Evaluation of a semi-automated software program for the identification of vertebral fractures in children

Introduction

Fractures are common in childhood and repeated fractures reflect the interacting effects of low bone mineral density (BMD) and/or physical activity [1]. Vertebral fractures (VFs) are a relatively common type of osteoporotic fracture. The detection of one or more vertebral compression (crush) fractures (identified by a 20% reduction in vertebral body height) is indicative of bone fragility irrespective of the reported BMD [1]. Although a lot of recent research has been conducted regarding the occurrence of osteoporotic VF in adults, relatively less attention has been paid towards pediatric VF, largely on account of the lack of an accepted standardized diagnostic technique in children [2].

In the absence of major trauma, reduced BMD in children and adolescents is the major cause of VF; indeed the finding of a VF is a main diagnostic feature of low BMD in children [1]. The low BMD may be primary (e.g. osteogenesis imperfecta) or secondary [1, 3]. For example, the STOPP studies have implicated glucocorticoids as a significant cause of secondary fractures in children and shown an incidence of vertebral fractures in those with a new diagnosis of acute lymphoblastic leukemia of 16% [4, 5]. Unlike osteoporotic fractures of the limbs, VFs are typically silent and if untreated may lead to progressive loss of vertebral body height and potential spinal deformity. If VFs are diagnosed early, however, bisphosphonate treatment can help to treat existing fractures and reduce future fracture risk [6].

Assessment of VFs in children is performed using standard lateral spine radiographs and, currently, these are interpreted using a subjective visual assessment method to identify loss of height/change in shape consistent with VF. This approach is hampered by significant inter and intraobserver variability [2; 7, 8], which is likely to be reduced if a more objective assessment method is applied. Semi-automated software programs such as SpineAnalyzer (Optasia Medical, Cheadle, UK) may be the solution; but, so far, limited studies have been carried out to evaluate these programs in children. The potential added value of these programs is that non-radiologists may be trained to use them, freeing up radiologists’ time for more specialized tasks.

The purpose of this study was to assess the observer reliability and diagnostic accuracy in children and adolescents, of the semi-automated 6-point technique developed for VF diagnosis in adults, using a semi-automated software program (SpineAnalyzer). This software records percentage loss of vertebral body height and classifies fractures based on the Genant system [9].

Materials and methods:

Study population

This study involved the retrospective analysis of images obtained as part of a larger prospective study of 250 children recruited between November 2011 and February 2014 [7]. All images used in this study were of patients recruited from our single center. The mean age of the 137 subjects at the time of image acquisition was 12.0 years (range 5 to 15) and 45 (33%) were male. The majority, 199 (80%) had suspected reduction in BMD (including children with osteogenesis imperfecta, inflammatory bowel disease, rheumatologic conditions, cystic fibrosis and celiac disease). The remaining 51 (20%) patients were recruited from spine clinic.

Lateral spine imaging

Lateral images of the thoracolumbar spine were acquired using one of two Phillips Healthcare machines (TH3 Digital or TH Bucky Diagnost, Guildford, UK) following the European guidelines for imaging the spine in children as previously described [7]. The subjects were asked to remain in the lateral decubitus position with flexed knees and hips. Depending on the size of each child being examined, thoracolumbar or separate thoracic and lumber spine images were obtained. As outlined in a previous study, the tube-to-film distance was set at 100 cm, and the films were centered at T7 and L3 for the thoracic and lumbar views, respectively [10]. The average exposures for thoracic, lumbar and thoracolumbar spine radiographs were 75, 84 and 74kV respectively.

Image analysis

Lateral spine images were analyzed independently by five observers (a radiologist, two radiographers, and two medical students), who attempted readings for all 137 cases, with each observer being blinded to the other evaluations. Prior to commencing the study, the four non-radiologists were trained on use of the software by the radiologist, learning from non-study spine radiographs. A previous consensus arrived at by three pediatric radiologists using a simplified algorithm-based qualitative (ABQ) technique (i.e. with no software involved) served as the reference standard [10].

