Published in final edited form as:

Bone. 2014 October ; 67: 152–155. doi:10.1016/j.bone.2014.06.039.

Methods and Reliability of Radiographic Vertebral Fracture Detection in Older Men: The Osteoporotic Fractures in Men Study

Peggy M. Cawthon[1], **Jane Haslam**[2], **Robin Fullman**[1], **Katherine W. Peters**[1], **Dennis Black**[4], **Kristine E. Ensrud**[7], **Steven R. Cummings**[1], **Eric S. Orwoll**[5], **Elizabeth Barrett-Connor**[6], **Lynn Marshall**[5], **Peter Steiger**[2], and **John T. Schousboe**[3] **for the Osteoporotic Fractures in Men (MrOS) Research Group'**

[1]California Pacific Medical Center Research Institute

[2]Optasia Medical

[3]Park Nicollet Institute for Research and Education; Division of Health Policy and Management, University of Minnesota

[4]University of California San Francisco

[5]Oregon Health and Science University

[6]University of California San Diego

[7]University of Minnesota and Minneapolis VA Health System

Abstract

We describe the methods and reliability of radiographic vertebral fracture assessment in MrOS, a cohort of community dwelling men aged ≥65 yrs.

Lateral spine radiographs were obtained at Visit 1 (2000-2) and 4.6 years later (Visit 2). Using a workflow tool (SpineAnalyzer™, Optasia Medical), a physician reader completed semi-quantitative (SQ) scoring. Prior to SQ scoring, technicians performed "triage" to reduce physician reader workload, whereby clearly normal spine images were eliminated from SQ scoring with all levels assumed to be SQ=0 (no fracture, "triage negative"); spine images with any possible fracture or abnormality were passed to the physician reader as "triage positive" images. Using a quality assurance sample of images (n=20 participants; 8 with baseline only and 12 with baseline and follow-up images) read multiple times, we calculated intra-reader kappa statistics and percent

Corresponding author: Peggy M Cawthon, 185 Berry Street, Suite 5700, San Francisco, CA 94107, Phone 415 600 7426, pcawthon@sfcc-cpmc.net.

Disclosure: Optasia Medical provided the Spine Analyzer software without charge through a cooperative agreement with California Pacific Medical Center.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

agreement for SQ scores. A subset of 494 participants' images were read regardless of triage classification to calculate the specificity and sensitivity of triage.

Technically adequate images were available for 5958 of 5994 participants at Visit 1, and 4399 of 4423 participants at Visit 2. Triage identified 3215 (53.9%) participants with radiographs that required further evaluation by the physician reader. For prevalent fractures at Visit 1 (SQ≥1), intra-reader kappa statistics ranged from 0.79-0.92; percent agreement ranged from 96.9%-98.9%; sensitivity of the triage was 96.8% and specificity of triage was 46.3%.

In conclusion, SQ scoring had excellent intra-rater reliability in our study. The triage process reduces expert reader workload without hindering the ability to identify vertebral fractures.

1. Introduction

Vertebral fractures are the most common of all osteoporotic fractures, and are considered the hallmark of osteoporosis.[1] The importance of vertebral fractures as an independent risk factor for future osteoporotic fractures is well understood.[2-8] Age, bone mineral density and prevalent vertebral fracture are the strongest predictors of future fractures.[2,3,5] Even mild fractures can have serious consequences, but are easily overlooked.[2,5,6] Measurement of vertebral deformity provides invaluable information and helps score the severity of fracture.

Several methods of describing osteoporotic fractures in lateral spine images have been developed.[9] Methods include semi-quantitative (SQ) techniques that involve a degree of subjective judgment by a trained expert radiologist.[10,11] The Genant SQ method[10] has become the standard for fracture assessment,[12] but there still remains significant subjectivity, particularly for mild fractures and relatively sparse data for its use in men. The assessment of fracture is time consuming; therefore, to reduce workload, we implemented a “triage” method whereby grossly normal spine images were eliminated from SQ scoring with all levels assumed to be SQ=0 (not fractured).

