Establishing the Level of Digitization for Wrist and Hand Radiographs for the Third National Health and Nutrition Examination Survey

Yechiam Ostchega, L. Rodney Long, Gin-Hua Goh, Rosemarie Hirsch, Loralie D. Ma, William W. Scott, Jr., Wayne Johnson, and George R. Thoma

In the third National Health and Nutrition Examination Survey (NHANES III) conducted by the National Center for Health Statistics, Centers for Disease Control and Prevention, radiographs of the hands and knees were taken of participants 60 years and older as part of the study of arthritis and musculoskeletal conditions. The purpose of the study was to decide the digitizing resolution to be used for these radiographs. A set of wrist and hand radiographs ($N = 49$) was graded by two radiologists for degree of bone erosions and served as a "gold standard." The radiographs were then digitized at three resolution levels; low-resolution 150 μm ($2001 \times 1634 \times 12$ bit matrix); intermediate-resolution 100 μm ($3000 \times 2400 \times 12$ bit matrix); and high-resolution 50 μm ($4900 \times 3000 \times 12$ bit matrix). A comparison of the digital images versus the gold standard reading was made at the three resolutions by two radiologists. Kappa statistics suggested fair ($K > .4$) to excellent ($K > .75$) agreement between the gold standard and the images at all levels. Intraclass correlation coefficient suggested high agreement between readers ($\text{ICC} > .5$), with minimal individual reader effect. Variance component estimates showed that the major contribution (78-83%) to scoring came from variability in the images themselves, not from the readers. The 100 μm resolution was selected over the 150 and 50 μm on the basis of practical considerations such as storage requirements, display time, and easier manipulation of the digital images by the readers.

Copyright © 1998 by W.B. Saunders Company

KEY WORDS: NHANES III, radiograph, hand, digital resolution

AS PART of the study of arthritis and musculoskeletal conditions in the third National

From the Division of Health Examination Statistics, National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, MD; National Library of Medicine, Bethesda, MD; Century Computing, Inc, Laurel, MD; Johns Hopkins University School of Medicine, Baltimore, MD; and the Office of Research Methodology, National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, MD.

Address reprint requests to Yechiam Ostchega, PhD, RN, National Center for Health Statistics, Centers for Disease Control and Prevention, Room 900, 6525 Belcrest Rd, Hyattsville, MD 20782.

*Copyright © 1998 by W.B. Saunders Company
0897-1889/98/1103-0002\$8.00/0*

Health and Nutrition Examination Survey (NHANES III),¹ radiographs of the wrist and hand (posterior-anterior view, cassette size 24 cm $\times$ 30 cm) were taken from individuals age 60 years and older. A decision was made to digitize all NHANES III radiographs for preservation and dissemination. A critical step before digitization was to determine the level of digitizing resolution to be used for these films. The biomedical informational goal was that the resultant digital images should allow the visual detection of features relevant to assessing osteo-arthritis (OA) and rheumatoid arthritis (RA). These radiographs are typical of those obtained in most epidemiological studies of hand arthritis (for example, the Baltimore Longitudinal Study on Aging). Moreover, methods developed should be useful to deal with the radiographs of many major studies of arthritis. Two previous studies^{2,3} had digitized similar film but for different purposes. In addition, the advent of higher resolution scanners made it possible to investigate additional levels of digitization. The objective of this study was to determine the digitization level, which would allow detection of features relevant to assessing OA and RA.

MATERIAL AND METHOD

A set of wrist and hand radiographs collected from 49 subjects in an independent study conducted among the Pima Indians was used. Epidemiologic studies have shown a higher prevalence of RA in Pima Indians than in the United States white population by both the 1961 Rome RA criteria and by the more stringent 1987 ACR criteria.^{4,5} This set was graded to consensus, in film form, by two musculoskeletal radiologists for degree of bone erosions. We adopted the scoring of the radiograph films as the "gold standard" for our study. The fundamental approach was to read each of the three digitized versions of these images to consensus, and then to compare the results for each digitization level to the gold standard of plain radiograph scores.

The set of study radiographs was digitized at the three levels. The grayscale resolution used was 12 bits of grayscale per pixel. The three levels were as follows: low-resolution 150 μm (2001×1634); intermediate-resolution 100 μm (3000×2400); and high-resolution 50 μm (4900×3000). There were 49 subjects, yielding 49 left hands and 49 right hands. Each of these 98 hand images was digitized at each of the three resolution levels to yield a total of 294 total individual hand images for scoring. Using the Atlas of Standard Radiographs of Arthritis⁶ as

a reference, the two radiologists trained in standardized radiograph reading methods graded the digitized images for the prevalence of bone erosions. The grading of the digitized images was done by the same two radiologists who graded the films; there was a time interval of more than 1 year between the film and digital image grading. The following scoring scale was used: 0 = none, 1 = doubtful, 2 = minimum, 3 = moderate, and 4 = severe. The Atlas was available to the readers at all times during the reading process.