As the first step in the semi-automated analysis using SpineAnalyzer, observers identify the T4 to L4 vertebral bodies by placing a point at or close to the center of each vertebral body and indicating to the software the highest identified vertebral body (for example, T4). Having indicated T4, the software program recognizes all identified vertebral bodies between T4 and L4 and automatically identifies six points corresponding to the four corners and the midpoints of the superior and inferior endplates of each vertebral body – observers modify the placement of these points as necessary. The software does not recognize vertebral bodies above T4 or below L4 (Fig 1).

Following placement of the six points, anterior, middle and posterior vertebral heights are automatically determined by the software. With the help of these measurements, the anterior: posterior, middle: posterior, posterior: posterior^{+1} and posterior: posterior $^{-1}$ height ratios are calculated (+1 and -1 indicate the vertebrae immediately above (+1) and below (-1) the vertebra of interest). The vertebral bodies are then classified according to their height ratios, based on the scoring system developed by Genant (Table 1 and Fig 1) [9].

For the purposes of this study, since the assessment only included lateral spine images, to maintain the consistency of vertebral level assignment between the five observers, the first vertebral body not associated with ribs was labelled as L1, while the lowermost vertebral body associated with ribs was labelled as T12. If the observer was unable to identify T12 and/or L1, (e.g. due to excessive coning), then that image was not scored.

Statistical Analysis

R software was employed for data analysis [11]. The frequency of readable vertebrae for each observer and for all vertebrae from T4 to L4 was calculated.

Diagnostic accuracy (sensitivity, specificity and 95% confidence interval) calculations of the observers' readings were calculated by comparing with a previously established consensus arrived at by three experienced pediatric radiologists using a simplified algorithm based qualitative scoring system (sABQ), Table 2 [10]. For diagnostic accuracy calculations, both sABQ and SpineAnalyzer scores of 0 and 1 were interpreted as, "no clinically significant fracture". Inter and intra observer agreement were calculated using kappa and intraclass correlation coefficient (ICC) respectively [12, 13].

<u>Approvals</u>

Local Research Ethics Committee approval was obtained for the main study from which the images were drawn but was not separately required for this study. The study was registered with our Research and Innovation Department prior to commencement.

Results

Prevalence of fractures

Overall, 20 (15 %) patients had one or more VF (vertebral height loss 20 % or more). Per-vertebra, 48 VFs were identified by three or more observers using SpineAnalyzer. The majority of these fractures were in the mid-thoracic region, with T7 being the most fractured level - 9 (19%).

Readability of radiographic lateral spine images within SpineAnalyzer software program

Of the possible total 1781 vertebrae, from T4 through to L4 (i.e. 13 vertebrae per subject in 137 subjects), 1310 (73.55%) were adequately visualized by Observer 1, 1370 (77%) by Observer 2, 1376 (77%) by Observer 3 and 1319 (74%) and 1344 (75%) by Observers 4 and 5 respectively (Fig 2). A total of 1187 (67 %) were adequately visualized by three or more observers, permitting comparison of morphology results. The visibility was relatively limited in the upper part of the thoracic spine; T4 was the least readable level, being adequately visualized by all observers on 423 (62%) radiographs.

Sensitivity and specificity values of the observers' readings with their 95% confidence intervals are presented in Table 3. T6 had the highest and L3 the lowest sensitivity, while L4 had the highest and T11 the lowest specificity. Overall sensitivity was 18% (95% CI, 14 – 22), while overall specificity was 97% (95% CI, 97 – 98).

The average kappa for interobserver agreement in respect to vertebral readability between the five observers for each of the 13 vertebrae ranged from 0.05 to 0.47 (95% CI, -0.19, 0.76). Table 3 shows the agreement (average kappa score) between the five observers using SpineAnalyzer. T4 had the lowest and T12 the highest agreement. Average intraobserver agreement ranged from 0.25 to 0.61. Table 3 also shows that overall, there was poor/fair agreement for the 13 vertebrae, with the only exception being T5, for which agreement was good. Table 4 compares results of this current study with those of the only other study to date that has assessed the 6-point technique in children [8] and with those of the largest published study to compare VFA with radiographs for diagnosis of VF in children [7].