Thus, the aim of this paper was to describe the methods and reliability of radiographic vertebral fracture assessment in the Osteoporotic Fractures in Men (MrOS) study, a cohort of community dwelling men aged 65 years and older. We evaluated the intra-reader reliability of the SQ method and determined the sensitivity and specificity of the “triage” process.

2. Methods

2.1 Study description and participants

MrOS is a study of 5,994 men initially recruited at six clinical centers in the United States between 2000-2002. Descriptions of the cohort have been published elsewhere.[13,14] Briefly, to be eligible to participate, community-dwelling men must have been aged≥65 years, ambulatory, and not have had bilateral hip replacements. At baseline and Visit 2 (an average of 4.6±0.4SD years after baseline) participants provided medical information and had lateral thoracic and lumbar radiographs taken. As previously described, height, weight, body mass index, and bone mineral density at the spine, total hip and femoral neck were measured; t-scores were calculated using female normative values.[15]

2.2 Lateral spine radiograph acquisition

Lateral thoracic and lumbar spine radiographs were taken with a tube-to-film distance of 40 inches, using a breathing technique, with thoracic films centered at T7 and lumbar films centered at L3. At the baseline exam, all participants had conventional (film-based) radiographs of the lumbar and thoracic spine; at Visit 2, four clinical centers acquired conventional radiographs and two clinical centers acquired the images digitally. All conventionally acquired film images were scanned to digital format; all images were read digitally. At baseline all 5994 men had thoracic and lumbar radiographs of which 5958 were technically adequate. At Visit 2, 4,423 men provided both a lumbar and thoracic radiograph of which 4,399 were technically adequate. Participants who did not have radiographs taken at Visit 2 had either died (N=571) or terminated study participation (N=85) before the visit; were living but did not attend any part of Visit 2 (N=109); or only provided questionnaire data at the visit (N=806).

2.3 Vertebral fracture assessment: Overview

The general process for review of spine images in MrOS is as follows. First, all spine images were assessed for quality and underwent a “triage process” by trained technicians, the purpose of which was to eliminate grossly normal images from SQ assessment, thereby reducing the number of images that needed to be read by the physician reader and limiting costs. Once triage was complete, all films from participants with a possible fracture or abnormality (“triage positive”) were evaluated by a physician reader using the SQ method of Genant[10] using a software workflow tool (SpineAnalyzer™, Version 3.2, Optasia Medical Ltd., Cheadle, UK).

2.4 Triage: Detailed Workflow

Paired (baseline and follow-up) participant radiographs were assigned to one of the following three categories based on review by a technician:[3]

1. normal
2. uncertain (one or more possible deformities, anatomical problems or marginal quality film)
3. likely presence of at least one vertebral deformity (SQ≥1) at either baseline or follow-up

The triage classification was performed using all available images. (That is, Visit 1 and 2 images were reviewed unless Visit 2 images were not obtained; otherwise only Visit 1 images were used.) The classification was performed by two radiology technicians who had been previously trained and certified by a radiologist.[3] Men with all images falling into category 1 (no vertebral deformities at any visit, “triage negative”) were classified as not fractured, and their radiographs were not further assessed for SQ scoring. Participants with any images classified into categories 2 or 3 were considered “triage positive”, and had SQ scoring completed on all available images by the physician reader. To evaluate the process of triage, 494 participants were selected at random, and the physician reader completed SQ scoring on these images regardless of triage status.

2.6 Triage: Sensitivity and Specificity

We calculated the sensitivity and specificity of the triage process using the random sample of 494 participants. We used the participant as the unit of analysis (rather than the x-ray or vertebrae). This is because all of a single participants' x-ray images (up to four images) were designated either “triage positive” or “triage negative” as they were reviewed at the same time.

2.5 Vertebral Fracture Assessment: Detailed Workflow

All “triage positive” images were read using SpineAnalyzer™ Software Version 3.2 (Optasia Medical Ltd., United Kingdom). SpineAnalyzer™ is a semi-automatic clinical workflow tool which facilitates the visual or quantitative assessment of vertebral body deformities between vertebral levels T4 and L4 in lateral DXA or digital/digitized x-ray images. The software may also be used to record Genant SQ scores based on visual assessment.