Only the digitized images were presented to the two readers in a random fashion, using independent randomizing schemes for each reader. No zooming or other image processing capability was used. All readings were made using software developed by the National Library of Medicine (NLM), and described in the Technology section below.

One reader scored the images in two work sessions separated by a 6-week interval; the second reader scored the images in three work sessions over the same interval. One work session was held to reach consensus on divergent scores; this session was held about 10 weeks after the initial scoring session. In this session, grading disagreements were read to a consensus, according to the following protocol: all "erosions absent" (grades 0-1) vs. "erosions present" (grades 2-4) grade disagreements were resolved between the readers by redisplaying the images of disagreement and reaching consensus after review and discussion. (Differences in readings within the "absence" or "presence" categories were not resolved by consensus; in these cases, the lower of the two grades was used as the final grade.)

Data Analysis

Kappa statistics for nominal scaled data were used to calculate observer agreement at each digitization level to the gold standard.⁷ Both raw grades (0-4) and dichotomized grades resolved by consensus (grades 0-1 = erosions absent; grades 2-4 = erosions present) were compared against the gold standard. Because simple kappa statistics are applicable only to dichotomized data, the ordinal scores from 0 to 4 were dichotomized. Landis and Koch⁸ have suggested that a kappa greater than .75 represents an excellent agreement beyond chance, values below .40 represent poor agreement, and values between .40 and .75 represent fair to good agreement. We used these classifications to categorize the degree of agreement between the digitized image scores and the gold standard plain radiographs scores at each digitization level.

Intraclass correlations (ICC), and variance components analysis were used to assess the agreement of raw scores between readers.^{9,10}

Technology

The images were displayed on an E-Systems Megascan monitor with a resolution of 2048×2560 pixels, using a linear function to map the 12-bit input pixels to 8-bit values of grayscale for display. A Graphical User Interface (GUI) was provided on a second (Sun 1152 $\times$ 900) monitor that was also attached to the workstation. The workstation computers were Sun SPARC 10 and SPARC 20 machines. The Megascan monitors were driven by the MD5 S-Bus image controller board from Dome Imaging. Using these workstations, the time to display the largest images was found to be approximately 2

minutes. By moving to a SPARC 10 workstation with 256 MB of RAM, this wait time was reduced to about 80 seconds.

The readings were carried out in an engineering laboratory environment, with overhead lights out, but no attempt was made to isolate or remove all reflected light sources. The protocol allowed the readers to freely alter the monitor contrast and brightness using manual controls.

RESULTS

Out of the 294 images which were read, 57 (19%) were read to a consensus. Table 1 shows the first measures of agreement between the readers and the gold standard. This table presents the kappa statistics computed using the raw and the dichotomized raw score values. For both readers the doubtful-minimum bone erosion category seems to be the most difficult. Between the two readers there is even a certain reversal of scoring agreement in these categories: for the first reader, the highest kappa for doubtful-minimum is at 50 μ m ($K = .46$ and $K = .61$); for the second reader the highest kappa for these categories is at 150 μ m ($K = .5$ and $K = .5$). Moreover, for both readers the doubtful-minimum categories represent the lowest kappa agreement with the gold standards. It seems that the readers had problems distinguishing between the doubtful and minimum categories.

In addition, Table 1 represents the raw dichoto-

Table 1. Comparison of Raw and Dichotomized Raw Scores for Right and Left Hand Digitized Images (N = 98) to Gold Standard at Three Levels of Resolution
CI = Confidence Interval

Resolutions	Reader One		Reader Two	
	kappa	95% CI	kappa	95% CI
150 μ m				
None	.65	(.49, .82)	.62	(.45, .80)
Doubtful	.25	(.02, .48)	.50	(.28, .72)
Minimum	.17	(.00, .41)	.50	(.24, .75)
Moderate	.48	(.30, .67)	.62	(.46, .79)
Severe	.59	(.32, .87)	.70	(.47, .92)
dichotomized	.67	(.53, .82)	.86	(.75, .96)
100 μ m				
None	.68	(.52, .83)	.53	(.34, .73)
Doubtful	.24	(.01, .48)	.48	(.26, .69)
Minimum	.39	(.14, .64)	.38	(.15, .61)
Moderate	.56	(.38, .74)	.71	(.56, .86)
Severe	.63	(.36, .90)	.88	(.71, 1.0)
dichotomized	.73	(.60, .87)	.77	(.65, .90)
50 μ m				
None	.65	(.50, .79)	.63	(.46, .81)
Doubtful	.46	(.22, .69)	.40	(.19, .61)
Minimum	.61	(.39, .83)	.15	(.00, .38)
Moderate	.68	(.52, .83)	.48	(.30, .66)
Severe	.78	(.55, 1.0)	.78	(.57, .99)
dichotomized	.76	(.63, .88)	.70	(.56, .84)

mized scores agreement with the gold standard; that is, the raw scores were collapsed into two categories, "erosions absent" (grades 0-1) and "erosions present" (grades 2-4); the gold standard scores were similarly collapsed. There was no apparent strong correlation shown in these results between the digitization level and reader agreement. Stated differently, the agreement was fair ($K > .4$) to excellent ($K > .75$) on all three resolutions of digitization.

