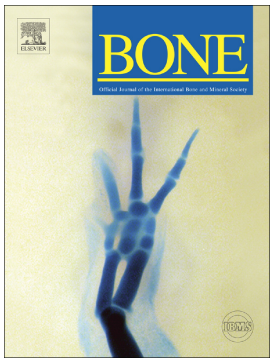
Diagnostic performance of morphometric vertebral fracture analysis (MXA) in children using a 33-point software program

Alqahtani, Fawaz F.; Crabtree, Nicola J.; Bromiley, Paul A.; Cootes, Timothy; Broadley, Penny; Lang, Isla; Offiah, Amaka C.

DOI:
10.1016/j.bone.2020.115249

Document Version
Peer reviewed version

Citation for published version (Harvard):
Alqahtani, FF, Crabtree, NJ, Bromiley, PA, Cootes, T, Broadley, P, Lang, I & Offiah, AC 2020, 'Diagnostic performance of morphometric vertebral fracture analysis (MXA) in children using a 33-point software program', *Bone*, vol. 133, 115249, pp. 1-8. https://doi.org/10.1016/j.bone.2020.115249



Diagnostic performance of morphometric vertebral fracture analysis (MXA) in children using a 33-point software program

Fawaz F. Alqahtani, Nicola J. Crabtree, Paul A. Bromiley, Timothy Cootes, Penny Broadley, Isla Lang, Amaka C. Offiah

PII: S8756-3282(20)30029-6

DOI: https://doi.org/10.1016/j.bone.2020.115249

Reference: BON 115249

To appear in: *Bone*

Received date: 18 September 2019

Revised date: 14 January 2020

Accepted date: 19 January 2020

Please cite this article as: F.F. Alqahtani, N.J. Crabtree, P.A. Bromiley, et al., Diagnostic performance of morphometric vertebral fracture analysis (MXA) in children using a 33-point software program, *Bone*(2018), https://doi.org/10.1016/j.bone.2020.115249

Diagnostic performance of morphometric vertebral fracture analysis (MXA) in children using a 33-point software program

Fawaz F Alqahtani[1,2], Nicola J Crabtree[3], Paul A Bromiley[4], Timothy Cootes[4], Penny Broadley[5], Isla Lang[5], Amaka C Offiah[1,5]

Correspondence: Professor Amaka C Offiah

University of Sheffield

Academic Unit of Child Health

Damer Street Building

Western Bank

Sheffield

S10 2TH

Phone: 0114 271 7557

E–mail: a.offiah@sheffield.ac.uk

1 Academic Unit of Child Health, University of Sheffield, Sheffield, UK

2 Department of Radiological Sciences, College of Applied Medical Sciences, Najran University, Najran, Kingdom of Saudi Arabia (Permanent address)

3 Department of Endocrinology and Diabetes, Birmingham Women's and Children's NHS Foundation Trust, Birmingham, UK

4 Division of Informatics, Imaging and Data Sciences, University of Manchester, Manchester, UK

5 Radiology Department, Sheffield Children's NHS Foundation Trust, Sheffield, UK

Declarations of Interest

None

Diagnostic performance of morphometric vertebral fracture analysis (MXA) in children using a 33-point software program

Abstract

Background

There is significant inter and intraobserver variability in diagnosing vertebral fractures in children.

Purpose

We aimed to evaluate the diagnostic accuracy of morphometric vertebral fracture analysis (MXA) using a 33-point software program designed for adults, on dual-energy x-ray absorptiometry (DXA) images of children.

Materials and Methods

Lateral spine DXA images of 420 children aged between 5 and 18 years were retrospectively reviewed. Vertebral fracture assessment (VFA) by an expert pediatric radiologist using Genant's semiquantitative scoring system served as the gold standard. All 420 DXA scans were analyzed by a trained radiographer, using semi-automated software (33-point morphometry). VFA of a random sample of 100 DXA was performed by an experienced pediatric clinical scientist. MXA of a random sample of 30 DXA images were analyzed by three pediatric radiologists and the pediatric clinical scientist. Diagnostic accuracy and inter and intraobserver agreement (kappa statistics) were calculated.

Results

Overall sensitivity, specificity, false positive (FP) and false negative (FN) rates for the radiographer using the MXA software were 80%, 90%, 10%, and 20% respectively and for mild fractures alone were 46%, 92%, 8%, and 54% respectively. Overall sensitivity, specificity, FP, and FN rates for the four additional observers using MXA were 89%, 79%, 21%, and 11% respectively and for mild fractures alone were 36%, 86%, 14%, and 64% respectively. Agreement between two expert observers was fair to good for VFA and MXA [kappa = 0·29 to 0·76 (95% CI: 0·17 – 0·88) and 0·29 to 0·69 (95% CI: 0·17 – 0·83)] respectively.

Conclusion

MXA using a 33-point technique developed for adults is not a reliable method for the identification of mild vertebral fractures in children. A pediatric standard is required which not only incorporates specific vertebral body height ratios but also the age-related physiological changes in vertebral shape that occur throughout childhood.

Keywords

Vertebral fracture assessment; Fracture; Pediatric; DXA; Morphometric vertebral fracture analysis

1. Introduction

Osteoporotic fractures may occur in children and adolescents with low bone mineral density (BMD) either as a primary condition (e.g., osteogenesis imperfecta),(1) or secondary to various disorders and medications including acute lymphoblastic leukemia, rheumatic disorders, inflammatory bowel disease, Duchenne muscular dystrophy, and glucocorticoid therapy.(2,3)

Vertebral fractures represent a significant proportion of all osteoporotic fractures and thus, given a lack of major trauma or local disease, presentation with one or multiple vertebral fractures is a strong indicator of bone fragility in children and is the basis on which osteoporosis in this age group is defined.(4) Most vertebral fractures are not identified clinically, which may be problematic, given the high levels of morbidity they may be associated with. Children with risk factors for low BMD are regularly screened to assess prevalent and incident vertebral fractures, because BMD itself is not predictive of presence of vertebral fractures. Indeed, some specialist groups have formalized annual spine imaging through inclusion in their guidelines e.g. Duchenne muscular dystrophy,(5) and following bone marrow transplantation.(6) Thus, a method of accurate detection of these fractures must be devised to allow prompt therapeutic intervention.

Until very recently, lateral spine radiographs were the main method for identifying vertebral fractures. However, the latest bone densitometers have made it possible to conduct vertebral fracture assessment (VFA) from dual-energy X-ray absorptiometry (DXA) scans. This technique is considered preferable due to similar (although poor) sensitivity and specificity when compared to radiographs, as well as the advantage of reduced radiation dose.(7-10) The available scoring systems for VFA in adults have also been evaluated for utilization in children: these systems include Genant's semiquantitative technique (SQ);(2, 9,11) the algorithm–based qualitative (ABQ) technique,(12) and software programs that allow morphometric analysis (MXA).(8-10,13) Results have been variable, with the largest studies showing low diagnostic accuracy of VFA and MXA, particularly for mild fractures, which are the most important to detect in order to prevent the complications associated with progression.(5,6,11)

The newest generation of bone densitometers are capable of enhancing the diagnostic utility of DXA through integration with semi–automated software that helps to diagnose vertebral fractures. In terms of recent refinements to MXA, the shape-based statistical modelling technique for semi-automated quantitative morphometry has been devised for detection of fractures in adults,(14) and this technical development may also improve analysis in children in terms of efficiency and accuracy.

