Published in final edited form as:

Bone. 2009 September ; 45(3): 560–567. doi:10.1016/j.bone.2009.04.250.

Bone fracture risk estimation based on image similarity

Wenjun Li¹, John Kornak^{1,2}, Tamara B. Harris³, Joyce Keyak⁴, Caixia Li¹, Ying Lu^{1,2}, Xiaoguang Cheng⁵, and Thomas Lang¹

¹Department of Radiology and Biomedical Imaging, University of California, San Francisco, San Francisco, CA 94143, USA

²Department of Epidemiology and Biostatistics, University of California, San Francisco, CA, USA

³Intramural Research Program, Laboratory of Epidemiology, Demography and Biometry, National Institute on Aging, Bethesda, MD 20892, USA

⁴Department of Orthopaedic Surgery, University of California, Irvine, Irvine, CA 92697, USA

⁵Department of Radiology, Beijing Ji Shui Tan Hospital, Beijing, China

Abstract

We propose a fracture risk estimation technique based on image similarity. We employ image similarity indices to determine how images are similar to each other in their 3D bone mineral density distributions. Our premise for fracture risk estimation is that if a given scan is more similar to scans of subjects known to have fractures than to scans of control subjects, this subject is likely to have a higher degree of fracture risk. To test this hypothesis, we analyzed hip QCT scans of 37 patients with hip fractures and 38 age-matched controls. We divided the scans randomly into two groups: the Model Group and the Test Group. For each scan in the the Test Group, the difference between the mean value of its image similarities to the Model fracture group and the mean value of its image similarities to the Model control group was used as index of fracture risk. We then used the estimated fracture risk indices to discriminate the fractured patients and controls in the Test Group. A test scan with a larger mean value of image similarities with respect to the Model fracture group was classified as a scan from a fractured patient, otherwise it was classified as a scan from a control subject. Based on ROC analysis, we compared the discrimination performances using image similarity measures with that obtained by using bone mineral density (BMD). When using BMD measured in the femoral neck, with the optimal BMD cutoff, the sensitivity and specificity were 86.5% and 73.7%. For the image similarity measures, the sensitivity ranged between 86.5% and 100%, and specificity ranged between 63.2% and 76.3%. By combining BMD with image similarity measures, the sensitivity and specificity reached 94.6% and 76.3% using linear discriminant analysis (LDA) algorithm, or 91.9% and 81.6% using recursive partitioning and regression trees (RPART) algorithm. In the RPART approach, the AUC value of the ROC curve was 0.923, higher than the AUC value of 0.835 when using BMD alone (p-value: 0.0046). Our results showed that combining BMD with image similarity measures resulted in improved hip fracture risk estimation.

Keywords

osteoporosis; proximal femur; QCT; mutual information; image registration

© 2009 Elsevier Inc. All rights reserved.

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.

INTRODUCTION

Skeletal fracture is the most serious clinical outcome associated with osteoporosis in the elderly [4,22]. BMD measurements based on dual x-ray absorptiometry (DXA)[25] or volumetric quantitative computed tomography (vQCT) [15,16] have been widely used for osteoporosis diagnosis and treatment monitoring[1,2]. However, BMD alone cannot adequately identify subjects at high risk of fracture [26,27]. BMD measures the average mineralization in a specified region of interest (ROI) such as the femoral neck, the trochanter, or the entire proximal femur. However, in reality, a fracture event is a mechanical failure stemming from the inability of the whole bone structure to withstand the forces exerted in a fall. In this study, we propose and test a hip fracture risk estimation technique based on measures of “image similarity”. Our fracture risk estimation is based on the idea that if a given scan is more similar to scans of fractured patients than scans of control subjects, this subject is likely to have a high degree of fracture risk. To test this hypothesis, we have retrospectively analyzed a cohort of Chinese women described in previous publications[5,13,18]. We compared fracture risk estimation efficacy based on image similarity measures to that obtained by BMD. We also tested whether combining BMD and image similarity measures resulted in improved fracture risk estimation.