In MrOS, the software was used as a workflow tool to visualize the image, assign vertebral levels and evaluate for SQ as would be done on any lateral spine image. No automated readings of SQ, from SpineAnalyzer™ or elsewhere, were used in this study. Although not evaluated herein, SpineAnalyzer™ may be also used by qualified medical professionals to annotate full vertebral outlines and standard 6-point morphometry for each vertebral level, from which vertebral heights and deformity ratios can be calculated. However, in MrOS, the reliability of the vertebral outlines and 6 point morphometry has not yet been evaluated.

The general workflow for the software as employed in MrOS is as follows:

1. Load and display digitized subject images with baseline and follow-up images positioned side-by-side on the display.
2. Place a single point at the approximate center of each vertebra to be analyzed.
3. Label vertebral levels to initiate automatic vertebral outline search.
4. Review and if necessary, manually correct vertebral outlines and morphometry points. This step records 6-point and 95-point quantitative morphometry data for each vertebral level in the study database.
5. The physician reader records SQ score for each vertebra or reason why SQ score cannot be assessed.

Longitudinal images were reviewed together. Vertebral levels were read on the same image type (lumbar or thoracic) at baseline and follow-up. For example, if L1 could not be visualized on the lumbar image at baseline, it would be read on the thoracic image at both baseline and follow-up.

2.6 Semi-quantitative (SQ) scoring of deformity: Detailed Workflow

After identifying each vertebral level, a single physician reader (JTS) classified each vertebra using the Genant SQ method in which the vertebra is rated according to the severity of loss of height and other qualitative features.[3,10] This method was applied as has been extensively described previously,[10] with the exception that mild fractures (with SQ score=1) required the presence of depression of most of the endplate. The reason for this is to distinguish wedged vertebrae from non-fracture deformities (such as short vertebral height) and degenerative changes, the presence of which lower the reliability of SQ readings.[16] Non-osteoporotic deformities were recorded, such as congenital anomalies, Schuermann's disease, or degenerative disease according to study protocol. Occasionally, these non-osteoporotic deformities made it challenging to determine if a vertebral fracture was present or not. These conditions did not prevent the physician reader from providing an SQ score, as long as the endplates and cortical margins were distinct. For example, a vertebral level with a wedge shape with adjacent degenerative disc disease and anterior osteophytes would nonetheless be read as fractured, if there was also endplate depression and/or clear buckling of the cortex. Infrequently, individual vertebra levels could not be scored because of poor image quality (level not entirely within radiograph, indistinct endplates or cortical margins).

2.7 SQ: Prevalence and Reliability

We report prevalence SQ score and non-osteoporotic deformities. Chi-square tests were used to compare the prevalence of SQ score ≥1 by presence of non-osteoporotic deformity.

To assess reliability of the SQ scoring method, a convenience sample for a quality assurance (QA) reading set (N=20) with a high proportion of fractures was used. Of the participants in this dataset, 8 contributed baseline images (lumbar and thoracic) and 12 contributed both baseline and follow-up images (2 lumbar and 2 thoracic images) for a total of 64 images. In all, these 64 images represented a maximum of 416 vertebrae. In the final MrOS dataset, the distribution of SQ scores in these vertebrae was as follows: 14 unreadable, 370 with SQ = 0; 4 with SQ = 1; 22 with SQ = 2 and 6 with SQ = 3. Intra-reader reliability of SQ scoring was assessed using three repeat reads of this dataset by a single reader (JTS). The reads were performed on three occasions: at study inception, during an interim reading and at study close out, with a period of approximately 12 months in between each read. Image order was not re-randomized for each repeat read. The physician reader was aware that he was reading a QA set of images, but was unable to access data from previous readings to compare or change results.

Kappa statistics were calculated for each pairing of repeat reads (Repeat 1 vs. Repeat 2, Repeat 1 vs. Repeat 3 and Repeat 2 vs. Repeat 3) for two outcomes: a mild or worse fracture (SQ≥1) vs. normal (SQ=0); and moderate or severe fracture (SQ≥ 2) vs. normal or mild (SQ=0,1). Each vertebral level contributed one record to the dataset for the reliability analyses (that is, there was one record per vertebrae per participant at each visit). Kappa values of 0.75 or greater are considered to represent excellent reproducibility.[17]

3. Results

3.1 Participant characteristics and prevalence and incidence of fractures

Among those with technically adequate radiographs at baseline (n=5958), average age was 73.6 ± 5.9 SD years. On average, men tended to be overweight (Table 1).