The aim of this study was to evaluate the diagnostic accuracy of MXA through the use of a novel semi-automated 33-point morphometric software tool, "AVERT™", in a cohort of children with chronic disease, using the latest iDXA imaging technology in the hands of various observers compared to the reference standard of a visual SQ method applied by an experienced pediatric radiologist.

2. Materials and Methods

2.1 Study population

The Picture Archiving and Communication System of Sheffield Children's Hospital was searched for all lateral spine iDXA images performed between November 2011 and November 2016 in children aged between five and 18 years old. All 2800 images were divided into yearly cohorts based on age and 15 lateral spine iDXA images were randomly selected for each year of age and both sexes, giving

a total of 420 iDXA lateral spine images which were anonymised and included in the study. Bone mineral density (BMD) for both lumber spine (L2–L4) and total body less head (TBLD) were performed as part of the same investigation and allowed comparison of diagnostic accuracy of MXA in relation to BMD. DXA results were automatically generated from the Lunar GE iDXA scanner, and Z-scores matched for sex, age and ethnicity which (according to the manufacturer) are based on the UK AP Spine Reference Population (V12).

2.2 Ethics statement

The study protocol was approved by the Local Health Research Authority (HRA reference number: 210524). Informed consent was not obtained as we only reviewed hospital notes and existing DXA images as part of clinical care of patients. The study was also registered with the local Research and Innovation Department and conducted in accordance with the Declaration of Helsinki and the NHS Research Governance Framework.

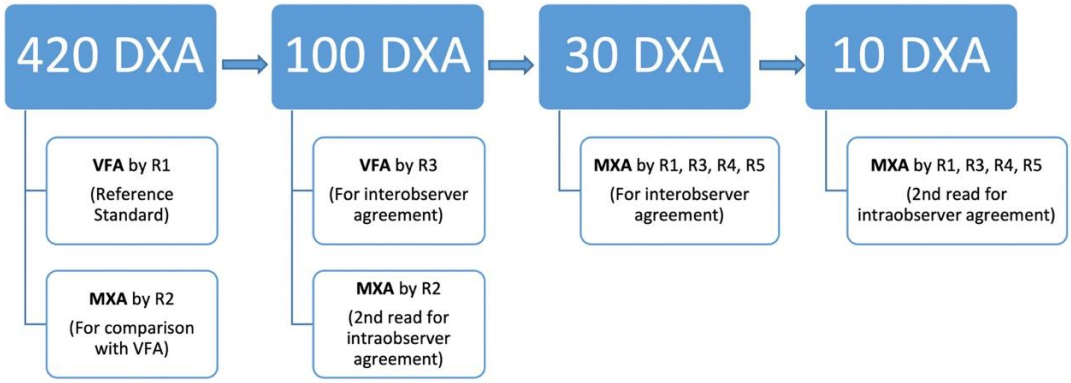
2.3 Lateral spine imaging

Lateral spine DXA scans were acquired using a Lunar GE iDXA machine (GE Healthcare Lunar iDXA, Buckinghamshire, UK), following the manufacturer's recommendations. Briefly, the child was positioned in the decubitus position on the scanning table, with their knees flexed upwards towards the chest, so that the spine was parallel to the table with their arms above their head and away from the area to be scanned. Foam padding was used to obtain and maintain the required position.

2.4 Image analysis

All images were analyzed using AVERT™ (Optasia Medical, Cheadle Hulme, Greater Manchester, UK). AVERT™ is a software program based on a 33-point morphometric technique and uses the latest appearance modelling technology (random forest regression voting constrained local models) developed by the University of Manchester.(15) Figure 1 is a flow chart of the reporting pathway described in more detail below.

Figure 1. Flow chart summarizing the reporting pathway



2.4.1 *Reference standard (420 VFA, R1)*

For the reference standard, identification of vertebral fractures was performed on the 420 VFA by visually assessing the T4 to L4 vertebrae, relying on an experienced pediatric radiologist (R1) i.e., with no software involved, as is the current clinical standard. Quantitative measurements only took place at the reader's discretion. Vertebrae were categorized by this visual semi–quantitative (SQ) method as 0 "non-fractured", 1 "mild fracture", 2 "moderate fracture", and 3 "severe fracture" based on Genant's classification.(16) Grades 0, 1, 2, and 3 entail loss of height of ≤20%, 21% to 25%, 26% to 40%, and ≥ 41% respectively. Vertebral fractures are manifested by a variety of alterations in shape, including "wedge", "biconcavity", or "crush", depending on the site of maximum reduction in vertebral height (anterior, middle, or generalized respectively). Additionally, vertebrae diagnosed as fractured but with ≤ 20% reduction in height and vertebrae with loss of height diagnosed as being related to physiological wedging were reported by the pediatric radiologist (on the basis of 20 years' experience in pediatric radiology).

For consistency of vertebral level detection between observers, prior to study commencement at the stage of image anonymization, R2 placed a marker at T12 for all images, confirmed by R1. The lowermost vertebral body associated with a pair of ribs was always designated as T12.

2.4.2 *Diagnostic accuracy of MXA (420 iDXA, R2)*

A radiographer (R2) used AVERT™ to perform MXA on the 420 selected DXA images. Prior to commencing the study, the radiographer was trained to use the software program by experts from the University of Manchester, who participated in developing the software (the training was provided by a research fellow in computer vision and an expert radiologist, using 72 non-study spine images).

2.4.3 *Intraobserver agreement of MXA (100 iDXA, R2)*

To evaluate intraobserver agreement of MXA for R2, DXA images of 100 subjects were randomly selected from the study population for a second read. In order to reduce recall bias, the repeat scoring was performed after an interval of approximately 30 days.

Interobserver agreement of VFA (100 iDXA, R3)

To evaluate interobserver agreement of VFA, an experienced pediatric clinical scientist (R3) independently used the SQ grading scale for visual assessment (VFA) of the same 100 iDXA used for R2's second read. The results were compared to the reference standard to assess interobserver agreement of VFA.