MATERIALS AND METHODS

Thirty-seven women with fractures of the hip (twenty-nine patients with neck fractures and eight patients with trochanteric fractures) aged 65 or older (age: 74.4 ± 5.5 years, weight: 59.2 ± 10.4 kg, height: 159.5 ± 4.8 cm, body mass index 23.3 ± 3.8 kg/m²) were recruited from the emergency room, Department of Traumatology and Orthopedic Surgery, Beijing Ji Shui Tan hospital. In order to minimize changes in BMD and body composition factors due to the fracture, only those subjects whose fractures had occurred within the last 48 hours were accepted into the study. Hip fracture subjects were referred to the emergency room CT scanning service and to the study by their orthopaedic surgeons. If the patient was female and over 65 years old, then the radiologist explained the study to the subject and asked the subject if she was willing to participate in the study. If the subject agreed to participate, she was asked to sign the consent form, fill out a questionnaire about the circumstances of the fall and other information regarding skeletal and other health conditions. Subjects were included in the study if they had experienced a hip fracture that resulted from a low-energy fall from a standing or sitting height, and prior to the fracture, were fully ambulatory, community-dwelling adults. Subjects were excluded from the study if they reported having the following medical conditions: previous fracture, history of metastatic disease, hyper- or hypoparathyroidism, Cushing's syndrome, diabetes or collagenosis. Subjects were excluded from the study if they had undergone treatment with corticosteroids or osteoporosis medications including hormone replacement therapy or bisphosphonates. All scanning of fractured subjects was carried out prior to fixation of the fracture. In addition to these subjects, thirty-eight women over 65 years old in good health (age: 70.6 ± 5.0 years, weight: 58.8 ± 9.6 kg, height: 157.9 ± 5.0 cm, body mass index 23.5 ± 3.4 kg/m²) and with no conditions affecting bone metabolism (established from the same questionnaire) were invited from the surrounding community to participate in the study. The healthy controls contacted the hospital in response to informational briefings held at two local senior community centers. If they fit the inclusion and exclusion criteria of the study, they were invited to participate in the study and upon their arrival at the hospital, provided informed consent. On average, the control subjects were approximately 4 years younger than the fractured subjects (p-value: 0.003). The two groups were not significantly different in weight, height and body mass index. The study was reviewed and approved by the Internal Review Boards of the Beijing Ji Shui Tan hospital and the University of California, San Francisco. Informed Consents were obtained for all study participants.

All QCT acquisitions utilized a GE CT Pro FII CT scanner (GE Medical Systems, Beijing, China). Subjects were positioned supine on the CT table. An Image Analysis QCT calibration phantom (Image Analysis, Columbia KY, USA) was placed under the subject between the hips. The superior aspect of the helical scan was 5 mm above the acetabulum and the inferior limit 5 mm below the lesser trochanter. Scan parameters were 3 mm section thickness (pitch = 1), 120 kVp, 200 mAs, with an in-plane pixel size of 0.84 mm. The CT images were archived to DICOM CD and forwarded to University of California, San Francisco for analysis. The CT grayscale values were represented in Hounsfield unit. In general, the left hip was analyzed. For subjects with hip fractures, the software analyzed the contralateral unfractured hip. When a right hip was analyzed, a left-right flipping of the image was performed first.

Before calculating image similarity between any scans, we first transformed all the hip QCT images into a common hip space using image registration. In this study, as we did in previous image registration studies[17,19,20], one scan was randomly selected to serve as the reference (target) scan (for this particular study, it was a scan from one control subject). All other scans, from fractured patients and controls, were registered to this reference. The registration transformations included rotations, translations and scaling (independent transformations in x , y , and z directions). The algorithm was based on maximizing normalized mutual information [30], using simplex optimization under a multi-resolution scheme[20]. After all the scans were registered toward the reference scan, image similarity between any two images could be calculated, as described below.

Our image similarity measures were adapted from the image registration field [8,9,28,29,33]. The measures used for this study could be divided into two categories. The first was the summation of voxel-wise differences[8], and the second was mutual information[7,32]. For the first category, we calculated the sum of the absolute intensity differences (SAD), and the sum of coefficients of variation (SCV). The SAD and the SCV measures are the opposite of image similarity, i.e., if the difference measured between two images is greater, the images are less similar.

The SAD measure calculates the summation of the absolute intensity differences between the corresponding voxels of two images, as formulated below

$$SAD = \frac{\sum_{i,j,k} |CT_{ijk}^1 - CT_{ijk}^2|}{N}$$

where the summation runs over the voxels inside the proximal femur and N is the total number of voxels. CT_{ijk}^1 and CT_{ijk}^2 are the grayscale values of the two images at voxel (i,j,k) . The SAD defined above is actually the mean (instead of summation) of the absolute differences. However, we use the term SAD (as well as SCV used below) in this paper to make it consistent with what has been used in the literature[8].

The SCV measure calculates the sum of the CV values. At each voxel, the CV is defined as

$$CV = \frac{\sqrt{(CT_1 - \overline{CT})^2 + (CT_2 - \overline{CT})^2}}{\overline{CT}}$$

where CT_1 and CT_2 are the grayscale intensities of the two images, and $\overline{CT}$ is the mean of these two. The CV value would be zero if there was no difference between the two images at that

voxel, otherwise this value serves as an indication of the difference between the two images at that voxel. SCV is the summation (and taking the average) of such CV values:

$$SCV = \frac{\sum_{i,j,k} CV_{ijk}}{N}$$

where N is again the number of voxels in the proximal femur.

Mutual information (MI) was proposed independently by Collignon et al. [7] and Viola and Wells [32] for applications in image registration. It is derived from the frequency of co-occurrence of values in two images. For images M and N , the mutual information is defined as

$$MI(M, N) = H(M) + H(N) - H(M, N)$$

where $H(M)$ and $H(N)$ are the marginal entropies derived from the probabilities of occurrence of intensities in images M and N , respectively, and $H(M, N)$ is the joint entropy, as given by

$$H(M) = - \sum_{m \in M} p\{m\} \log p\{m\}$$

$$H(N) = - \sum_{n \in N} p\{n\} \log p\{n\}$$

and

$$H(M, N) = - \sum_{m \in M} \sum_{n \in N} p\{m, n\} \log (p\{m, n\}),$$

where $p\{m\}$ and $p\{n\}$ are the marginal probability distribution functions (PDFs) of M and N respectively, and $p\{m, n\}$ is the joint probability distribution function of M and N . The joint PDF is the normalized joint histogram, i.e., the joint histogram divided by the number of voxels in the overlapping regions of images M and N . The marginal PDFs can be thought of as the projections of the joint PDF onto the axes corresponding to intensities in images M and N respectively [8]. The summations run over the proximal femora (excluding soft tissue and the acetabulum). More detailed description about the general algorithm of calculating mutual information can be found in [7,28-30,32], and its adaptation to bone QCT images in [20].