Figure 3 illustrates examples of good and poor observer agreement, while Figure 4 illustrates differences in diagnostic outcome due to early ossification of the apophyses causing minor observer differences in placement of the six points. Figure 5 demonstrates false positive and false negative results of SpineAnalyzer.

Discussion:

One or multiple VF without high-energy trauma or local disease is indicative of osteoporosis in children. Early and accurate diagnosis is important to allow appropriate treatment to commence.

There is a relatively low observer reliability for current techniques of VF diagnosis in children; with reported kappa values for inter and intraobserver reliability ranging from 0.39 to 0.59 and 0.33 to 0.84 respectively [2,7,8]. A recent study in adults showed an agreement between SpineAnalyzer and readers ranging from 0.96 to 0.97 [17]. The authors suggested that SpineAnalyzer is an accurate tool for measuring vertebral height and identifying VFs in adults. The purpose of this current study was to evaluate the accuracy and reliability of the semi-automated 6-point technique for diagnosing VF in children. To our knowledge, this evaluation is the largest to assess vertebral morphometry in children using semi-automated 6-point technique software, with only one other study on the same subject published to date [8].

Compared to our results, observer reliability has been shown to be higher in studies of the diagnostic accuracy of VF detection in adults using both visually-based scoring systems and software [14-17]. A recent study on children [2], based on the observation of radiographic images utilizing Genant's semi-quantitative (SQ) technique, showed higher inter-kappa agreement for VF diagnosis (k=0·45 to 0·54) than both our corresponding SpineAnalyzer calculations (k = 0.05 to 0.47) and those of Crabtree et al (k = 0.36 to 0.41) [8]. Results of the three studies should be directly comparable, given that the SpineAnalyzer categories are based on Genant's scoring system. It seems that small differences between observers in point placement account for the reduced observer reliability of SpineAnalyzer, compounded by the fact that the final categorization is based on ratios and not simple measurements. This is supported by the fact that the pediatric study from which images for this report were drawn also obtained a higher level of interobserver agreement (k = 0·394 to 0·455) when utilizing a simplified algorithm-based qualitative (sABQ) technique for vertebral morphometry [10].

Agreement between the observers reached a maximum kappa of 0.47 (95% CI, 0.18, 0.76) with the greatest level of agreement being at T12 and L4 (fair to moderate) whilst the least was at T4 (slight to poor). At each vertebral level, there was diversity in the interobserver agreement and readability of the vertebra (Fig.4). Results suggest that the observers could visualize the lower vertebral levels for point placement more adequately and that the calculations were correspondingly more precise than those made for the upper vertebral levels, underlining the difficulty in applying SpineAnalyzer for the upper thoracic spine. These findings support those of previous studies reporting that identification of vertebrae in the mid and upper thoracic spine is one of the major challenges in identifying VF in children [2; 3]. Reasons for poor visibility include the summation caused by intrathoracic tissues and shoulders; poor image quality; and patient positioning. Therefore, the patient positioning protocol and radiographic parameters selected for imaging larger patients play an important role in improving image quality and visibility, in order that upper thoracic vertebrae can be assessed. In this regard, it should be noted that lateral spine DXA allows improved visibility of the upper thoracic spine compared to radiographs [7], which may account for the improved observer reliability of SpineAnalyzer in the study by Crabtree et al [8] compared to this current study. Finally, variability in observer reliability may be related to differences in identifying T12/L1. In future studies, this limitation can be countered by having a marker placed adjacent to an agreed vertebra so that all observers recognize the same vertebral levels.