At baseline, 689 men (11.6%) had at least one mild or worse prevalent vertebral fracture (SQ≥1). Of these, 448 men (7.5%) had at least one fracture with an SQ score≥2; and 132 men (2.2%) and at least one fracture with an SQ score>3. The number of moderate or severe (SQ>=2) fractures per participant was as follows: 5510 (92.5%) had no moderate or severe fractures; 361 (6.1%) had only 1 moderate or severe fracture; 87 (1.5%) had 2 or more moderate or severe fractures. Only 2.4% of participants (N=145) had at least one vertebral level that could not be evaluated at baseline.

At baseline, many participants (48.1%, N=2855) had a non-osteoporotic deformity; mostly degenerative changes (N=2818, 47.3%) with fewer men with congenital anomalies (N=14, 0.2%), Scheurmann's disease (N=32, 0.5%) or other conditions (N=102, 1.7%). When analyzed as one-record per vertebral level per participant, of the 77,221 levels evaluated, 60,415 (78.2%) levels had a SQ score = 0 and did not have a non-osteoporotic deformity; 621 (0.8%) had a SQ score≥1 and did not have a non-osteoporotic deformity; 15830 (20.5%) had a SQ score = 0 and but had a non-osteoporotic deformity; and 355 (0.5%) had both a non-osteoporotic deformity and an SQ score ≥1. The prevalence of fracture (SQ score≥1) was higher in those levels with (2.2%) compared to without (1.0%) a non-osteoporotic deformity ($p<0.05$).

Among 4397 participants with baseline and follow-up images, 197 (4.5%) had an incident fracture (an increase in SQ score of ≥1, a new or worsening fracture) between visits. Only 1.8% of participants (N=81) had at least one vertebral level that could not be evaluated at follow-up.

3.2 Reliability of technician triage

Triage identified 3215 (53.9%) participants with radiographs that required further evaluation by SQ (i.e., “triage positive”). The sensitivity of the triage process was excellent, as assessed in a random sample of 494 participants. At Visit 1, for prevalent fractures of SQ≥1, the sensitivity of triage was 96.8% and the specificity was 46.3%. Thus, there were very few participants with fractures that were undetected (N=3,false negatives) while many images that were “triage positive” were not confirmed as fractures (N=214, false positives). At Visit 1, for prevalent fractures with an SQ≥2, the sensitivity of triage was 98.3% and the specificity was 42.9%. Similar results were seen for Visit 2 prevalent fractures (SQ≥1: sensitivity: 97.3%, specificity: 43.8%; SQ≥2: sensitivity: 97.9%, specificity: 40.3%).

3.3 Reliability of SQ scoring

The range of kappa scores for intra-reader agreement for SQ=0 vs. SQ≥1 (shaded cells) and for SQ=0 or 1 vs. SQ≥2 are reported in Table 2.

4. Discussion

The sensitivity and specificity of the triage approach has not previously been reported for men. The triage process used in this study had excellent sensitivity while reducing the workload of the physician reader by almost half. Although the specificity of triage was poor, this is acceptable, as the decision rules for triage were established to eliminate as many false negatives as possible (that is, missing fractures that should have been deemed triage positive) at the potential expense of increasing false positives (identifying images “triage positive” when in fact no fracture exists).

The reliability of SQ scoring in this study was excellent and better than or similar to other reports. The original Genant description of the SQ method showed an intra-reader kappa of 0.93 and percent agreement of 98% for an experienced reader and a kappa of 0.76 and percent agreement of 98% for an inexperienced reader.[10] Subsequent studies have reported a similar range of values (Grados et al.,[18] kappa: 0.95, percent agreement: 95%; Wu et al.,[19] kappa: 0.80-0.81, percent agreement; 94.2%-94.4%) which are similar to our range of kappa values (0.79-0.92) and percent agreement (96.9%-98.9%). Kappa statistics vary based on the underlying distribution of fractures, and since the prevalence of fractures varied across studies, direct comparison of the kappa scores across studies is limited.