Not like BMD or geometric parameters that can be measured based on a scan itself, the image similarity measures were obtained by comparing the scan with respect to some reference scans. Such reference scans formed a Model Group, which included both fracture patients and controls. We partitioned the 75 scans (38 controls and 37 fractures) into two groups: Group I and Group II. The random partitioning was performed in a stratified way so that Group I had 19 controls and 19 fractures, and the remainder formed Group II. At first, Group I served as the Model Group, while Group II served as the Test Group. As a further validation, the roles of Group I and Group II were switched and the calculations repeated. Hereafter we will refer to these two tests as the Initial-order Test and the Switched-order Test.

Mutual information calculations were carried out as follows. For each QCT scan from the Test Group, for example scan t , we calculated the mutual information between this proximal femur and the proximal femur of each of the scans in the Model Group. Then for scan t , we calculated the mean mutual information (after obtaining all the individual mutual information values) between scan t and the scans from the control subjects in the Model Group, denoted as

$\overline{MI}^{control}$. Similarly, we calculated the mean mutual information between scan t and the scans from the fracture subjects in the Model Group, denoted as $\overline{MI}^{fx}$. The difference:

$$\Delta\overline{MI} = \overline{MI}^{control} - \overline{MI}^{fx}.$$

If scan t were from a control subject, and if our hypothesis is true, its similarities with respect to other control subjects would be higher than its similarities with respect to the fracture subjects. Thus, we would expect $\Delta\overline{MI}$ to be positive for this control subject. Alternatively, if scan t were from a fracture patient, we would expect $\Delta\overline{MI}$ to be negative.

After obtaining the $\Delta\overline{MI}$ values for all the scans from the Test Group, we used them to discriminate the fractured patients from controls. A test scan with a positive $\Delta\overline{MI}$ was classified as a scan from a control subject, while a test scan with a negative $\Delta\overline{MI}$ was classified as a scan from a fractured patient. We then calculated the false positive and false negative rates, as well as the sensitivity and specificity. Here false positive refers to a control subject being classified as a fractured subject, and false negative refers to a fractured subject being classified as a control subject. For either the Initial Test or the Switched-order Test, the above procedure was carried out independently. The mean values of the obtained two sets of sensitivity and specificity values were used as the overall sensitivity and specificity for the mutual information measure.

The same procedure was repeated for the SAD and the SCV measures. Since the SAD and SCV measures are in the opposite direction of image similarity (i.e., larger difference implies less similar), we defined $\Delta\overline{SCV}$ as $\overline{SCV}^{fx} - \overline{SCV}^{control}$ (similarly for $\Delta\overline{SAD}$), i.e., we switched the order of subtraction compared to the definition of $\Delta\overline{MI}$.

To form a comparison standard to the image similarity approach, we also used BMD to discriminate the fractured patients from the controls. BMD can be measured in different regions of interest. We used the BMD measured in the femoral neck, the region of interest suggested in the FRAX model[10]. We also calculated the Z-scores based on the BMD average and standard deviation of the control subjects of our cohort. To evaluate the fracture discrimination performance, we plotted receiver operating characteristic (ROC) curve and calculated the area under curve (AUC). Furthermore, from the ROC curve, the optimal BMD cutoff was derived. With such an optimal cutoff, the sum of the false positives and false negatives minimizes. In an ROC curve, the optimal cutoff also appears to be the point closest to the upper-left corner. For the BMD method, we tested three approaches to estimate the fracture discrimination performance. 1) Performing ROC analysis separately for Group I and Group II alone; 2) using the optimal cutoff derived from one group as a training parameter for the other group; 3) considering the two groups together to estimate an overall performance. Although all the three approaches could give the same results if the sample size was large enough, due to our limited sample size and the variance between the two groups, the former two approaches gave different results for Group I and Group II. Thus, to avoid bias, the results from the third approach were considered as the overall BMD performance when the BMD method was compared to the image similarity methods.

We also explored whether BMD and the image similarity measures could be combined to improve fracture discrimination. We used two approaches: linear discriminant analysis (LDA) [24,31] and recursive partitioning and regression trees (RPART)[3]. For the LDA approach, a linear combination of BMD and similarity measures was used as a new measure for fracture discrimination. To derive the linear combination, we used the *lda* package in the R statistical language and supplemental libraries [23]. For the RPART approach, the test images were classified by decision trees derived from the *rpart* package in R, and the predicted values

obtained for the test images were used to plot the ROC curve. In both the LDA and the RPART approaches, we ran the corresponding algorithms with all the scans together, since the fracture risk estimations were derived independently, i.e., when a scan's fracture risk was estimated in either the Initial-order Test or the Switched-order Test, the estimation was not affected by whether or not this scan served as a model scan in a separate test.