2.4.4 *Observer agreement of MXA (30 iDXA, R1, R3, R4, R5)*

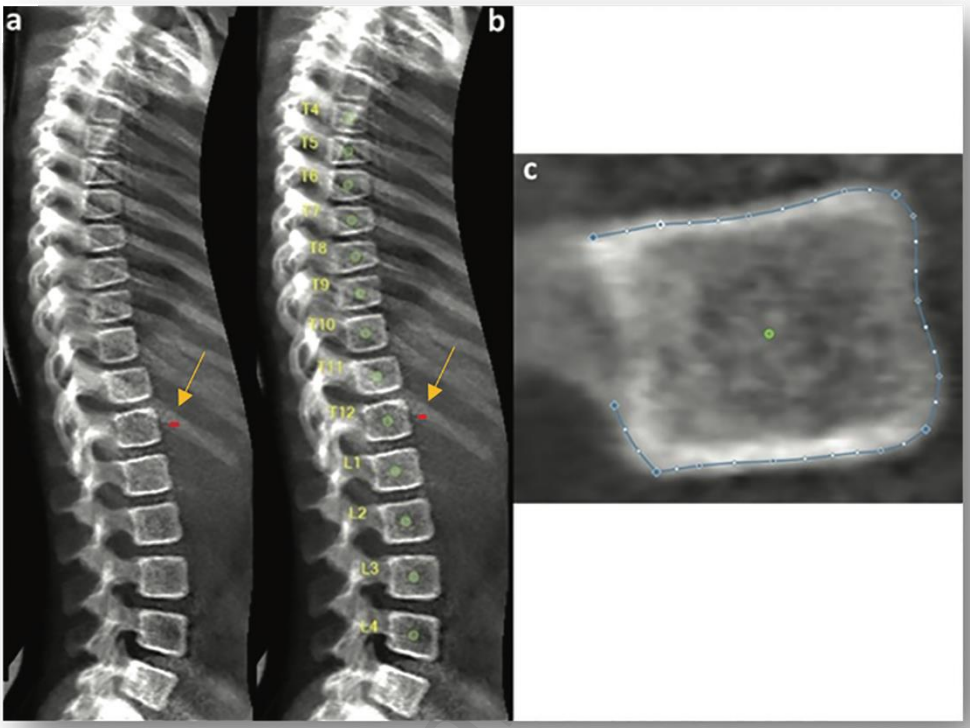
To ascertain observer agreement of MXA more widely, three consultant pediatric musculoskeletal radiologists (R1, R4, R5), each with a minimum of 13 years' experience, and an experienced clinical scientist (R3), independently performed MXA on 30 iDXA images randomly selected from the 100 interpreted by R2. Images were analyzed in random order without accessing the subject's clinical information, and also blinded to any previous analyses. Following an interval of at least 2 weeks, 10 of the 30 iDXA images were randomly selected for a second read by the same four observers to allow calculation of intraobserver agreement of MXA.

Sensitivity, specificity, false positive and false negative rates were calculated for all grades of fracture and for mild fractures alone.

2.5 Morphometric analysis technique

The first step in MXA required the observers to identify all vertebrae from T4 to L4 by manually placing a single point at the center of each vertebral body, then the software identified the vertebral bodies accordingly (i.e. T4 as the highest and L4 as the lowest vertebra) (Figure 2).

Figure 2. Technique used to perform semi-automated quantitative morphometry measurements (AVERT™)



a) Lateral iDXA scan of the entire spine of a 9 -year-old female with osteogenesis imperfecta; b) identified vertebral bodies from T4 to L4; c) 33 points placed to outline T12. The arrow points to the T12 marker that ensured consistency between readers for vertebral level identification (lowest vertebral body associated with a rib).

Subsequently, the program automatically outlined each labelled vertebra with 33 measurement points: eleven on the upper endplate, eight on the anterior margin, eleven on the lower endplate, and three on the posterior margin (leading to 33 points for each vertebral body). The observers reviewed the images and, if necessary, modified these points. From these confirmed points, the software then computed the anterior, middle and posterior (h^a, h^m and h^p) heights and calculated the wedge ratio (h^a/ h^p), biconcave ratio (h^m/h^p) and crush ratio (h^p/h^{p+2} or h^p/h^{p-2}), where +2 and -2 indicate the four neighboring vertebrae, i.e. the two immediately above [+2] and the two immediately below [-2] the vertebra under examination. Based on the semi–quantitative (SQ) scoring system developed by Genant, vertebrae were classified according to their height loss ratios as normal or mild, moderate, or severe fracture for height loss of < 21%, 21%–25%, 26%–40% and ≥41% respectively.

2.6 Statistical analysis

We report demographic and bone densitometry data (bone mineral density (BMD, g/cm2) and Z-score for both L2–L4 and TBLH). The frequency of vertebral fracture severity for each observer and for all vertebrae from T4 to L4 was calculated. Inter and intraobserver agreement and associated 95% confidence intervals (CI) were calculated using the kappa statistic. Diagnostic accuracy of observers (sensitivity, specificity, false positive and false negative rates) was calculated. Analyses were performed both at the subject and at the individual vertebral level. A previous survey of 14 members of the British Paediatric and Adolescent Bone Group showed that the majority would instigate treatment only in the presence of one or more vertebrae with height loss of greater than 25% PLUS pain (7). Therefore, we analyzed prevalent vertebral fractures in three groupings: (1) Any fracture (mild, moderate and severe), (2) clinical fracture (moderate and severe), and (3) mild fracture. Vertebral levels that could not be visualized were excluded from analysis.

Statistical analyses were conducted using SPSS statistics software version 24 (IBM, Armonk, NY, USA) and Microsoft® Excel 2016.

3. Results

We included 420 lateral iDXA scans in children aged between 5 and 18 years (30 per year of age being the typical number used to train software); 210 (50 %) were male; 380 (90%) had osteogenesis imperfecta, 12 (3%) Duchenne muscular dystrophy, 8 (2%) polyostotic fibrous dysplasia, and 20 (5%) other conditions including anorexia nervosa, diabetes mellitus, juvenile dermatomyositis and coeliac disease. Descriptive and clinical data are presented in Table 1.

Table 1. Summary of demographic and bone densitometry data of study subjects (mean and SD), n = 420*

Age (Year)	Height (cm)	SD	Weight (kg)	SD	L2–L4 BMD (g/cm^3)	SD	L2–L4 BMD (z–score)	SD	TBLH BMD (g/cm^3)	SD	TBLH BMD (z–score)	SD
5	105·5	10·85	17·42	3·30	0·559	0·13	-0·857	1·79	0·489	0·16	-0·687	0·72
6	111·85	8·90	20·07	3·76	0·616	0·10	-0·739	1·35	0·601	0·08	-0·478	0·90
7	116·77	13·57	23·91	6·28	0·669	0·11	-0·306	1·37	0·661	0·07	-0·106	1·19
8	126·37	8·53	30·44	11·39	0·735	0·24	-0·120	1·75	0·712	0·13	0·093	1·48
9	133·31	12·68	32·95	13·38	0·668	0·07	-0·736	1·05	0·659	0·11	-0·927	1·17
10	138·04	8·23	35·12	8·62	0·737	0·13	-0·507	1·35	0·763	0·10	-0·433	1·11
11	142·23	10·99	38·75	16·78	0·867	0·33	-0·538	1·40	0·802	0·12	-0·377	1·13
12	143·72	11·73	40·30	10·75	0·824	0·12	-0·908	0·80	0·805	0·13	-0·758	0·97
13	156·17	9·46	47·26	10·95	0·863	0·35	-0·127	1·79	0·902	0·16	-0·227	1·53
14	158·34	11·09	52·34	20·75	0·948	0·23	-0·360	1·51	0·966	0·12	-0·260	1·21
15	160·21	9·90	49·74	5·14	1·080	0·20	-0·442	1·39	0·970	0·11	-0·567	0·92
16	161·49	6·08	60·91	13·31	1·147	0·23	-0·287	1·77	0·969	0·12	-0·34	1·49
17	165·38	9·04	58·95	11·62	1·111	0·15	-0·806	1·20	0·950	0·10	-0·863	0·78
18	166·06	9·50	60·23	7·70	1·057	0·21	-1·217	1·63	0·950	0·10	-0·958	1·07

** 15 females and 15 males in each age group*

BMD= bone mineral density, TBLH = total body less head

3.1 *Diagnostic accuracy of MXA (420 iDXA)*

Vertebral fracture assessment (VFA) of 5460 individual vertebrae was performed by R1 using the visual SQ method (this was the gold standard read) and by R2 using the 33-point MXA technique; of these, 4% were not evaluable by either method because of either poor visualization or poor image quality, including movement artefact. The majority of unevaluable vertebrae for both techniques were located in the upper thoracic spine (Figure 3) and were unrelated to BMD (Tables 2 and 3).

