



This is a repository copy of *Evaluation of a semi-automated software program for the identification of vertebral fractures in children.*

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/116988/>

Version: Accepted Version

Article:

Alqahtani, F.F., Messina, F., Kruger, E. et al. (5 more authors) (2017) Evaluation of a semi-automated software program for the identification of vertebral fractures in children. *Clinical Radiology*, 72 (10). 904.E11-904.E20. ISSN 0009-9260

<https://doi.org/10.1016/j.crad.2017.04.010>

Article available under the terms of the CC-BY-NC-ND licence
(<https://creativecommons.org/licenses/by-nc-nd/4.0/>)

Reuse

This article is distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) licence. This licence only allows you to download this work and share it with others as long as you credit the authors, but you can't change the article in any way or use it commercially. More information and the full terms of the licence here: <https://creativecommons.org/licenses/>

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.



eprints@whiterose.ac.uk
<https://eprints.whiterose.ac.uk/>

1 Evaluation of a semi-automated software program for the 2 identification of vertebral fractures in children

3 4 Introduction

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

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

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

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

31 **Materials and methods:**

32 **Study population**

33 This study involved the retrospective analysis of images obtained as part of a larger prospective study of 250
34 children recruited between November 2011 and February 2014 [7]. All images used in this study were of
35 patients recruited from our single center. The mean age of the 137 subjects at the time of image acquisition was
36 12.0 years (range 5 to 15) and 45 (33%) were male. The majority, 199 (80%) had suspected reduction in BMD
37 (including children with osteogenesis imperfecta, inflammatory bowel disease, rheumatologic conditions, cystic
38 fibrosis and celiac disease). The remaining 51 (20%) patients were recruited from spine clinic.

39 **Lateral spine imaging**

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

47 **Image analysis**

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

54 As the first step in the semi-automated analysis using SpineAnalyzer, observers identify the T4 to L4 vertebral
55 bodies by placing a point at or close to the center of each vertebral body and indicating to the software the

56 highest identified vertebral body (for example, T4). Having indicated T4, the software program recognizes all
57 identified vertebral bodies between T4 and L4 and automatically identifies six points corresponding to the four
58 corners and the midpoints of the superior and inferior endplates of each vertebral body – observers modify the
59 placement of these points as necessary. The software does not recognize vertebral bodies above T4 or below L4
60 (Fig 1).

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

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

70 **Statistical Analysis**

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

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

79 **Approvals**

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

83

84 **Results**

85 **Prevalence of fractures**

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

89 **Readability of radiographic lateral spine images within SpineAnalyzer software program**

90 Of the possible total 1781 vertebrae, from T4 through to L4 (i.e. 13 vertebrae per subject in 137 subjects), 1310
91 (73.55%) were adequately visualized by Observer 1, 1370 (77%) by Observer 2, 1376 (77%) by Observer 3 and
92 1319 (74%) and 1344 (75%) by Observers 4 and 5 respectively (Fig 2). A total of 1187 (67 %) were adequately
93 visualized by three or more observers, permitting comparison of morphology results. The visibility was
94 relatively limited in the upper part of the thoracic spine; T4 was the least readable level, being adequately
95 visualized by all observers on 423 (62%) radiographs.

96 Sensitivity and specificity values of the observers' readings with their 95% confidence intervals are presented in
97 Table 3. T6 had the highest and L3 the lowest sensitivity, while L4 had the highest and T11 the lowest
98 specificity. Overall sensitivity was 18% (95% CI, 14 – 22), while overall specificity was 97% (95% CI, 97 –
99 98).

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

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

111

112 **Discussion:**

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

115 There is a relatively low observer reliability for current techniques of VF diagnosis in children; with reported
116 kappa values for inter and intraobserver reliability ranging from 0.39 to 0.59 and 0.33 to 0.84 respectively
117 [2,7,8]. A recent study in adults showed an agreement between SpineAnalyzer and readers ranging from 0.96 to
118 0.97 [17]. The authors suggested that SpineAnalyzer is an accurate tool for measuring vertebral height and
119 identifying VFs in adults. The purpose of this current study was to evaluate the accuracy and reliability of the
120 semi-automated 6-point technique for diagnosing VF in children. To our knowledge, this evaluation is the
121 largest to assess vertebral morphometry in children using semi-automated 6-point technique software, with only
122 one other study on the same subject published to date [8].