To illustrate our technique of comparing image similarities, we computed average images of subjects with fractures and average images of controls. In an average image, the grayscale value of a voxel, for example (i, j, k) , was the mean of the grayscale values at (i, j, k) for images used to generate this average image (after the scans were registered to the common hip space). We computed four average images from the subgroups of the fractured patients and the controls in Group I and Group II. And for this illustration purpose, we used the SAD measure. For two average images, we derived a 3D map that showed their voxel-wise absolute differences. We observed whether such a map derived from the fracture/control average images differed from the map derived from the control/control average images, as well as the map derived from the fracture/fracture average images. We would like to point out that our purpose of comparing such difference maps was to help illustrate the concept of comparing image similarities. Such difference maps derived from average images were not intended to be interpreted as a statistically rigorous measure of bone differences. A more rigorous comparison between bones from fractured patients and controls based on voxel-by-voxel t-statistics can be found in [19].

RESULTS

Fig. 1 shows the maps of voxel-wise differences between average images. In the left two columns, we compared the average image of Group II controls (denoted as “control (II)” in the figure) to the average image of Group I fractured patients (denoted as “fx (I)”), and the average image of Group I controls (denoted as “control (I)”), respectively. The control-fracture differences, as shown in Fig. 1A and Fig. 1B (in the coronal and axial views, respectively), were visibly greater (brighter) than the control-control differences (Fig. 1C and Fig. 1D), which were close to zeros (much darker). Similarly, in the right two columns, we compared the average image of Group II fractured patients to the average image of Group I controls and the average image of Group I fractured patients. We found that the fracture-control differences, as shown in Fig. 1G and Fig. 1H, were visibly greater than the fracture-fracture differences as shown in Fig. 1I and Fig. 1J.

When using BMD values to discriminate the 38 fractured patients and 37 controls, according to ROC analysis, the AUC value was 0.835. The optimal BMD cutoff determined by the ROC analysis is shown in Fig. 2. The corresponding BMD was -0.479 (Z-score). At this optimal cutoff, the sensitivity and specificity were 86.5% and 73.7%, respectively. The discrimination results can also be seen in the scatter plots of Fig. 2. The subjects in the green regions were correctly classified. To aid visual comparison with later figures, the subjects are shown in four sub-groups in the scatter plots. In total, for the fractured patients, the number of subjects that were incorrectly classified by the discrimination procedure, i.e., the false negatives, were 5 (out of 37), with a false negative rate of 13.5%. The false positives were 10 (out of 38), with a false negative rate of 26.3%. These results are summarized in Table 1 for comparisons with the image similarity measures.

For the BMD method, in addition to the above analysis where all the 75 scans were considered together, we also tried to analyze Group I and Group II separately. Based on Group I data alone, the AUC value was 0.807. At its optimal cutoff, the sensitivity and specificity were 89.5%, and 68.4%, respectively. Based on Group II alone, the AUC value was 0.871. At its optimal cutoff, the sensitivity and specificity were 94.4%, and 73.7%, respectively. Furthermore, we tried to use the optimal cutoff derived from one group as the training parameter for the other

group. With this approach, the sensitivity and specificity obtained were 89.5% and 47.4% for Group I, and 83.3% and 78.9% for Group II. Due to the group variance as seen from these results, the discrimination performance using the 75 scans together were considered as the overall performance for the BMD method.

When using SCV as the image similarity measure, for the Initial-order Test, Fig. 3 shows the distributions of $\Delta\overline{SCV}$ for the test scans from the fracture subjects (Fig. 3A) and the control subjects (Fig. 3B). We used the sign of $\Delta\overline{SCV}$ to classify the fracture patients and controls. All subjects falling into the green regions were correctly classified. Although not every scan's result supported our hypothesis, a general agreement can be seen. In particular, for the scans from the 18 fractured patients, all had the expected negative $\Delta\overline{SCV}$ values. For the 19 control subjects, all but 3 scans showed the expected positive $\Delta\overline{SCV}$ values. Therefore, for the Initial-order Test, except for these 3 subjects, all other subjects in this test group supported our hypothesis, i.e., an image from a fractured patient would show higher image similarities with other fractured patients, while an image from a control subject would show higher image similarities with other control subjects. After switching the roles of Group I and Group II, i.e., for the Switched-order Test, as shown on the right, we again see that all fractured patients had the expected negative $\Delta\overline{SCV}$ values (Fig. 3C). However, there were 9 out of the 19 controls subjects that did not have the expected positive $\Delta\overline{SCV}$ values (Fig. 3D). By taking the mean values of the sensitivity and specificity values from the two rounds of tests (the Initial-order and the Switch-order Tests), we obtained an overall sensitivity of 100%, and a specificity of 68.4%. Similar results were obtained when using SAD as the image similarity measure, where all 37 fractured patients were correctly classified, and out of the 38 controls, 14 were incorrectly classified (5 for the Initial-order Test and 9 for the Switched-order Test), as tabulated in Table 1.