Similarly, the scores resolved by consensus were dichotomized and compared to the dichotomized gold standards. The scores resolved by consensus are the raw scores, modified as necessary by the readers in the cases where the readers initially disagreed between "erosions present" and "erosions absent." The results are shown in Table 2, which suggests good to excellent agreement with the gold standard with a slight advantage for the 100 μ m level ($K = .82, (.70, .93)$).

As a final measure of reader agreement, and to understand sources of disagreement, the intraclass correlation coefficient (ICC) and the variance component estimates are shown in Table 3. The ICC for all levels was very good (near unity), suggesting high agreement between readers, with minimal individual reader effect. The variance component estimates show that the major contribution (78-83%) to overall scoring variance came from variability in the images themselves. Of interest is the fact that the highest "error" per cent (22%) was at the 50 micron level.

DISCUSSION

The National Center for Health Statistics faced two major challenges with the NHANES III radiographic skeletal images: (1) to provide for the long-term storage of the films, and (2) to provide an image retrieval and viewing system for wide dissemination of the information contained in the radiographs. Conversion of the film to a digital form was a feasible and attractive solution because computer technology could offer reasonable op-

Table 3. Intraclass Correlation Coefficients (ICC), 95% Confidence Intervals (CI), and Variance Components Estimate for the Two Readers at Three Resolutions, Based on Raw Scores

Resolutions	Estimated source of variance components & percent			
	ICC (95% CI)	Image (%)	Reader (%)	Residual (%)
150 μ m	.91 (.87, .94)	1.77 (83)	.022 (1)	.33 (16)
100 μ m	.89 (.84, .93)	1.61 (81)	.032 (2)	.33 (17)
50 μ m	.88 (.83, .92)	1.51 (78)	.005 (0)	.41 (22)

tions for permanent storage and easy access to the digitized images.

The findings of this study indicate that all three levels of digitization would provide the required digital preservation of the biomedical information of interest. Clearly, there are a number of limitations to this study. First, we based our assessments of the quality of the images on bone erosions. We chose this radiographic feature because (1) it is critical in the determination of the presence of radiographic RA (adults), and (2) other critical features for distinguishing radiographic OA and RA are much more readily discernible. A reasonable question to pose is whether we would obtain similar results if the feature of interest required finer visual discrimination as in, perhaps, bone resorption. Second, no attempt was made to read the images more than once; and, the number of images read was very small. Finally, it was suggested by the readers that a bias toward the highest resolution images might have been introduced by not having the lower resolution images zoomed to have the same spatial resolution as the highest resolution images. This hypothesis, however, did not appear to be supported by the analysis of the collected results.

The study did not identify a particular digitization level as superior for preserving the particular biomedical data of interest, namely, visual detectability of bone erosions. Therefore, we based our choice of digitization level on practical considerations, consistency with findings in related studies, and a subjective evaluation of the need to allow a prudent margin in the resolution choice to compensate for limitations of the study and for currently unforeseen requirements on the data by future researchers.

Important practical considerations include stor-

Table 2. Comparison of Dichotomized Scores Resolved by Consensus for Right and Left Hand Digitized Images (N = 98) to Gold Standard at Three Resolutions

Resolutions	kappa	95% CI
150 μ m	.78	(.65, .90)
100 μ m	.82	(.70, .93)
50 μ m	.80	(.68, .92)

Table 4. Practical Considerations in Digitizing the Images at the Three Levels

Resolution (μm)	File Size (Mbytes)	Time to Digitize (minutes)	Time to Display (sec)	Both Hands on One Image	Requires Scrolling
150	7.0	2	15	Yes	No
100	15.5	3	30	Yes	Some
50	62.0	10	60×2	No	Yes