Figure 3. Total number of unevaluable vertebrae for VFA=231 (4%) and MXA (AVERT™) =243 (4%)

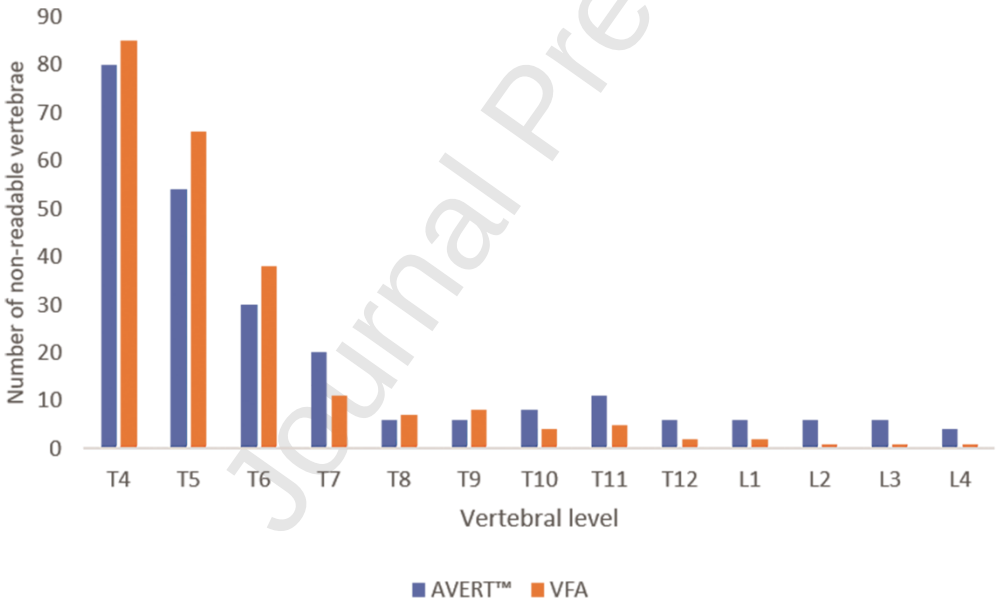


Table 2. Prevalence of vertebral fractures in the study cohort (n= 420 patients, 5460 vertebrae)

	VFA (R1) = gold standard		**MXA– AVERT™ (R2)**	
	Per vertebra	***Per subject***	***Per vertebra***	***Per subject***
No fracture	4564 (81%)	191 (45%)	4412 (78%)	157 (37%)
Mild fracture (21% to 25% loss of height)	216 (4%)	98 (23%)	441 (8%)	204 (49%)
Moderate fracture (26% to 40% loss of height)	124 (2%)	67 (16%)	317 (5%)	155 (37%)
Severe fracture (≥ 41% loss of height)	54 (1%)	29 (7%)	47 (1%)	27 (7%)
Non-readable vertebra	231 (4%)	80 (19%)	243 (4%)	98 (23%)
Fractures	77 (1%)	32 (7%)	N/A	N/A

(loss of height ≤ 20%)*				
Physiological wedge	136 (3%)	35 (8%)	N/A	N/A
Possible fracture	58 (2%)	14 (14%)	N/A	N/A

** A height reduction of ≤ 20% that was nevertheless considered to represent a fracture rather than normal variation*

Table 3. Prevalence of mild vertebral abnormality per BMD Z-score category

Diagnosis (R1)	**Total Number Vertebrae (%) including T4 to L4 per BMD Z Score Category**				**Total N = 420**
	≤ -2.0 SD* N = 46**	**-1.9 to 0.0 SD N = 249**	**0.1 to 1.9 SD N = 102**	**≥ 2.0 SD N = 23**	
Normal	472 (87%)	2843 (94%)	1187 (95%)	236 (89%)	4738 (93%)
Physiological Wedging	36 (6.5%)	87 (3%)	16 (1%)	0 (0%)	139 (3%)
Mild Fracture	36 (6.5%)	99 (3%)	52 (4%)	29 (11%)	216 (4%)
Total	544 (100%)	3029 (100%)	1255 (100%)	265 (100%)	5093 (100%)

A. Gold standard VFA

Diagnosis (R2)	**Total Number Vertebrae (%) including T4 to L4 per BMD Z Score Category**				**Total N = 420**
	≤ -2.0 SD* N = 46**	**-1.9 to 0.0 SD N = 249**	**0.2 to 1.9 SD N = 102**	**≥ 2.0 SD N = 23**	
Normal	445 (89%)	2672 (91%)	1087 (91%)	246 (92%)	4450 (91%)
Mild Fracture	54 (11%)	272 (9%)	107 (9%)	22 (8%)	455 (9%)
Total	499 (100%)	2944 (100%)	1194 (100%)	268 (100%)	4905 (100%)