Compared to the consensus read of the radiological experts, overall sensitivity of the semi-automated 6-point technique was only 18% (95%CI of 14 – 22) while overall specificity was 97% (95%CI of 97 – 98). These findings are likely a result of a high degree of subjectivity in placing the original six semi-automated points used by the software to identify VF. This is despite the training given prior to commencing the study. The sensitivity results may also be low because identifying VF using SpineAnalyzer is based only on the loss of height of vertebral bodies, while the sABQ method is a visual method which considers alterations in the vertebral endplates that may be non-fracture related. Interpretation of SpineAnalyzer measurements is based on a grading system derived from analysis of thoracolumbar spine radiographs of 57 postmenopausal women and developed for adults [8]. Nevertheless, the Genant scoring system has been used with satisfactory results in a number of pediatric studies [18,19] and therefore we suggest that the placement of only 6 points is insufficient to capture vertebral morphometry in children and placement of further points may be required.

Another factor that affects sensitivity of the software is observer skill and experience. Although in theory no medical knowledge/specialized skills are required to identify the four corners of the vertebral bodies and center of inferior and superior endplates, small differences in placement affect the overall height ratios and factors confounding point placement and/or fracture categorization include visibility of vertebrae, early ossification of apophyses, physiological wedging and non-fracture related irregularities of vertebral endplates. Observers in this study included a musculoskeletal consultant radiologist, 2 radiographers and 2 medical students. Despite the training received, the disparate experience of the observers may be a weakness of the study, particularly given the confounding influence of physiological variations on point placement. This will need to be considered if such programs are to be used for role extension. If the 6-point or any semi-automated systems are to be more accurate and reliable, then a precise algorithm is required describing where the points should be placed if, for example, the apophyses are unossified and having ossified, prior to fusion. The difficulty in reproducible point-placement is also reflected by the low intraobserver reliability, even for the experienced radiologist. While the purpose of this current study was specifically to address the reliability of SpineAnalyzer amongst non-radiologists, in retrospect, and particularly given the poor observer reliability, it would have been interesting to have recruited and compared the results of at least two pediatric (or musculoskeletal) radiologists. This limitation of the current study is a future objective.

We conclude that although it appears useful in adults, from whose radiographs and for whom it was developed, due to its low inter and intraobserver reliability and sensitivity, currently the six-point technique comparing vertebral height ratios is neither satisfactorily accurate nor reliable for VF diagnosis in children. We suggest that the system needs training on pediatric images, with a specific algorithm designed to determine point placement, incorporate overall vertebral body shape and that the classification be based on a grading system specifically designed to differentiate physiological variation from VF.

Acknowledgment:

[redacted]

References

1 Bishop N, Arundel P, Clark E, et al. Fracture Prediction and the Definition of Osteoporosis in Children and Adolescents: The ISCD 2013 Pediatric Official Positions. J Clin Densitom. 2014;17(2):275-280.

2 Siminoski K, Lentle B, Matzinger MA, Shenouda N, Ward LM, Canadian SC. Observer agreement in pediatric semiquantitative vertebral fracture diagnosis. Pediatr Radiol. 2014;44(4):457-466.

3 Kyriakou A, Shepherd S, Mason A, Ahmed SF. A critical appraisal of vertebral fracture assessment in paediatrics. Bone. 2015;81:255-259.

4 Huber AM, Gaboury I, Cabral DA, et al. Prevalent Vertebral Fractures Among Children Initiating Glucocorticoid Therapy for the Treatment of Rheumatic Disorders. Arthritis Care & Research. 2010;62(4):516-526. doi: 10.1002/acr.20171. PubMed PMID: 20391507; PubMed Central PMCID: PMC3958950.

5 Halton J, Gaboury I, Grant R, et al. Advanced Vertebral Fracture Among Newly Diagnosed Children With Acute Lymphoblastic Leukemia: Results of the Canadian Steroid-Associated Osteoporosis in the Pediatric Population (STOPP) Research Program. J Bone Miner Res. 2009;24(7):1326-1334. Epub 2009/02/13. doi: 10.1359/jbmr.090202 10.1359/jbmr.090202 [pii]. PubMed PMID: 19210218.

6 Shaw NJ. Management of osteoporosis in children. Eur J Endocrinol. 2008;159:S33-S39.

7 [redacted].

