already limit the range of applicable clinical scenarios for these images. Finally, the use of images with CR 16:1 is associated with a high rate (54%) of images that are clearly degraded.

**Diagnostic accuracy.** Previous compression studies (10,11) often have attempted to find a CR for which image degradation could be measured with a given statistical confidence (e.g., $p < 0.05$). This statistical test does not, however, answer the central question for a study on compression viability, because the remaining images with lower CRs cannot automatically be considered as equivalent. Therefore, this approach was avoided by applying an equivalence test (16,17).

The equivalence test is usually preceded by the explicit a priori definition of a tolerance limit delta, where delta defines the increase in error rate one is willing to accept. It turned out, however, to be impossible to obtain concrete a

**Table 9.** Diagnostic Errors in the DC+ and in the DC− Group of Raters

| | DC+ Rater Group (5 Raters) | | | DC− Rater Group (13 Raters) | | |
|---|---|---|---|---|---|---|
| CR* | Percent QD2 Scores | Percent Inadequate Evaluations* | Limit for Equivalence ($p = 0.05$) | Percent QD2 Scores | Percent Inadequate Evaluations* | Limit for Equivalence ($p = 0.05$) |
| CR 1:1 | 0 | 6.7, s.e.e.: 1.3 | n.a. | 0 | 9.0, s.e.e.: 1.4 | n.a. |
| CR 6:1 | 0 | see CR 1 | n.a. | 0 | see CR 1 | n.a. |
| CR 10:1 | 0.8 | 7.5, s.e.e.: 1.3 | 8.9 | 0.6 | 9.6, s.e.e.: 1.4 | 10.4 |
| CR 16:1 | 7.2 | 13.9, s.e.e.: 2 | 16.5 | 4.3 | 13.3, s.e.e.: 1.9 | 16.1 |

*The number of inadequate evaluations is the sum of the baseline errors plus the number of diagnostic errors (QD2) at the compression ratio indicated. The limits for equivalence given are based on Student one-sided *t*-test for equivalence, compare Figure 5.

priori values for delta from the clinical committee of the study. The requirement "avoidance of any additional feature detection error" is of course not a possible basis of an equivalence test. In order to avoid this problem, this study performed an a posteriori derivation of the parameter delta from a plot that presents the delta dependency of the significance of the statistical test (Fig. 5). The data for the DC− group of raters at CR 10:1 show that two of 325 evaluations were scored as a change in clinical decision making (the error rate increases from 9% to 9.6%). For the corresponding equivalence test at $p = 0.05$ (Fig. 5), this means that one has to consider an increase of the error rate from 9% (baseline) to 10.4% (CR 10:1) as negligible in order to be able to accept the two error distributions as equivalent. Together with the results on QA discussed in the last paragraph, this seems to represent an unambiguous basis for the decision to use or to discard the use of images with CR 10:1 in a given clinical scenario such as primary decision making or secondary review. At CR 16:1, the compression-induced increase in the diagnostic error rate was 4.3%, making this CR probably unacceptable for most clinical scenarios.

For the other rater group (DC+), changes in clinical decisions were reported at CR 10:1 at a higher rate (4/150 vs. 2/325) than in the DC− group. This has to be seen, however, in the context of a much higher variability of the data of this rater group and the group's small size. This variability is exemplified by the 5.6% of oversensitive evaluations that reported a "clearly degraded image quality" (QA2) even in original/original comparisons (Table 5) compared with 0.6% for the DC− raters. From this and similar phenomena at CR 16:1 (Fig. 4), one may infer that the DC+ raters might partly have used excessive settings of the contrast and brightness controls. Phase I of this study used DC− conditions expressly to reduce this source of rater variation. It might be advisable to supply cardiac diagnostic workstations with a digital gray scale test pattern to allow recalibration of these controls as a strategy for avoiding inappropriate contrast and brightness settings.

**Study limitations.** The study applied only one scheme of image compression, the JPEG standard, although others such as wavelet compression (15) might be advantageous. The reason was the lack of a standardized algorithm for wavelet compression. Although the JPEG quality factor is the parameter that characterizes image quality, the study instead used the JPEG CRs as independent variables for image quality assessment, thus introducing images representing a range of quality factors at a given CR. The consensus panel chose to over-represent cine runs with low GQ (without compression) and difficult clinical cases, and this must have resulted in relatively high estimates of error rates, both without and with compression (Fig. 2). The side-by-side design of Phase III of the study, while having improved reliability in the detection of changes in image quality, may also have entailed some disadvantages. Nine of the original images could not be fitted into this format, because either the ROI was too large or the movement of the vessels was too extended. Also, in the side-by-side design, the raters had to decide on the detectability of the features on the compressed side while seeing the uncompressed side. This prior knowledge available in Phase III might have introduced some bias.