B. MXA

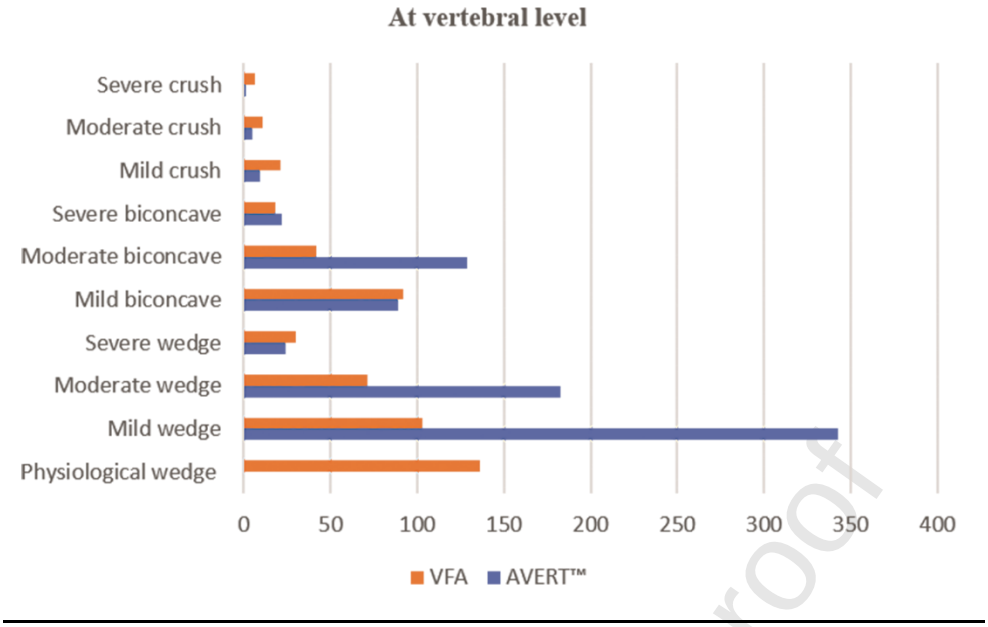
** SD = standard deviation*

*** N = total number of patients in each category*

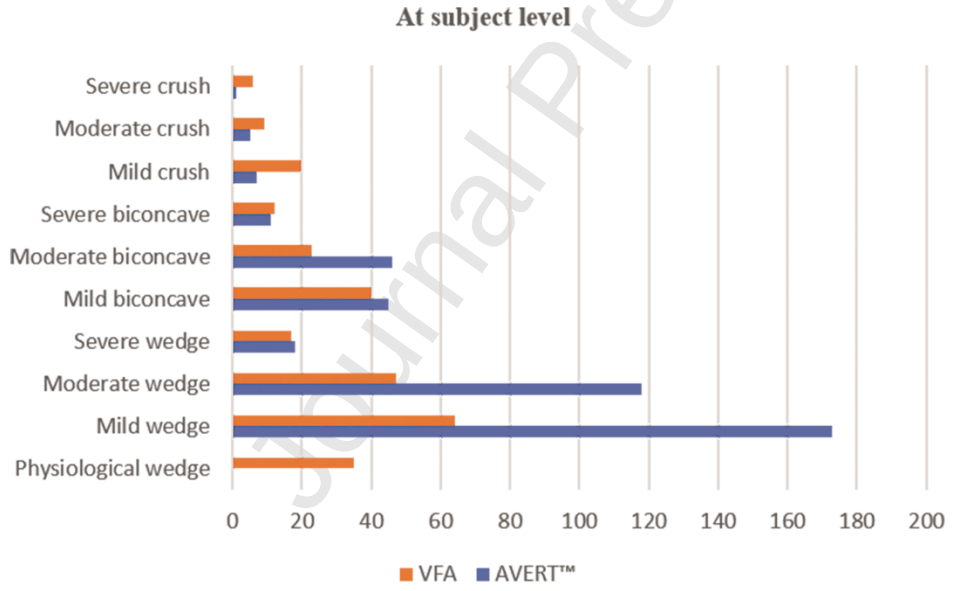
Among the 420 subjects, 191 (45%) had no fracture by the gold standard visual SQ method, while mild, moderate and severe fractures were identified in 98 (23%), 67 (16%), and 29 (7%) subjects respectively. Isolated physiological wedging (with no fracture) was identified in 35 (8%) children. MXA identified more children with mild and moderate vertebral fractures than the gold standard but almost the same number of severe vertebral fractures. Table 2 shows the number and grading of the evaluated vertebrae by the two techniques.

Figures 4a and 4b show the number, severity, and shape of vertebral fractures by the two methods at the vertebral and subject levels respectively, as well as the physiological wedges identified by VFA.

Figure 4. Number of vertebral fracture shapes identified using both techniques



(a)

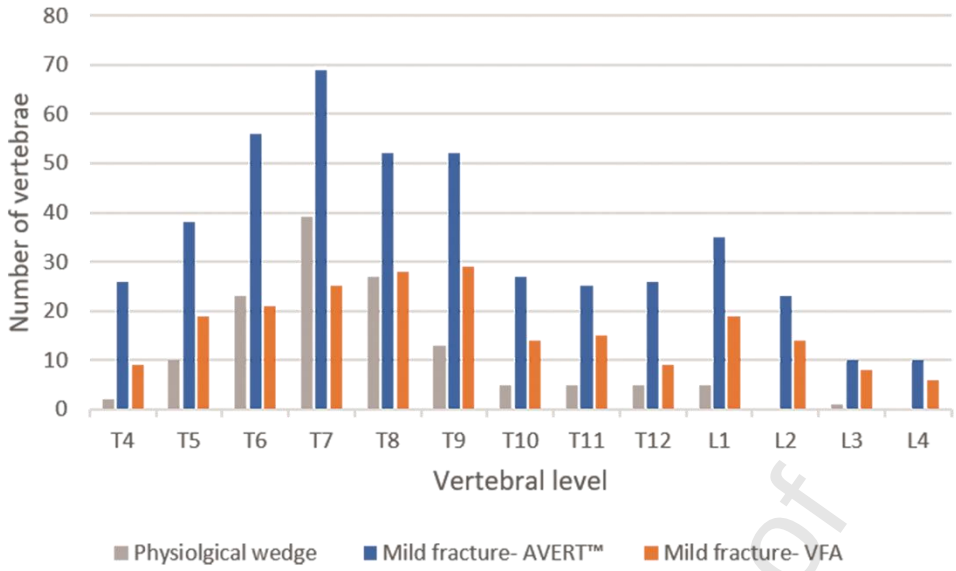


(b)

(a) at vertebral level and (b) at subject level (note that AVERT™ does not have the ability to diagnose physiologically wedged vertebrae)

The location of mild fractures and physiologically wedged vertebrae is shown in Figure 5.

Figure 5. Number and location of mild vertebral fractures identified by both techniques compared to number of physiologically wedged vertebrae identified by VFA.



The figure illustrates that in the mid-thoracic region, the number of mild fractures identified by AVERT™ was comparable to the sum of the mild fractures and physiological wedges identified by the visual SQ method (e.g. at T7 and T8, AVERT™ identified 69 and 52 mild fractures, respectively; whereas the sum of the mild fractures and physiologically wedged vertebrae identified by VFA were 64 and 54, respectively).

The diagnostic accuracy and observer agreement of AVERT™ for the "any fracture" (≥ 21% loss of height), "clinical fracture" (≥26% loss of height), and "mild fracture" (21% to 25% loss of height) groups are presented in Table 4.

Table 4. Diagnostic accuracy of MXA for detecting vertebral fractures (n = 420 patients, 5460 vertebra)

Vertebral Level	Number of Evaluable Vertebra /iDXA Scans	Any fracture				Clinical fracture*				Mild fracture**			
		Sensitivity	Specificity	Kappa	95% CI	Sensitivity	Specificity	Kappa	95% CI	Sensitivity	Specificity	Kappa	95% CI
T4	332	16/19 (84%)	289/313 (92%)	0·50	0·33 – 0·66	4/4 (100%)	318/328 (97%)	0·43	0·12 – 0·69	7/9 (79%)	304/323 (94%)	0·37	0·21 – 0·55
T5	364	27/34 (79%)	286/330 (87%)	0·44	0·31 – 0·57	8/10 (80%)	329/354 (93%)	0·34	0·15 – 0·52	3/19 (16%)	310/345 (90%)	0·03	-0·17 – 0·28
T6	388	39/45 (87%)	289/343 (84%)	0·48	0·37 – 0·58	15/19 (79%)	347/369 (94%)	0·49	0·30 – 0·65	8/21 (38%)	320/369 (87%)	0·14	0·01 – 0·29
T7	400	46/52 (88%)	273/348 (78%)	0·42	0·32 – 0·52	18/23 (78%)	343/377 (90%)	0·42	0·28 – 0·55	11/25 (44%)	319/375 (85%)	0·15	0·03 – 0·33
T8	404	45/52 (86%)	290/352 (82%)	0·46	0·36 – 0·56	19/20 (95%)	325/384 (84%)	0·32	0·20 – 0·44	7/28 (25%)	336/376 (89%)	0·09	-0·14 – 0·21
T9	411	32/52 (64%)	317/361 (88%)	0·41	0·29 – 0·52	12/17 (71%)	383/394 (97%)	0·53	0·32 – 0·69	11/29 (38%)	344/382 (89%)	0·20	0·09 – 0·33
T10	409	21/29 (72%)	347/380 (91%)	0·45	0·30 – 0·58	10/11 (91%)	381/398 (96%)	0·49	0·28 – 0·67	8/14 (57%)	377/395 (95%)	0·36	0·21 – 0·44