123 Compared to our results, observer reliability has been shown to be higher in studies of the diagnostic accuracy
124 of VF detection in adults using both visually-based scoring systems and software [14-17]. A recent study on
125 children [2], based on the observation of radiographic images utilizing Genant's semi-quantitative (SQ)
126 technique, showed higher inter-kappa agreement for VF diagnosis ($k=0.45$ to 0.54) than both our
127 corresponding SpineAnalyzer calculations ($k = 0.05$ to 0.47) and those of Crabtree et al ($k = 0.36$ to 0.41) [8].
128 Results of the three studies should be directly comparable, given that the SpineAnalyzer categories are based on
129 Genant's scoring system. It seems that small differences between observers in point placement account for the
130 reduced observer reliability of SpineAnalyzer, compounded by the fact that the final categorization is based on
131 ratios and not simple measurements. This is supported by the fact that the pediatric study from which images for
132 this report were drawn also obtained a higher level of interobserver agreement ($k = 0.394$ to 0.455) when
133 utilizing a simplified algorithm-based qualitative (sABQ) technique for vertebral morphometry [10].

134 Agreement between the observers reached a maximum kappa of 0.47 (95% CI, 0.18, 0.76) with the greatest
135 level of agreement being at T12 and L4 (fair to moderate) whilst the least was at T4 (slight to poor). At each
136 vertebral level, there was diversity in the interobserver agreement and readability of the vertebra (Fig.4). Results
137 suggest that the observers could visualize the lower vertebral levels for point placement more adequately and
138 that the calculations were correspondingly more precise than those made for the upper vertebral levels,
139 underlining the difficulty in applying SpineAnalyzer for the upper thoracic spine. These findings support those
140 of previous studies reporting that identification of vertebrae in the mid and upper thoracic spine is one of the
141 major challenges in identifying VF in children [2; 3]. Reasons for poor visibility include the summation caused

142 by intrathoracic tissues and shoulders; poor image quality; and patient positioning. Therefore, the patient
143 positioning protocol and radiographic parameters selected for imaging larger patients play an important role in
144 improving image quality and visibility, in order that upper thoracic vertebrae can be assessed. In this regard, it
145 should be noted that lateral spine DXA allows improved visibility of the upper thoracic spine compared to
146 radiographs [7], which may account for the improved observer reliability of SpineAnalyzer in the study by
147 Crabtree et al [8] compared to this current study. Finally, variability in observer reliability may be related to
148 differences in identifying T12/L1. In future studies, this limitation can be countered by having a marker placed
149 adjacent to an agreed vertebra so that all observers recognize the same vertebral levels.

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

161 Another factor that affects sensitivity of the software is observer skill and experience. Although in theory no
162 medical knowledge/specialized skills are required to identify the four corners of the vertebral bodies and center
163 of inferior and superior endplates, small differences in placement affect the overall height ratios and factors
164 confounding point placement and/or fracture categorization include visibility of vertebrae, early ossification of
165 apophyses, physiological wedging and non-fracture related irregularities of vertebral endplates. Observers in
166 this study included a musculoskeletal consultant radiologist, 2 radiographers and 2 medical students. Despite the
167 training received, the disparate experience of the observers may be a weakness of the study, particularly given
168 the confounding influence of physiological variations on point placement. This will need to be considered if
169 such programs are to be used for role extension. If the 6-point or any semi-automated systems are to be more
170 accurate and reliable, then a precise algorithm is required describing where the points should be placed if, for
171 example, the apophyses are unossified and having ossified, prior to fusion. The difficulty in reproducible point-

172 placement is also reflected by the low intraobserver reliability, even for the experienced radiologist. While the
173 purpose of this current study was specifically to address the reliability of SpineAnalyzer amongst non-
174 radiologists, in retrospect, and particularly given the poor observer reliability, it would have been interesting to
175 have recruited and compared the results of at least two pediatric (or musculoskeletal) radiologists. This
176 limitation of the current study is a future objective.

177

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

184

185

186 **Acknowledgment:**

187

188 [REDACTED]

189 [REDACTED]

190 [REDACTED]

191 [REDACTED]

192

193

194

195

196

197

198

199

200

201

202 **References**

203

204 1 Bishop N, Arundel P, Clark E, et al. Fracture Prediction and the Definition of Osteoporosis in Children
205 and Adolescents: The ISCD 2013 Pediatric Official Positions. *J Clin Densitom.* 2014;17(2):275-280.