When using mutual information as the image similarity measure, as shown in Fig. 3, for the Initial-order Test, the false negatives were 4 (out of 18) for the fractured patients (Fig. 3E), and the false positives were 3 (out of 19) for the controls (Fig. 3F). For the Switched-order Test, there was 1 (out of 19) false negative for the fractured patients (Fig. 3G) and 6 (out of 19) false positives for the controls (Fig. 3H). By combining the Initial-order and the Switched-order Tests, the overall false negative rate was 13.5% (5/37) and the overall false positive rate was 23.7% (9/38). The corresponding overall sensitivity and specificity were 86.5% and 76.3%, respectively.

In Fig. 4 the mutual information and the BMD measurements (using the BMD results based the 75 subjects together, as mentioned earlier) are plotted together. The BMD cutoff ($= -0.479$) and the MI cutoff ($= 0$) were the cutoffs used earlier when BMD or $\Delta\overline{MI}$ was used individually for discrimination. By combining BMD and $\Delta\overline{MI}$ the new (BMD, MI) cutoff, derived from LDA, is also shown as the dashed line. With this cutoff, the false negatives were 2 and the false positives were 10. Furthermore, by combining BMD and all the three image similarity measures (MI, SAD, SCV), a new cutoff integrating all the four measures was derived by LDA (which corresponded to a "line" in a 4-dimensional space), which resulted in a sensitivity of 94.6% and a specificity of 76.3%, respectively. Interestingly, in the combination case of (BMD, MI, SAD, SCV), when we replaced (SAD, SCV) by either SAD or SCV alone, the sensitivity and specificity remained unchanged.

Fig. 5 shows the results derived from the RPART algorithm when using BMD, MI and SCV for fracture discrimination. The sensitivity and the specificity were 91.9% and 81.6%, respectively. Similar to the LDA algorithm, when using SAD, or the combination of (SAD, SCV) to replace SCV, the same sensitivity and specificity were obtained.

Table 1 summarizes the fracture discrimination results using BMD, image similarity and their combination for hip fracture discrimination. We compared the AUC value obtained using the BMD method with those obtained using the other methods. The difference between the BMD and the RPART methods showed statistical significance.

For our image analysis, we used a Dell workstation with a Pentium 4 processor, 2.4 Ghz CPU and 1.5Gb RAM. The average CPU time for registering a hip pair was 2.3 minutes. After image registration, for a single scan, the time needed to perform the image similarity calculations and to estimate its fracture risk was generally negligible (around 2 seconds). The image registration and image similarity calculations were automatic, no manual operation was needed.

DISCUSSION

In this study, we adapted image similarity, a measurement that is routinely used in image registration, to estimate hip fracture risk. In contrast to conventional bone densitometry, which measures average intensity across an image region, our image similarity approach aims to capture information about the spatial distribution of bone mineral in a given hip compared to that of hips from subjects with known fracture risk. Our fracture risk estimation is based on the hypothesis that the image similarities between high-risk/high-risk subjects or low-risk/low-risk subjects will be greater than the image similarities between high-risk/low-risk subjects. To test this hypothesis, we analyzed hip QCT scans from fractured patients and controls, and demonstrated that the image similarities between the control/control subjects and between the fracture/fracture subjects were indeed higher than that between the fracture/control subjects. Based on this difference in image similarity, we were able to perform fracture risk estimation and to apply it to discriminate hip scans of fractured patients from those of control subjects. According to ROC analysis, the image similarity approaches showed higher AUC values compared to the BMD approach. In particular, with our relatively small sample size, when BMD was combined with the image similarity measures (in the Recursive Partitioning and Regression Trees approach), compared to using BMD alone, the AUC value was improved with statistical significance (p-value: 0.0046).

The motivation for exploring image similarity based fracture risk estimation was its potential to capture bone fragility features related to spatial similarity that have not been captured by conventional bone densitometry. This can be seen from a two-voxel example. If two voxels switch their grayscale values, based on BMD measurement, the pre- and post-switch scenarios cannot be distinguished (since the average BMD stays the same). However, an image similarity measure (such as any of those used in this study) would show that the two scenarios are different. Therefore, image similarity measures carry additional spatial information that is missed by BMD measurement.

It is interesting to notice that when we combined the BMD and the image similarity measures, in both combination algorithms (LDA and RPART), when SAD (or SCV) was included, replacing it by the other, SCV (or SAD), or by the combination of (SAD, SCV), the sensitivity and specificity remained unchanged. This is not very surprising, instead, this indicates that SAD and SCV essentially carry similar information. They both belong to the first image similarity category, i.e., the summation of voxel-wise differences. Including either of them or including both of them did not seem to change the discriminatory power. On the other hand, by adding image similarity information to BMD, both sensitivity and specificity were improved compared to using BMD alone. A plausible explanation to this would be that the image similarity measures do carry bone fragility information not contained by BMD. In addition to BMD, our new fracture risk estimation approach may also be integrated with other fracture risk factors, such as those clinical factors recommended in the FRAX model, as well as

functional measurements such as bone turnover markers, or even other image based fracture risk factors.