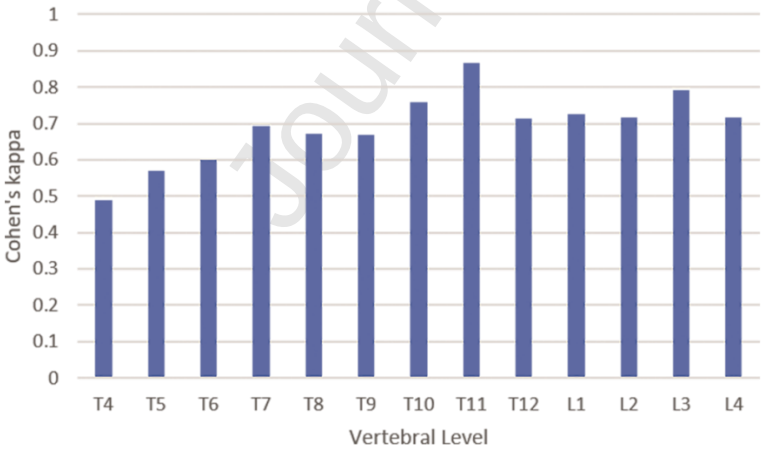
T11	407	24/27 (89%)	348/380 (92%)	0·53	0·39 – 0·65	10/10 (100%)	376/397 (95%)	0·45	0·24 – 0·62	7/15 (47%)	375/392 (95%)	0·31	0·17 – 0·48
T12	412	23/25 (92%)	357/387 (92%)	0·54	0·40 – 0·67	13/14 (93%)	385/398 (97%)	0·61	0·42 – 0·77	6/9 (67%)	384/403 (95%)	0·32	0·19 – 0·43
L1	412	39/42 (93%)	335/370 (90%)	0·63	0·52 – 0·72	18/19 (95%)	372/393 (95%)	0·58	0·41 – 0·71	11/20 (55%)	369/392 (94%)	0·36	0·22 – 0·49
L2	414	25/30 (83%)	361/384 (94%)	0·60	0·45 – 0·72	13/14 (93%)	388/400 (97%)	0·65	0·43 – 0·80	6/14 (43%)	383/400 (96%)	0·29	0·13 – 0·36
L3	413	15/22 (68%)	384/391 (98%)	0·66	0·46 – 0·82	8/10 (80%)	399/403 (99%)	0·72	0·45 – 0·89	3/8 (38%)	398/405 (98%)	0·31	0·19 – 0·48
L4	415	10/17 (59%)	389/398 (98%)	0·53	0·30 – 0·70	5/7 (71%)	404/408 (99%)	0·61	0·23 – 0·87	3/6 (50%)	402/409 (98%)	0·36	0·22 – 0·49
Average		**80%**	**90%**			**87%**	**95%**			**46%**	**92%**		
Subject level	420	131/133 (98%)	142/287 (49%)	0·37	0·30 – 0·43	72/78 (92%)	227/342 (66%)	0·43	0·35 – 0·51	93/96 (97%)	198/324 (61%)	0·41	0·27 – 0·55

**Clinical fracture = moderate and severe (≥26%) vertebral height reduction; ** Mild fracture = 21% to 25% vertebral height reduction*

3.2 *Intraobserver agreement of MXA (100 iDXA)*

There was fair to excellent intraobserver agreement, with kappa ranging from 0·49 to 0·87 (95% CI 0·37 – 0·98), with the lowest agreement level identified at T4. Figure 6 summarizes intraobserver agreement of MXA for R2.

Figure 6. Intraobserver (R2) agreement of MXA/AVERT™



3.3 *Observer agreement of MXA (30 iDXA)*

In respect to the "any fracture" grade, there was fair to good interobserver agreement between the additional four raters when they used AVERT™, with kappa ranging from 0·39 to 0·53 (95% CI 0·17 – 0·67). In contrast, there was a slightly higher agreement level when only "clinical fractures" were

considered, with kappa ranging from 0·48 to 0·67 (95% CI 0·33 – 0·78). Finally, there was poor agreement when only "mild fractures" were considered, with kappa ranging from 0·10 to 0·29 (95% CI -0·09 – 0·41). Intraobserver agreement for the same four readers for "any fracture" ranged from moderate to good, with mean kappa values for R1, R3, R4, and R5 of 0·55, 0·60, 0·68, and 0·58, respectively; for "clinical fractures", kappa values were 0·59, 0·82, 0·89, and 0·67 and for "mild fractures" kappa values were 0·67, 0·61, 0·51, and 0·58 respectively. Table 5 summarizes inter- and intraobserver agreement of MXA for the four observers.

Table 5. Summary of inter and intraobserver agreement for MXA (n=30)

Interobserver agreement	**Observer**	**Kappa**		
Any fracture (≥ 21% loss of height)		Mean	Min	Max
	R1 vs R3	0·39	0·20	0·69
	R1 vs R4	0·44	0·23	0·57
	R1 vs R5	0·53	0·38	0·73
	R3 vs R4	0·41	0·20	0·85
	R3 vs R5	0·39	0·15	0·58
	R4 vs R5	0·42	0·11	0·70
	Agreement across four observers	Fleiss' kappa = **0·44**		
Clinical fracture (≥ 26% loss of height)				
	R1 vs R3	0·50	0·30	0·76
	R1 vs R4	0·52	0·24	0·73
	R1 vs R5	0·67	0·42	0·92
	R3 vs R4	0·48	0·26	0·79
	R3 vs R5	0·56	0·36	0·96
	R4 vs R5	0·49	0·16	0·81
	Agreement across four observers	Fleiss' kappa = **0·52**		
Mild fracture (21% to 25% loss of height)				
	R1 vs R3	0·21	0·07	0·45
	R1 vs R4	0·21	0·01	0·43
	R1 vs R5	0·29	0·04	0·61
	R3 vs R4	0·19	0·04	0·56
	R3 vs R5	0·10	0·07	0·27
	R4 vs R5	0·15	0·03	0·51
	Agreement across four observers	Fleiss' kappa = **0·21**		
Intraobserver agreement	**Observer**	**Kappa**		
Any fracture (≥ 21% loss of height)				
		Mean	Min	Max
	R1	0·55	0·16	1·00
	R3	0·60	0·28	1·00
	R4	0·68	0·11	1·00
	R5	0·58	0·13	1·00
	Agreement across four observers	**0·60**	**0·11**	**1·00**
Clinical fracture (≥ 26% loss of height)				
	R1	0·59	0·19	1·00
	R3	0·82	0·44	1·00
	R4	0·89	0·56	1·00
	R5	0·67	0·18	1·00
	Agreement across four observers	**0·74**	**0·18**	**1·00**
Mild fracture (21% to 25% loss of height)				
	R1	0·67	0·21	1·00
	R3	0·61	0·11	1·00
	R4	0·51	0·01	1·00
	R5	0·58	0·01	1·00
	Agreement across four observers	**0·59**	**0·01**	**1·00**

The average sensitivity, specificity, false positive, and false negative rates for the four observers were 89%, 79%, 21%, and 11% at the vertebral and 98%, 52%, 48%, and 2% at the subject level for any fracture grade. When only mild fractures were considered, the average sensitivity, specificity, false positive, and false negative rates were 36%, 86%, 14%, and 64% at the vertebral and 88%, 35%, 65%, and 12% at the subject levels respectively.