8. Crabtree N, Chapman S, Hogler W, Hodgson K, Chapman D, Bebbington N, Shaw NJ, Vertebral fractures assessment in children: Evaluation of DXA imaging versus conventional spine radiography Bone (2017, In Press) doi http://dx.doi.org/10.1016/j.bone.2017.01.006.

9 Genant HK, Wu CY, Vankuijk C, Nevitt MC, Vertebral Fracture Assessment Using A Semiquantitative Technique. J Bone Miner Res. 1993;8(9):1137-1148.

10 [redacted]

11 R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/. Accessed 25 January 2017

12 Landis JR, Koch GG, Measurement Of Observer Agreement For Categorical Data. Biometrics 1977;33(1):159-174.

13 Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess. 1994;6(4):284.

14 Ferrar L, Jiang G, Schousboe JT, DeBold CR, Eastell R, Algorithm-based qualitative and semiquantitative identification of prevalent vertebral fracture: Agreement between different readers, imaging modalities, and diagnostic approaches. J Bone Miner Res. 2008;23(3):417-424.

15 Kim YM, Demissie S, Eisenberg R, Samelson EJ, Kiel DP, Bouxsein ML. Intra-and inter-reader reliability of semi-automated quantitative morphometry measurements and vertebral fracture assessment using lateral scout views from computed tomography. Osteoporos Int. 2011;22(10):2677-2688.

16 Oei L, Ly F, El Saddy S, et al. Multi-functionality of computer-aided quantitative vertebral fracture morphometry analyses. Quant Imaging Med Surg. 2013;3(5):249-255.

17 Birch C, Knapp K, Hopkins S, Gallimore S, Rock B, SpineAnalyzer (TM) is an accurate and precise method of vertebral fracture detection and classification on dual-energy lateral vertebral assessment scans. Radiography. 2015;21(3):278-281.

18 Lentle B, Ma J, Jaremko JL, et al. The Radiology of Vertebral Fractures in Childhood Osteoporosis Related to Glucocorticoid Administration. J Clin Densitom. 2016;19(1):81-88.

19 Cummings EA, Ma J, Fernandez CV, et al. Incident Vertebral Fractures in Children With Leukemia During the Four Years Following Diagnosis. J Clin Endocrinol Metab. 2015;100(9):3408-3417.

Tables

Table 1. Genant grading system for vertebral fracture (VF) [9]

Table 2. Simplified algorithm based qualitative scoring system [10]

Table 3: Sensitivity, specificity interobserver (kappa) and intraobserver (ICC) reliability of SpineAnalyzer for vertebral fracture diagnosis in children

Table 4. Summary of diagnostic accuracy and observer reliability of SpineAnalyzer in children

Figure Legends

Fig. 1 Lateral thoracolumbar spine radiograph, illustrating the six semi-automatically identified points used to outline the vertebral bodies and the deformity result produced by the SpineAnalyzer program

Fig. 2 Number of readable vertebrae for each observer. There is a trend towards increasing readability from the upper thoracic to the lumbar spine

Fig. 3a Observer agreement: all five observers identified a severe T8 fracture. Similarly, the T11 fracture was identified by all, but graded as mild by two observers, moderate by one and severe by two

Fig. 3b Lack of observer agreement: T5 - T7 were deemed non-evaluable by one observer and graded as no fractures by one observer, mild fractures by two and moderate fractures by one

Fig. 4 Effect of minor alterations in point placement for T11 in the same patient in which there is early apophyseal ossification. 4a (no manipulation), b (posterior manipulation) and c (middle manipulation) were classified by SpineAnalyzer as normal, while 4d (anterior manipulation) was scored by SpineAnalyzer as a mild fracture

Fig. 5a False positive SpineAnalyzer result. Wedging of T7 and T8 as indicated by SpineAnalyzer was reported by the consensus expert panel as physiological, rather than pathological wedging

Fig. 5b False negative SpineAnalyzer result. T11, T12 and L2 were reported by the consensus expert panel as fractured but were scored normal by SpineAnalyzer