**Conclusions.** The sensitive methods applied in Phase III of the compression study allowed to resolve subtle quality degradations in QA even at the lowest ratio, CR 6:1. At this compression factor, however, one can still expect equivalence with the original cine runs. If the highest ratio (CR 16:1) were used, one would have to tolerate a significant increase in diagnostic error rate. At CR 10:1, intermediate results were obtained that provide a numerical basis to decide on the applicability of compression in a given clinical scenario when combined with the results from Phases I and II (13,14). The final decision whether to use compressed coronary angiograms for certain scenarios can be made by the informed user or by a guideline panel of the ACC and the ESC.

**Reprint requests and correspondence:** Dr. Rüdiger Brennecke, II Medical Clinic, Johannes-Gutenberg-University Hospital, D55131 Mainz, Germany. E-mail: incis.rb@uni-mainz.de.

## REFERENCES

1. Simon R, Brennecke R, Hess O, Meier B, Reiner H, Zeelenberg C. Recommendations for digital imaging in angiocardiography. Eur Heart J 1994;15:1332–4.
2. ACC/ACR/NEMA Ad Hoc Group. American College of Cardiology, American College of Radiology and industry develop standard for digital transfer of angiographic images. J Am Coll Cardiol 1995;25: 800–2.
3. Pennebecker WB, Mitchell JL. JPEG Still Image Compression Standard. New York: Van Nostrand Reinhold, 1993:78.
4. Rigolin VH, Robiolio PA, Spero LA, et al. Lossy JPEG compression of digital coronary angiograms does not affect visual or quantitative assessment of coronary stenosis severity. Am J Cardiol 1996;78:131–5.
5. Baker WA, Hearne SE, Spero LA, et al. Lossy (15:1) JPEG compression of digital coronary angiograms does not limit detection of subtle morphological features. Circulation 1997;96:1157–64.
6. Koning G, van Meurs BA, Haas H, Reiber JHC. Effects of data compression on quantitative coronary measurements. Cathet Cardiovasc Diagn 1995;34:175–85.
7. Koning G, Beretta P, Zwart P, Hekking E, Reiber JHC. Effect of lossy data compression on quantitative coronary measurements. Int J Card Imaging 1997;13:261–70.
8. Kirkeeide R, Beretta P, Smalling RW, Anderson HV, Schroth G, Gould KL. Diagnostic content is unaffected by 12:1 image compression (abstr). J Am Coll Cardiol 1997;29 Suppl A:35.
9. Fritsch JP, Negwer F, Renneisen U, Brennecke R, Meyer J. Visual and quantitative analysis of coronary angiograms after irreversible data compression (abstr). Eur Heart J 1994;15 Suppl:46.
10. Karson TH, DeFranco A, Evans DJ, et al. Replacement of cine film by digital angiography: clinical evaluation of a real-time optical storage system (abstr). J Am Coll Cardiol 1993;2156A.
11. Silber S, Dörr R, Zindler G, Mühling H, Diebel T. Impact of various compression rates on interpretation of digital coronary angiograms. Int J Cardiol 1997;60:195–200.
12. Robinson PJA. Radiology's Achilles' heel: error and variation in the interpretation of the Röntgen image. Br J Radiol 1997;70:1085–98.

13. Tuinenburg JC, Koning G, Hekking, et al. American College of Cardiology/European Society of Cardiology international study of angiographia data compression phase II: the effects of varying JPEG data compression levels on the quantitative assessment of the degree of stenosis in coronary angiography. J Am Coll Cardiol 2000;35:1380–7.
14. Kerensky RA, Cusma JT, Kubilis P, et al. American College of Cardiology/European Society of Cardiology international study of angiographic data compression phase I: the effects of lossy data compression of diagnostic features in digital coronary angiography. J Am Coll Cardiol 2000;35:1370–9.
15. Goldberg MA, Pivovarov M, Mayo-Smith WW, et al. Application of wavelet compression to digitized radiographs. AJR 1994;163: 463–8.
16. Chester S. Review of equivalence and hypothesis testing. Am Statistical Ass Proc Biopharmaceutical Section 1986:177–82.
17. Ho SY, Zhu G, Zhao W. Statistical testing for equivalence. Am Statistical Ass Proc Biopharmaceutical Section 1993:240–4.