3.4 *Observer agreement of VFA (100 iDXA)*

Of the possible total of 1300 vertebrae, from T4 to L4 (i.e. 13 vertebrae per subject in 100 subjects); 1267 (97%) were adequately visualized by R1, and 1269 (98%), and 1248 (96%) by R2 and R3 respectively. The number and severity of vertebral fractures at the vertebral and subject levels for each observer are shown in Table 6.

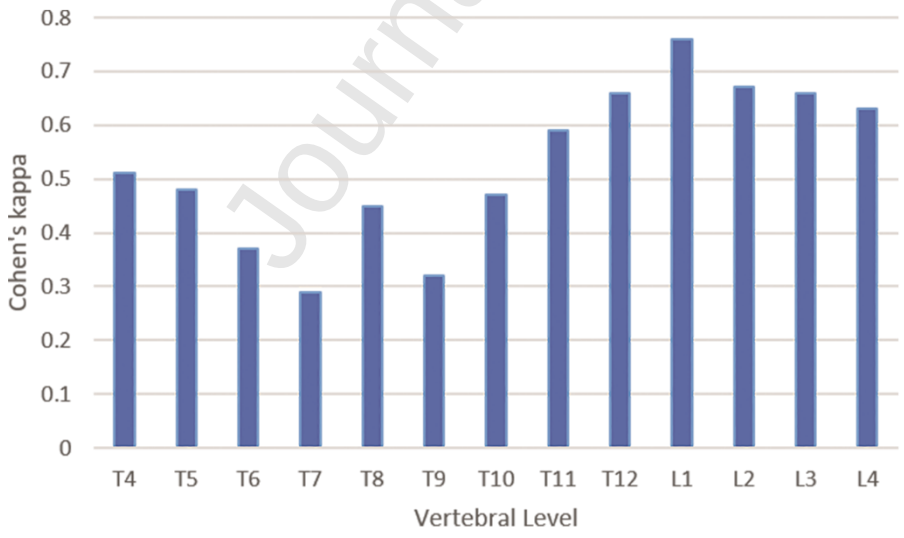
Table 6. Fracture prevalence by observer and technique for 100 randomly selected images

	VFA (R1) = Gold standard		MXA– AVERT™ (R2)		VFA (R3)	
	Per vertebra	*Per subject*	*Per vertebra*	*Per subject*	*Per vertebra*	*Per subject*
No fracture	822 (63%)	32 (32%)	902 (69%)	11 (11%)	782 (60%)	14 (14%)
Mild fracture (21% to 25% loss of height)	149 (11%)	56 (56%)	176 (14%)	70 (70%)	208 (16%)	72 (72%)
Moderate fracture (26% to 40% loss of height)	97 (7%)	35 (35%)	153 (11%)	61 (61%)	130 (10%)	45 (45%)
Severe fracture (≥ 41% loss of height)	39 (3%)	19 (19%)	38 (3%)	20 (20%)	62 (4%)	22 (22%)
Non-readable vertebrae	66 (5%)	16 (16%)	31 (2%)	19 (19%)	70 (5%)	15 (15%)
Fractures (with loss of height ≤ 20%)*	55 (4%)	25 (2%)	N/A	N/A	48 (3%)	23 (2%)

** A height reduction of ≤ 20% that was nevertheless considered to represent a fracture rather than normal variation*

Although the numbers of mild and moderate vertebral fractures varied between all observers, the number of severe fractures was comparable. A similar pattern was observed at the subject level. Figure 7 summarizes the interobserver agreement of VFA between R1 and R3.

Figure 7. Interobserver *(R1, R3)* agreement of VFA



4. Discussion

This study aimed to determine the diagnostic accuracy and inter and intraobserver agreement of morphometric vertebral fracture analysis (MXA) using a 33-point software program (designed for adults) on a large cohort of children with conditions predisposing to vertebral fracture. Results of

MXA were compared to the visual SQ technique for vertebral fracture identification from iDXA scans (VFA). Results demonstrate that MXA is only as good as VFA in identifying severe vertebral fractures with reduced diagnostic accuracy for detecting mild vertebral fractures.

The overall sensitivity, specificity, false positive, and false negative rates for R1 and R3 to R5 were 89%, 79%, 21%, and 11% at the vertebral and 98%, 52%, 48%, and 2% at the subject levels. Five previous studies that used 6-point MXA and VFA have shown sensitivity and specificity ranging from 18% to 94% and 71% to 100% respectively for MXA; and 63% to 95% and 85% to 100% respectively for VFA for analysis at the vertebral level. While at the subject level, sensitivity and specificity range from 43% to 94% and 85% to 97% respectively for MXA; and 78% to 95% and 72% to 100% respectively for VFA.(5–8,11) These results are generally lower, except for a higher specificity at subject level than has been shown by the current study. This may be due to the high number of subjects with physiological wedging in the current study by the reference standard which were diagnosed as mild fractures by MXA, thus causing an increase in the false positive rate.

The results of observer agreement of MXA in this current study are slightly higher than those of a previous study,(6) where the evaluation was conducted by three readers (an experienced clinical scientist, a senior radiographer and a clinical scientist unfamiliar with MXA).(8) In that study, kappa scores ranged from 0·13 to 0·32 when compared to VFA. On the other hand, our results show a slightly lower agreement level when compared to another recent study,(10) where kappa reached 0·79 (95% CI 0·62 – 0·92) and 0·55 (95% CI 0·40 – 0·68) at the vertebral and subject levels, respectively. It should be noted that the study was based on only 20 subjects, and the gold standard was radiographic images reported by a non-radiologist reader.(10) Another study used Hologic QDR Physician's viewer software (version 7·02) to perform MXA on lateral DXA scans of 58 children and adolescents, using six-point software. This reported higher agreement at both the vertebral and subject levels (a kappa score of 0·72 (95 % CI 0·65 – 0·78), and 0·73 (95 % CI 0·55 – 0·91) respectively) when compared to the visual SQ method using conventional radiographs and performed by two experienced skeletal radiologists.(9) Notably, no comparison was established in that study between MXA and visual SQ for VFA. Finally, our current findings are better than those of a recent study on radiographic images of 137 children, in which five observers utilized a six-point software program (SpineAnalyzer™, Optasia Medical, Cheadle, UK); kappa for interobserver reliability ranged from 0·05 to 0·47 (95% CI: -0·19 – 0·76) and the intraclass correlation coefficient for intraobserver reliability ranged from 0·25 to 0·61.(13)

Despite improvement in diagnostic accuracy of 33-point MXA compared to 6-point MXA and VFA, our results show overall low diagnostic accuracy and observer reliability when only mild fractures are considered. Our results suggest that a large contributory factor is the inability of the software to distinguish normal physiological wedging (i.e. developmental morphological variability that occurs throughout childhood) from mild fractures, particularly in the thoracic region. As a consequence, the rate of mild and moderate fracture was relatively higher for MXA than for the reference standard. Another major limitation of MXA is the inability of the software program to identify fractures when height loss is below 20%, as identified in 32 subjects (8%) in this study.