In clinical practice and research, besides CT imaging, DXA imaging has also been widely used for bone mineral density measurements and fracture risk estimation. We also estimated fracture discrimination using simulated DXA images derived from our QCT scans. We generated simulated DXA images using an approach described in [14], where the 3D QCT images were projected in the coronal direction to generate the simulated DXA images. Similar to the QCT images, we measured the areal BMDs in the femoral neck, and applied such areal BMDs to discriminate the subjects. According to ROC analysis, the AUC value was 0.686. The sensitivity and specificity at the optimal cutoff were 59.5% and 76.3%, respectively. This AUC value was considerably lower than those obtained using the QCT volumetric BMD and the image similarity measures. However, the performance using such simulated DXA images does not necessarily represent the performance using true DXA images, since after the numerical projection, our simulated DXA images generally showed poorer delineation of the proximal femur compared to true DXA images.

The image similarity measures tested in this study can be roughly categorized into two types, one based on the summation of voxel-wise image differences, including the SAD and the SCV measures, while the other is mutual information. Our results showed that the first type had the best sensitivity, and the two types had similar specificities. However, mutual information could be more robust with respect to CT calibration. Mutual information is calculated based on histogram/entropy analysis of intensity co-occurrence distribution of two images. Such co-occurrence distribution is not necessarily affected by image calibration. In fact, in image registration applications, mutual information has been used to calculate image similarity between images from different modalities, such as between a CT image and an MRI image [30]. We expect the mutual information approach to be potentially applicable to our image similarity based fracture estimation for images scanned from different CT scanners, or even scanned from different modalities. Therefore, the mutual information approach may be a more technically simple option and more feasible as a measure of proximal femoral fragility in clinical studies.

In this study, we selected hip fracture as an endpoint by which to develop and test our methods, since hip fracture is the most severe manifestation of bone fracture and it accounts for 72% of the total cost due to bone fractures [4]. However, our technique is easily applicable to vertebrae and to other skeletal sites (such as forearm and tibia). In addition, our image similarity analysis may be integrated with other image analysis techniques such as finite element modeling (FEM) [6,11,12,21]. Besides estimating bone strength (fracture load), FEM also generates 3D maps of stress and strain distributions under the external load, and Keyak et al. have also derived factor of safety (FOS) maps [12] to depict the regional fragility properties under the external load. Such regional maps of load bearing capacity may potentially contain information about bone fragility beyond what is contained in a single scalar value such as failure load. For example, the stress maps of two fractured patients may be more similar to each other than to the stress map of a control subject. Therefore, analyzing the image similarities between FEM maps may yield additional bone fragility information than estimating the fracture load alone.

Although our study has the fundamental strength of evaluating a novel image-based bone fragility marker against a clinically-relevant endpoint, it also has some limitations. One limitation is that due to the relatively small sample size, all the scans from the Model Group were treated equally when applied to estimate a test scan's fracture risk. In fact, the power of each model image for determining the fracture risk of a test scan is not necessarily the same. In other words, for some model images, the image similarity calculated between the model scan and the test scan may accurately reflect the fracture risk of the test scan, while for other

model images this may not be that accurate. Therefore, it would be ideal if we could have an additional Training Group to test which model images do have the ability to predict fracture risk for other scans. After such a training process, we would select only those model scans that do have the predictive power to form a new Model Group. Or alternatively, we can assign a weighting factor for each model image when considering its contribution to fracture risk estimation. We expect such a further study could be carried out once we have a larger study cohort available. Furthermore, in our study, there were only eight subjects with trochanteric fractures (the rest of the patients had neck fractures). In our modeling, the subjects of these two types of hip fractures were not distinguished, although there could be fracture type specific structural features as demonstrated in [18]. With a sufficient number of cases with both neck fractures and trochanteric fractures, we may be able to compare how the image similarity features differ between these two types of hip fractures, and potentially provide better predictive ability for each fracture type. In addition, it is worth pointing out that the sensitivity and specificity reported here may not be representative of the performance of this technique in the general population. There are several reasons for this. First, our study samples included roughly half fractured patients and half control subjects, while the fracture rate in the general population is much lower. Secondly, the subjects in this study were all Chinese women, who do not necessarily reflect fracture risk factors in other ethnicities. For example, the Asian population generally has a much lower fracture rate compared to Caucasian and Hispanic women, even lower than African American women [27]. Thirdly, the sensitivity and specificity used in this study were employed to determine the rates at which the subjects could be classified to the correct groups. In clinical diagnosis, however, the risk factors are mainly used to make a cost-effective treatment decision. The decision cutoff would not necessarily match the cutoff used in this study.

For future work, we expect to study this new technique's ability to predict bone fracture in a prospective manner, and to characterize the image similarity features associated with ethnicity, gender, and age differences, as well as to characterize changes in image similarity features due to disease, pharmacologic interventions, and exercise.

In conclusion, we have proposed and tested a new bone fracture risk estimation technique based on image similarity. By adding image similarity information to bone mineral density, we observed improved fracture discriminatory power compared to that obtained by bone densitometry alone. The image similarity based technique may also be integrated with other fracture risk factors to further improve our ability in evaluating bone fracture risk. Such an improvement could result in more accurate selection of individuals for intervention, and more effective monitoring of patient response to osteoporosis therapies.