age required, display time required, and manipulability of the digital images, ie, can both hands be digitized on one digital image, and can the entire image be viewed without scrolling. The need to scroll the image at the 50 μm level may explain the fact that the highest "error" per cent (22%) was at the 50 micron level. Shavelson and Webb¹¹ explain that in addition to random error a large "error" variance can be indicative of an interaction effect (between readers and particular images, in our case). This may explain why there was such a discrepancy between the readers for the minimum-moderate bone erosion category at 50 micron level (see Table 1): the 50 μm images were 3000×4900 pixels in size; this exceeded the size of the viewing area (2048×2560) on the monitor in both dimensions, requiring both horizontal and vertical scrolling to roam the image for viewing individual carpal, metacarpal, and phalanx structures. One reader may have been more affected than the other by the need for a high degree of manual interaction with the display device. There is also the consideration of the digitization labor itself, a one-time, but non-negligible expense; the higher the resolution, the more time is required to digitize a film. These considerations are summarized in Table 4, which shows the practical disadvantages of using the highest resolution 50 micron resolution. From the practical considerations, either the 100 (5.0 lp/mm) or 150 μm (3.3 lp/mm) levels were acceptable. In the same vein, Jonsson et al² compared film images with two levels of digital images: one level was acquired at 3.3 lp/mm, and the other at 5.0 lp/mm. The digital images were displayed on a 1024×1024 monitor and evaluated for soft tissue swelling, joint space narrowing, erosions, and particularly osteopenia. No significant differences were

found between the film and digital images for joint narrowing, erosions, or soft tissue swelling. Film was significantly better than the 3.3 lp/mm digital image for osteopenia. Finally, Sayre et al³ studied the diagnostic quality of full-frame discrete cosine transform data compression. This study required the preservation of biomedical information for resorption, which requires a high degree of visual acuity for detection. Hence, a priori, it could be argued that the digitization level (5.8 lp/mm) used in this study is adequate for NHANES III purposes. Among our digitization levels, the 100 μm images (5.0 lp/mm) approximate this level most closely.

Because it is likely that this national digital database will be created only once from this particular set of survey images, we believe that a prudent course is to use a higher resolution (within the constraints of the practical considerations listed above) than might actually suffice for the immediate goal of the present study, which was detecting bone erosions. This course of action would guard against errors in and provide additional resolution for potential future uses of NHANES III data.

Therefore, in balancing these considerations, the 50 μm resolution was eliminated on the basis of practical expense and implications for data storage and handling. The 100 μm resolution was selected over the 150 μm on the basis of consistency with the resolution in published literature, as well as prudence and an affordable margin of error beyond what the present study suggested as the minimum required resolution.

ACKNOWLEDGMENT

We greatly appreciate the help provided to us by the following individuals: Dr. H.K. Huang, Director, Laboratory for Radiology Informatics Department of Radiology, University of California, San Francisco; William C. Knowler, MD, DrPH, NIH/NIDDK, Phoenix Epidemiology and Clinical Research Branch; Peter H. Bennett, MB, FRCP, FFCM, NIH/NIDDK, Phoenix Epidemiology and Clinical Research Branch; and Stan Pillemer, MD, NIH/NIAMS.

REFERENCES

1. U.S. Department of Health and Human Services (DHHS). National Center for Health Statistics: Third National Health and Nutrition Examination Survey, 1988-94, Plan and Operations Procedures Manuals (CD-ROM). Hyattsville, MD, Centers for Disease Control and Prevention, 1996
2. Jonsson A, Borg A, Hannesson P, et al: Film-screen vs. digital radiography in rheumatoid arthritis of the hand an ROC analysis. *Acta Radiologica* 35:311-318, 1994
3. Sayre J, Aberle DR, Boechat MI, et al: Effect of data compression on diagnostic accuracy in digital hand and chest radiography. *Proc SPIE* 1653:232-240, 1992
4. Del Puente A, Knowler WC, Pettitt DJ, et al: High

incidence and prevalence of rheumatoid arthritis in Pima Indians. *Am J Epidemiol* 129:1170-1178, 1989

5. Jacobsson LTH, Knowler WC, Pillemer SR, et al: A cross-sectional and longitudinal comparison of the Rome criteria for active rheumatoid arthritis (equivalent to the American College of Rheumatology 1958 criteria) and the American College of Rheumatology 1987 criteria for rheumatoid arthritis. *Arthritis Rheum* 37:1479-1486, 1994
6. The Department of Rheumatology and Medical Illustration and Manchester Royal Infirmary and the Empire Rheumatism Council's Field Unit: *The Epidemiology of Chronic Rheumatism, Volume II*, in *Atlas of Standard Radiographs of Arthritis*. Philadelphia, PA, F.A. Davis Company, 1963
7. Fleiss JL: *Statistical Methods for Rates and Proportions*, (2nd ed). New York, NY, Wiley 1981
8. Landis JR, Koch GG: The measurement of observer agreement for categorical data. *Biometrics* 33:159-174, 1977
9. Shrout PE, Fleiss JL: Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin* 86:420-428, 1979
10. SAS Institute, Inc: *SAS/STAT User's Guide* (vol 2). Cary, NC, SAS Institute, 1990, pp 1661-1673
11. Shavelson RJ, Webb NM: *Generalizability Theory: A Primer*. Newbury Park, CA, Sage, 1991