This inability to differentiate normal physiological wedging from fracture also accounts for low diagnostic accuracy of VFA (5). Software that is developed on a healthy cohort of children which incorporates relevant variables related to age may be the solution to accurate and reliable diagnosis of mild vertebral fractures in children and will help to elucidate the diagnostic criteria for "physiological wedging".

It should be pointed out that observer reliability of MXA depends on point placement, which to a large extent affects thresholds for height ratios. In other words, only a very small alteration in point placement and therefore in height ratio (that would be insignificant clinically) can lead to two different fracture categories being reported by 2 observers or by the same observer at different times (e.g. 24·9% and 25·1% loss of height will be classified as mild and moderate fractures respectively). This is particularly important at the threshold between no vertebral fracture and mild vertebral fracture. We postulated that MXA reads would be of lower accuracy or not possible in children with reduced BMD. However, the study has demonstrated that importation of images to the software program does not significantly affect DXA image quality and MXA is possible even in children with BMD Z-scores below -2 standard deviations.

Considering individual vertebral levels, the L1–L4 region showed the highest kappa scores, indicating that the lower vertebral levels are more adequately visualized and more likely to be assessed correctly by all observers using the two methods. This is in line with previous research that has reported on the difficulty of identifying vertebral fractures in the mid and upper thoracic spine in children. (8, 9,11)

A limitation of this study is that the rating of only one experienced pediatric radiologist was used as reference rather than a consensus of several radiologists. It is possible for example, that the vertebrae diagnosed as showing “physiological wedging” were in fact mild fractures. However, only a single radiologist provides the clinical report, so in this respect the study design more closely resembles clinical practice. The subjectivity of positioning the points on each vertebral body is a limitation of any quantitative morphometric technique and cannot easily be avoided. This is further complicated in children who have age-dependent changes in vertebral body ossification. A clear guideline as to where the points should be positioned in children prior to full vertebral ossification is required. The strength of this current study is that it demonstrates the utility of the 33-point software program to conduct MXA in the hands of various observers, including three pediatric radiologists, a radiographer and a clinical scientist, all with varying degrees of experience. With a reliable software program, specifically designed for use in children, non-medical staff could be trained to perform MXA. However, as emphasized by a previous study, (17) a second read by a radiologist is required to reliably differentiate fractures from non-fracture deformities. It should be noted that while we categorized fractures as mild and “clinical” based on the likelihood of clinicians to treat, it has been shown that a single mild vertebral fracture is predictive of future multiple vertebral fractures; the so-called vertebral fracture “cascade” (18). This emphasizes the need to more reliably differentiate mild loss of height due to fracture from mild loss of height due to physiological variants and the need to review current treatment protocols in order to prevent the “cascade”.

5. Conclusion

MXA reaches only moderate agreement when compared to the visual SQ VFA technique, with fair to moderate inter and intraobserver agreement. Further studies in children of current MXA software are not warranted. In order to facilitate the detection of mild vertebral fractures in children, a pediatric standard is required which not only incorporates specific vertebral body height ratios but also the age-related physiological changes in vertebral shape that occur throughout childhood.

Acknowledgments

The first author (F. F. Alqahtani) was sponsored by Najran University, Ministry of Education, Kingdom of Saudi Arabia (KSA)

Declarations of interest

None

References

1. Bishop N. Characterising and treating osteogenesis imperfecta. *Early Hum Dev*. 2010; **86**:743–6.
2. Huber AM, Gaboury I, Cabral DA, et al. Prevalent Vertebral Fractures Among Children Initiating Glucocorticoid Therapy for the Treatment of Rheumatic Disorders. *Arthritis Care Res*. 2010; **62**:516–26.
3. 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**:1326–34.
4. 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**:275–80.
5. Birnkrant DJ, Bushby K, Bann CM, et al. Diagnosis and management of Duchenne muscular dystrophy, part 2: respiratory, cardiac, bone health, and orthopaedic management. *Lancet Neurol*. 2018;**17**:347-61.
6. Dietz AC, Savage SA, Vlachos A, et al. Late effects screening guidelines after hematopoietic cell transplantation for inherited bone marrow failure syndromes: consensus statement from the second pediatric blood and marrow transplant consortium international conference on late effects after pediatric HCT. *Biol Blood Marrow Transplant*. 2017;**23**:1422-8.
7. Adiotomre E, Summers L, Allison A, et al. Diagnostic accuracy of DXA compared to conventional spine radiographs for the detection of vertebral fractures in children. *Eur Radiol*. 2017; **27**:2188–99.
8. Crabtree N, Chapman S, Högler W, et al. Vertebral fractures assessment in children: evaluation of DXA imaging versus conventional spine radiography. *Bone*. 2017; **97**:168–74.
9. Diacinti D, Pisani D, D'Avanzo M, et al. Reliability of Vertebral Fractures Assessment (VFA) in Children with Osteogenesis Imperfecta. *Calcif Tissue Int*. 2015; **96**:307–12.
10. Kyriakou A, Shepherd S, Mason A, Ahmed SF. A critical appraisal of vertebral fracture assessment in paediatrics. *Bone*. 2015; **81**:255–9.
11. Siminoski K, Lentle B, Matzinger MA, Shenouda N, Ward LM, Canadian SC. Observer agreement in pediatric semiquantitative vertebral fracture diagnosis. *Pediatr Radiol*. 2014; **44**:457–66.
12. Adiotomre E, Summers L, Allison A, et al. Diagnosis of vertebral fractures in children: is a simplified algorithm-based qualitative technique reliable? *Pediatr Radiol*. 2016; **46**:680–8.
13. Alqahtani FF, Messina F, Kruger E, et al. Evaluation of a semi-automated software program for the identification of vertebral fractures in children. *Clin Radiol*. 2017; **72**:e904– 11.
14. Bromiley PA, Adams JE, Cootes TF. Automatic localisation of vertebrae in DXA images using random forest regression voting. In: Vrtovec T. et al. (eds) International Workshop on Computational Methods and Clinical Applications for Spine Imaging; 2015: Springer. Cham
15. Lindner C, Bromiley PA, Ionita MC, Cootes TF. Robust and accurate shape model matching using random forest regression-voting. *IEEE Trans Pattern Anal Mach Intell*. 2015; **37**:1862–74.
16. Genant HK, Wu CY, Vankuijk C, Nevitt MC. Vertebral fracture assessment using a semiquantitative technique. *J Bone Miner Res*. 1993; **8**:1137–48.
17. Zeytinoglu M, Jain RK, Vokes TJ. Vertebral fracture assessment: Enhancing the diagnosis, prevention, and treatment of osteoporosis. *Bone*. 2017; **104**:54–65.
18. Christiansen BA, Bouxsein ML Biomechanics of vertebral fractures and the vertebral fracture cascade Curr Osteoporos Rep 2010;8:198-204

Highlights

- Accuracy of software programs for vertebral fracture diagnosis in children is poor
- A 33-point software program has better diagnostic accuracy than a 6-point program
- The greatest challenge is differentiating physiological wedging from mild fractures
- Bespoke software (developed using a population of healthy children) is required