Acknowledgments

Development of the analytic technique was supported by NASA Grant NNJ04HF78G and NIH grant NIH-R42-AR45713. Acquisition of patients was supported by a grant from Eli Lilly.

REFERENCES

1. Black DM, Bilezikian JP, Ensrud KE, Greenspan SL, Palermo L, Hue T, Lang TF, McGowan JA, Rosen CJ. One year of alendronate after one year of parathyroid hormone (1-84) for osteoporosis. *N Engl J Med* 2005;353:555–65. [PubMed: 16093464]
2. Black DM, Greenspan SL, Ensrud KE, Palermo L, McGowan JA, Lang TF, Garnero P, Bouxsein ML, Bilezikian JP, Rosen CJ. The effects of parathyroid hormone and alendronate alone or in combination in postmenopausal osteoporosis. *N Engl J Med* 2003;349:1207–15. [PubMed: 14500804]
3. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees*: Wadsworth. 1984

4. Burge R, Dawson-Hughes B, Solomon DH, Wong JB, King A, Tosteson A. Incidence and economic burden of osteoporosis-related fractures in the United States, 2005-2025. *J Bone Miner Res* 2007;22:465–75. [PubMed: 17144789]
5. Cheng X, Li J, Lu Y, Keyak J, Lang T. Proximal femoral density and geometry measurements by quantitative computed tomography: Association with hip fracture. *Bone* 2007;40:169–174. [PubMed: 16876496]
6. Cody DD, Gross GJ, Hou FJ, Spencer HJ, Goldstein SA, Fyhrie DP. Femoral strength is better predicted by finite element models than QCT and DXA. *Journal of Biomechanics* 1999;32:1013–20. [PubMed: 10476839]
7. Collignon, A.; Maes, F.; Delaere, D.; Vandermeulen, D.; Suetens, P.; Marchal, G. Automated multi-modality image registration based on information theory. In: Bizais, Y.; Barillot, C.; Di Paola, R., editors. *Information Processing in Medical Imaging*. Kluwer; Dordrecht, The Netherlands: 1995. p. 263-274.
8. Hill DL, Batchelor PG, Holden M, Hawkes DJ. Medical image registration. *Phys Med Biol* 2001;46:R1–45. [PubMed: 11277237]
9. Hutton BF, Braun M. Software for image registration: algorithms, accuracy, efficacy. *Semin Nucl Med* 2003;33:180–92. [PubMed: 12931320]
10. Kanis, J. Assessment of osteoporosis at the primary health care level. Technical Report. WHO Collaborating Centre: University of Sheffield; UK: 2008.
11. Keyak JH, Rossi SA. Prediction of femoral fracture load using finite element models: an examination of stress- and strain-based failure theories. *Journal of Biomechanics* 2000;33:209–14. [PubMed: 10653034]
12. Keyak JH, Rossi SA, Jones KA, Skinner HB. Prediction of femoral fracture load using automated finite element modeling. *J Biomech* 1998;31:125–33. [PubMed: 9593205]
13. Lang T, Koyama A, Li C, Li J, Lu Y, Saeed I, Gazze E, Keyak J, Harris T, Cheng X. Pelvic body composition measurements by quantitative computed tomography: association with recent hip fracture. *Bone* 2008;42:798–805. [PubMed: 18234578]
14. Lang T, LeBlanc A, Evans H, Lu Y, Genant H, Yu A. Cortical and trabecular bone mineral loss from the spine and hip in long-duration spaceflight. *J Bone Miner Res* 2004;19:1006–1012. [PubMed: 15125798]
15. Lang TF, Keyak JH, Heitz MW, Augat P, Lu Y, Mathur A, Genant HK. Volumetric quantitative computed tomography of the proximal femur: precision and relation to bone strength. *Bone* 1997;21:101–8. [PubMed: 9213015]
16. Lang TF, Li J, Harris ST, Genant HK. Assessment of vertebral bone mineral density using volumetric quantitative CT. *Journal of Computer Assisted Tomography* 1999;23:130–7. [PubMed: 10050823]
17. Li W, Kezele I, Collins L, Zijdenbos A, Keyak J, Kornak J, Koyama A, Saeed I, LeBlanc A, Harris T, Lu Y, Lang T. Voxel based modeling and quantification of the proximal femur using inter-subject registration of quantitative CT images. *Bone* 2007;41:888–895. [PubMed: 17707712]
18. Li W, Kornak J, Harris T, Keyak J, Li C, Lu Y, Cheng X, Lang T. Identify fracture-critical regions inside the proximal femur using statistical parametric mapping. *Bone* 2009;44:596–602. [PubMed: 19130910]
19. Li W, Kornak J, Harris T, Keyak J, Li C, Lu Y, Cheng X, Lang T. Identify Fracture-critical Regions inside the Proximal Femur Using Statistical Parametric Mapping. *Bone*. 2009 10.1016/j.bone.2008.12.008.
20. Li WJ, Sode M, Saeed I, Lang T. Automated registration of hip and spine for longitudinal QCT studies: Integration with 3D densitometric and structural analysis. *Bone* 2006;38:273–279. [PubMed: 16199215]
21. Lotz JC, Hayes WC. The use of quantitative computed tomography to estimate risk of fracture from falls. *J Bone Joint Surg [Am]* 1990;72:689–700.
22. Orsini LS, Rousculp MD, Long SR, Wang S. Health care utilization and expenditures in the United States: a study of osteoporosis-related fractures. *Osteoporos Int* 2005;16:359–71. [PubMed: 15340799]
23. R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; 2004. ISBN 3-900051-00-3 <http://www.r-project.org>