We found that, by vertebral level, the prevalence of mild or worse fracture (SQ score ≥1) was greater in the presence of non-osteoporotic deformities. Thus, whenever possible, vertebral levels should be evaluated for SQ regardless of the presence of non-osteoporotic deformities. The reasons for the higher prevalence of vertebral fracture in levels with non-osteoporotic deformities are unclear as this could reflect true differences in fracture prevalence by presence or absence of non-osteoporotic deformities or it could reflect measurement error by non-osteoporotic fracture status; this area will be the subject of future analyses.

This study represents the largest cohort of community dwelling men in a non-referral population with state-of-the-art assessment of vertebral fracture status on lateral spine radiographs. However, some limitations must be noted. First, MrOS is a cohort of older, mostly white, community-dwelling men. The reliability and precision of these measures may be different in women, other races, and the institutionalized or infirm, as the prevalence of vertebral fractures and non-osteoporotic deformities that influence assessment of the vertebral body are likely to be different in these other populations. Second, our sample for repeat assessment was relatively small. Additionally, while particularly useful to research studies, these results may have limited applicability to clinical settings, as our study used a single reader with extensive specific training for assessment of vertebral fractures. Third, we used a small subset of participants with a high prevalence of moderate and severe fractures to evaluate the reliability of the physician reader. The higher number of moderate fractures and fewer number of mild fractures in this subset may have artificially inflated the degree of agreement as the mild fractures traditionally have the most disagreement.

5. Conclusions

We show excellent reliability for SQ scoring and demonstrate that the triage process reduces physician reader workload while maintaining the ability to accurately identify fractures, including mild fractures.

Acknowledgments

The Osteoporotic Fractures in Men (MrOS) Study is supported by National Institutes of Health funding. The following institutes provide support: the National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS), the National Institute on Aging (NIA), the National Center for Research Resources (NCRR), and NIH

Roadmap for Medical Research under the following grant numbers: U01 AR45580, U01 AR45614, U01 AR45632, U01 AR45647, U01 AR45654, U01 AR45583, U01 AG18197, U01 AG027810, and UL1 TR000128.