206 2 Siminoski K, Lentle B, Matzinger MA, Shenouda N, Ward LM, Canadian SC. Observer agreement in
207 pediatric semiquantitative vertebral fracture diagnosis. *Pediatr Radiol.* 2014;44(4):457-466.

208 3 Kyriakou A, Shepherd S, Mason A, Ahmed SF. A critical appraisal of vertebral fracture assessment in
209 paediatrics. *Bone.* 2015;81:255-259.

210 4 Huber AM, Gaboury I, Cabral DA, et al. Prevalent Vertebral Fractures Among Children Initiating
211 Glucocorticoid Therapy for the Treatment of Rheumatic Disorders. *Arthritis Care & Research.*
212 2010;62(4):516-526. doi: 10.1002/acr.20171. PubMed PMID: 20391507; PubMed Central PMCID:
213 PMC3958950.

214 5 Halton J, Gaboury I, Grant R, et al. Advanced Vertebral Fracture Among Newly Diagnosed Children
215 With Acute Lymphoblastic Leukemia: Results of the Canadian Steroid-Associated Osteoporosis in the
216 Pediatric Population (STOPP) Research Program. *J Bone Miner Res.* 2009;24(7):1326-1334. Epub
217 2009/02/13. doi: 10.1359/jbmr.090202 10.1359/jbmr.090202 [pii]. PubMed PMID: 19210218.

218 6 Shaw NJ. Management of osteoporosis in children. *Eur J Endocrinol.* 2008;159:S33-S39.

219 7 [REDACTED]

220 [REDACTED].

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

224 9 Genant HK, Wu CY, Vankuijk C, Nevitt MC, Vertebral Fracture Assessment Using A
225 Semiquantitative Technique. *J Bone Miner Res.* 1993;8(9):1137-1148.

226 10 [REDACTED]

227 [REDACTED]

228 11 R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical
229 Computing, Vienna, Austria. <https://www.R-project.org/>. Accessed 25 January 2017

230 12 Landis JR, Koch GG, Measurement Of Observer Agreement For Categorical Data. *Biometrics*
231 1977;33(1):159-174.

232 13 Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized
233 assessment instruments in psychology. *Psychol Assess.* 1994;6(4):284.

234 14 Ferrar L, Jiang G, Schousboe JT, DeBold CR, Eastell R, Algorithm-based qualitative and
235 semiquantitative identification of prevalent vertebral fracture: Agreement between different readers,
236 imaging modalities, and diagnostic approaches. *J Bone Miner Res.* 2008;23(3):417-424.

237 15 Kim YM, Demissie S, Eisenberg R, Samelson EJ, Kiel DP, Bouxsein ML. Intra-and inter-reader
238 reliability of semi-automated quantitative morphometry measurements and vertebral fracture
239 assessment using lateral scout views from computed tomography. *Osteoporos Int.* 2011;22(10):2677-
240 2688.

241 16 Oei L, Ly F, El Saddy S, et al. Multi-functionality of computer-aided quantitative vertebral fracture
242 morphometry analyses. *Quant Imaging Med Surg.* 2013;3(5):249-255.

243 17 Birch C, Knapp K, Hopkins S, Gallimore S, Rock B, SpineAnalyzer (TM) is an accurate and precise
244 method of vertebral fracture detection and classification on dual-energy lateral vertebral assessment
245 scans. *Radiography.* 2015;21(3):278-281.

246 18 Lentle B, Ma J, Jaremko JL, et al. The Radiology of Vertebral Fractures in Childhood Osteoporosis
247 Related to Glucocorticoid Administration. *J Clin Densitom.* 2016;19(1):81-88.

248 19 Cummings EA, Ma J, Fernandez CV, et al. Incident Vertebral Fractures in Children With Leukemia
249 During the Four Years Following Diagnosis. *J Clin Endocrinol Metab.* 2015;100(9):3408-3417.

250

251

252

253

254

255

256

257

258

259 **Tables**

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

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

262

263 **Table 3:** Sensitivity, specificity interobserver (kappa) and intraobserver (ICC) reliability of

264 SpineAnalyzer for vertebral fracture diagnosis in children

265

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

267 **Figure Legends**

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

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

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

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

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

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

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

286

287