24. Ripley, BD. Pattern Recognition and Neural Networks. Cambridge University Press; 1996.
25. Robbins J, Aragaki A, Kooperberg C, Watts N, Wactawski-Wende J, Jackson R, LeBoff M, Lewis C, Chen Z, Stefanick M, Cauley J. Factors associated with 5-year risk of hip fracture in postmenopausal women. JAMA 2007;298:2389–2398. [PubMed: 18042916]
26. Siris ES, Brenneman SK, Miller PD, Barrett-Connor E, Chen YT, Sherwood LM, Abbott TA. Predictive value of low BMD for 1-year fracture outcomes is similar for postmenopausal women ages 50–64 and 65 and older: Results from the National Osteoporosis Risk Assessment (NORA). Journal of Bone and Mineral Research 2004;19:1215–1220. [PubMed: 15231007]
27. Siris ES, Miller PD, Barrett-Connor E, Faulkner KG, Wehren LE, Abbott TA, Berger ML, Santora AC, Sherwood LM. Identification and fracture outcomes of undiagnosed low bone mineral density in postmenopausal women - Results from the National Osteoporosis Risk Assessment. Jama-Journal of the American Medical Association 2001;286:2815–2822.
28. Studholme C, Hill DL, Hawkes DJ. Automated 3-D registration of MR and CT images of the head. Med Image Anal 1996;1:163–75. [PubMed: 9873927]
29. Studholme C, Hill DL, Hawkes DJ. Automated three-dimensional registration of magnetic resonance and positron emission tomography brain images by multiresolution optimization of voxel similarity measures. Med Phys 1997;24:25–35. [PubMed: 9029539]
30. Studholme C, Hill DLG, Hawkes DJ. An overlap invariant entropy measure of 3D medical image alignment. Pattern Recognition 1999;32:71–86.
31. Venables, WN.; Ripley, BD. Modern Applied Statistics with S. Fourth Edition. Springer; 2002.
32. Viola, P.; Wells, WM. Alignment by maximization of mutual information. In: Grimson, E.; Shafer, S.; Blake, A.; Sugihara, K., editors. Proc. Int. Conf. Computer Vision. Los Alamitos, CA: 1995. p. 16–23.
33. Zitová B, Flusser J. Image registration methods: a survey. Image and Vision Computing 2003;21:977–1000.

Fig. 1.

Image differences (represented here by brightness of red overlay, and serving as an inverse map of image similarity) between average images. In the middle column are QCT sections in the coronal view (top) and the axial view (bottom), respectively. On its left and right are QCT sections superimposed by the image difference maps (derived from the SAD measure). In the left two columns, the average image of Group II controls was compared to the average image of Group I fractured patients, and the average image of Group I controls. In the right two columns, the average image of Group II fractured patients was compared to the average image of Group I controls, and the average image of Group I fractured patients.

Fig. 2.
ROC curve with optimal cutoff (left), and scatter plots of BMD discrimination.

Fig. 3. Distributions of image similarity differences. The cases falling to the highlighted green regions are those that matched the expectations.

Fig. 4. Linear discriminant analysis: combining BMD and the mutual information measure for fracture discrimination.

Fig. 5. Recursive partitioning and regression trees: combining BMD, MI, and SCV for fracture discrimination.

Table 1

Comparison of fracture discrimination results using BMD, image similarity measures, and their combination based on linear discriminant analysis (LDA) or recursive partitioning and regression trees (RPART)[†].

Measure	Sensitivity	Specificity	False negatives (out of 37 fractured patients)	False positives (out of 38 controls)	AUC (p-value; compared with BMD)
BMD (femoral neck)	86.5%	73.7%	5	10	0.835 (-)
Image Similarity	Summing voxel-wise differences	SCV	0	12	0.885 (0.097)
		SAD	0	14	0.883 (0.12)
	Mutual information (MI)		5	9	0.845 (0.81)
Combination (BMD + similarity measures)	LDA	76.3%	2	9	0.878 (0.11)
	RPART	81.6%	3	7	0.923 (0.0046)

[†] In this table, the AUC values derived from the corresponding ROC curves were used as an overall evaluation of the discrimination efficacy. In addition, for the image similarity measures (SCV, SAD, and MI), we calculated the sensitivity, specificity (as well as false positives and false negatives) using the natural cutoff (i.e., a scan was classified depending on whether it was more similar to fractured patients or more similar to controls, as described in the text). For the BMD method and its combinations with the image similarity measures (LDA, RPART), there was no such a natural cutoff, and the sensitivity and specificity (as well as false positives and false negatives) were the values corresponding to the optimal cutoffs derived from the ROC curves, as described in the text.