References

1. Cummings SR, Melton LJ. Epidemiology and outcomes of osteoporotic fractures. Lancet. 2002; 359(9319):1761–1767. [PubMed: 12049882]
2. Ross PD, Davis JW, Epstein RS, Wasnich RD. Pre-existing fractures and bone mass predict vertebral fracture incidence in women. Ann Intern Med. 1991; 114:919–923. [PubMed: 2024857]
3. Black DM, Palermo L, Nevitt MC, et al. Comparison of methods for defining prevalent vertebral deformities: the Study of Osteoporotic Fractures. J Bone Miner Res. Jun.1995 10:890–902. 1995. [PubMed: 7572313]
4. Burger H, van Daele PL, Algra D, et al. Vertebral deformities as predictors of non-vertebral fractures. Bmj. 1994; 309(6960):991–992. [PubMed: 7950721]
5. Ross PD, Genant HK, Davis JW, Miller PD, Wasnich RD. Predicting vertebral fracture incidence from prevalent fractures and bone density among non-black, osteoporotic women. Osteoporos Int. 1993; 3:120–126. 1993. [PubMed: 8481587]
6. Delmas PD, Genant HK, Crans GG, et al. Severity of prevalent vertebral fractures and the risk of subsequent vertebral and nonvertebral fractures: results from the MORE trial. Bone. Oct; 2003 33(4):522–532. [PubMed: 14555255]
7. Kanis JA, Johnell O, Oden A, et al. The risk and burden of vertebral fractures in Sweden. Osteoporos Int. Jan; 2004 15(1):20–26. [PubMed: 14593450]
8. Lindsay R, Silverman SL, Cooper C, et al. Risk of new vertebral fracture in the year following a fracture. JAMA. Jan 17; 2001 285(3):320–323. [PubMed: 11176842]
9. Guermazi A, Mohr A, Grigorian M, Taouli B, Genant HK. Identification of vertebral fractures in osteoporosis. Semin Musculoskelet Radiol. Sep; 2002 6(3):241–252. [PubMed: 12541202]
10. Genant HK, Wu CY, van Kuijk C, Nevitt MC. Vertebral fracture assessment using a semiquantitative technique. J Bone Miner Res. 1993; 8(9):1137–1148. [PubMed: 8237484]
11. Jiang G, Eastell R, Barrington NA, Ferrar L. Comparison of methods for the visual identification of prevalent vertebral fracture in osteoporosis. Osteoporos Int. Nov; 2004 15(11):887–896. [PubMed: 15071725]
12. Schousboe JT, Vokes T, Broy SB, et al. Vertebral Fracture Assessment: The 2007 ISCD Official Positions. J Clin Densitom. Jan-Mar;2008 11(1):92–108. [PubMed: 18442755]
13. Blank JB, Cawthon PM, Carrion-Petersen ML, et al. Overview of recruitment for the osteoporotic fractures in men study (MrOS). Contemporary clinical trials. Oct; 2005 26(5):557–568. [PubMed: 16085466]
14. Orwoll E, Blank JB, Barrett-Connor E, et al. Design and baseline characteristics of the osteoporotic fractures in men (MrOS) study--a large observational study of the determinants of fracture in older men. Contemporary clinical trials. Oct; 2005 26(5):569–585. [PubMed: 16084776]
15. Looker AC, Wahner HW, Dunn WL, et al. Updated data on proximal femur bone mineral levels of US adults. Osteoporos Int. 1998; 8(5):468–489. [PubMed: 9850356]
16. Schousboe JT, Debold CR. Reliability and accuracy of vertebral fracture assessment with densitometry compared to radiography in clinical practice. Osteoporos Int. Feb; 2006 17(2):281–289. [PubMed: 16172798]
17. Rosner, B. Fundamentals of Biostatistics. 2nd. Boston, MA: Duxbury Press; 1986.
18. Grados F, Roux C, de Vernejoul MC, Utard G, Sebert JL, Fardellone P. Comparison of four morphometric definitions and a semiquantitative consensus reading for assessing prevalent vertebral fractures. Osteoporos Int. 2001; 12(9):716–722. [PubMed: 11605736]
19. Wu CY, Li J, Jergas M, Genant HK. Comparison of semiquantitative and quantitative techniques for the assessment of prevalent and incident vertebral fractures. Osteoporosis Int. 1995; 5:354–370.

Table 1
Characteristics of MrOS men with radiographic vertebral fracture assessment

Characteristics	**Mean (SD) or N (%)**
Age (years)	73.6 (5.9)
Weight (kg)	83.1 (13.3)
Height (cm)	174.1 (6.8)
BMI (kg/m^2)	27.4 (3.8)
Lumbar spine BMD (g/cm^2)	1.170 (0.254)
Total hip BMD (g/cm^2)	0.958 (0.140)
Total hip t-score	0.13 (1.15)
Femoral neck BMD (g/cm^2)	0.784 (0.128)
Femoral neck t-score	-0.62 (1.07)
At least one prevalent vertebral fracture, SQ≥1	689 (11.6%)
At least one prevalent vertebral fracture, SQ≥2	448 (7.5%)
At least one incident fracture, change in SQ≥1 between Visit 1 and Visit 2	197 (4.5%)

Table 2

Kappa statistic (95% confidence interval) and percent agreement for three sets of intra-rater evaluations

		No fracture (SQ=0) vs mild or worse fracture (SQ≥1)		
		Reading 1	**Reading 2**	**Reading 3**
None or mild fracture (SQ=0,1) vs moderate or worse fracture (SQ≥2)	Reading 1	--	0.93 (0.86, 1.00) 98.9%	0.82 (0.71, 0.94) 97.5%
	Reading 2	0.92 (0.84, 1.00) 98.9%	--	0.79 (0.67, 0.91) 96.9%
	Reading 3	0.81 (0.67, 0.94) 97.8%	0.81 (0.67, 0.94) 97.8%	--